ABSTRACT

ZHANG, SHENGFAN. Modeling the Complexity of Breast Cancer under Conditions of Uncertainty. (Under the direction of Dr. Julie S. Ivy).

Breast cancer is the most common noncutaneous cancer in American women. It is associated with high mortality risk; however mortality is strongly correlated with the stage at detection, where cancers detected at an early stage have improved survival. There have been recent controversies regarding breast cancer mammography screening policies. In 2009 when the United States Preventive Services Task Force (USPSTF) recommended changing the screening policy from annual screening beginning at age 40 to biennial screening for women beginning at age 50 through age 74, it generated significant commentary and initiated discussion on the topic of over-diagnosis. In fact, many agencies, including the American Cancer Society (ACS), did not adopt the USPSTF screening policy.

While breast cancer does not develop in the same way for everyone, these current guidelines are designed for the average population, without regard to personal background including demographic information, health history, family history and other characteristics. This dissertation research addresses the complexity of breast cancer, including disease development and outcomes, particularly under conditions in which information regarding disease progression is not known with certainty. The goal of this research is to analyze disease risk at the individual level, and build a foundation for future research to develop more personalized screening policies.

Three important topics are addressed in this dissertation. The objective of the first study is to model the impact of comorbidity on breast cancer patient outcomes (e.g., length of stay and disposition). The 2006 AHRQ Nationwide Inpatient Sample (NIS) is used to analyze

the relationships among comorbidities (e.g., hypertension, diabetes, obesity, and mental disorder), total charges, length of stay, and patient disposition as a function of age and race. Multivariate statistical analysis and survival analysis are performed to explore the effect of various comorbidities on patient outcomes. A cluster analytic model is also developed on ICD-9 codes for identifying the specific conditions that are strongly associated with breast cancer. This study illustrates the interactions and relationship among various comorbidities and breast cancer. Moreover, this study will help to improve the understanding of expenditure patterns for population subgroups with several chronic conditions and to quantify the impact of comorbidities on patient outcomes. Lastly, it also provides insight for breast cancer patients with comorbidities as a function of age and race.

The second study models mortality risk for complex patient populations. A nonparametric cumulative incidence function is developed to estimate mortality probabilities from breast cancer and other causes as a function of patient age, race, cancer stage at diagnosis and breast cancer risk factors (breast density, estrogen and progesterone receptor status, and family history of breast cancer) using population-based data from the Carolina Mammography Registry. Special ln(-ln) transformed bounds for confidence intervals are used to compare mortality estimates associated with different risk groups. Left censoring is incorporated using a cancer growth model to quantify the lag between the actual start time of breast cancer and the diagnosis time (recorded cancer start date). This study quantifies breast cancer mortality in the presence of competing risks for complex patients. Breast cancer is a disease that behaves differently in different patient populations, these differences must be considered when making screening and treatment decisions.

In most models of screening and treatment decision making, breast cancer is modeled as a progressive disease that does not regress. However, there have been medical studies that suggest breast cancer may regress without treatment. While this has initiated considerable debate in the medical community, there have been limited analytical studies on the topic. The last study in this dissertation seeks to quantify the impact of breast cancer spontaneous regression on patient outcomes with respect to different mammography screening and treatment policies. A partially observable discrete-time Markov model is built that incorporates disease regression, allowing transition from the in situ (non-invasive) cancer stage to the cancer-free state. Policy evaluation is used to quantify (in terms of lifetime breast cancer mortality risk) the impact of the probability of breast cancer regression on sample path behavior as a function of screening policies and various treatment decision rules. The American Cancer Society and the United States Preventive Screening Task Force screening policies are evaluated and compared, under different treatment decision rules. The results suggest that a woman's lifetime breast cancer mortality will be affected by disease regression and it may be worthwhile for patients to wait for treatment under these conditions.

Modeling the Complexity of Breast Cancer under Conditions of Uncertainty

by
Shengfan Zhang

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Industrial Engineering

Raleigh, North Carolina

2011

APPROVED BY:

_____	_____
Julie S. Ivy Committee Chair	Stephen D. Roberts
_____	_____
Russell E. King	Donald E. K. Martin

DEDICATION

To

my dear parents,

my beloved Zhong,

my advisor Dr. Julie Ivy,

and

loving memory of my sister

BIOGRAPHY

Shengfan Zhang was born on December 21, 1981 in Shanghai, China. After receiving her Bachelor of Management from Fudan University (Shanghai, China) in 2004, she came to North Carolina State University (Raleigh, NC) for graduate study. She received a Master of Industrial Engineering in 2006, and continued PhD study from the summer of 2007. Since joining Dr. Julie Ivy's research group in 2008, Shengfan has been conducting research on mathematical-statistical modeling and decision analysis of stochastic systems with application to health care environment. Upon graduation, she will join the faculty of the Department of Industrial Engineering at the University of Arkansas, Fayetteville, AR.

ACKNOWLEDGEMENT

I would like to extend my deepest appreciation to my advisor, Dr. Julie Ivy, for her endless support and encouragement. It is a truly honor to have Dr. Ivy as my mentor, who cares for students, and is always there when I need her advice and guidance. Especially during the difficult time in my life, her support means a lot to me. Dr. Ivy has also influenced my career choice, and I am extraordinarily excited to have the opportunity to pursue a challenging profession that I love.

My sincere gratitude goes to Dr. Bonnie Yankaskas, the Principal Investigator at the Carolina Mammography Registry (CMR), which has been funded by the National Cancer Institute. CMR has provided the data for this dissertation research, and Dr. Yankaskas's mentorship has helped me gain a better knowledge of the dataset and the direction for healthcare research. I would also like to thank Dr. Kathleen Diehl and Dr. Fay Payton, who have been working with me over the past few years, for their comments and suggestions on my work.

I am grateful to have Dr. James Wilson as my mentor for the Preparing the Professoriate program at NC State, who also prepares recommendation letters for me during my academic job search. I also want to thank my committee members: Dr. Steve Roberts, for providing great comments for my job talk practice and writing reference letters; Dr. Russell King, for detailed editing on my dissertation and sharing with me the information about academic job five years ago; Dr. Donald Martin, for providing important feedback on my dissertation, and introducing statistical methods related to my work.

I would also like to thank the professors in the Department of Industrial and Systems Engineering, particularly Dr. Paul Cohen, Dr. Brian Denton, Dr. Salah Elmaghraby, Dr. Yahya Fathi, Dr. Shu-Cherng Fang, Dr. Thom Hodgson, Dr. Simon Hsiang, and Dr. Reha Uzsoy for their help and support in many ways. I want to thank my friends at NC State, particularly my peers in 375/373 Daniels Hall, Claire, Jenn, Bjorn, Emine, Jingyu, Yuan, Daniel, Sean, Hamed, and Jeremy. We have fun together for the past three years. Many special thanks to Jinhui Liu, Weiwei Kuang, Yan Du, Zhen Shen, Deyao Ren, Wei Wei, Hui Wang, Yingying Wang, Ziyu Xiao. I would not be where I am today without your help and support.

It is difficult for me to express how thankful I am for my parents who support me in any way they can. Lastly, I want to say thank you from my heart to my dear husband Zhong. I would not be able to finish my graduate studies if you were not there to support and encourage me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

1.1 Background and Motivation

Since the nineteenth century when Rudolf Virchow founded cellular pathology [1], the understanding of cancer and its potential to harm people has continued to improve. Increased knowledge of cellular pathology has resulted in the development of more precise diagnostic capability and better treatment options.

Cancer is a group of diseases characterized by uncontrolled cell division leading to the growth of abnormal tissue. It is often believed to be a progressive disease that advances through different stages [2]. Researchers have found that patients in the early stage of a progressive disease generally have a better prognosis and are more successfully treated [3]. This is particularly true for breast cancer.

Breast cancer is often defined as a progressive disease in which malignant cancer cells form in the tissues of the breast. According to the American Cancer Society, one in eight women has a chance of developing breast cancer in her lifetime, and the chance of dying from breast cancer is about 1 in 35. In 2010, approximately 207,090 new cases of invasive breast cancer were diagnosed in women, and there were around 39,840 deaths from breast cancer in women [4]. The standard taxonomy for categorizing breast cancer is given by the American Joint Committee on Cancer (AJCC) staging system based on tumor size and spread of the disease [5]. According to this taxonomy, patients with smaller tumors are more likely to be in the early stage of the disease. The staging of breast cancer at the time of

diagnosis has prognostic value for the patient as increasing patient stage is associated with increased risk of recurrence and decreased breast cancer survival. Early diagnosis is also associated with higher probability of successful treatment [4].

Mammography is currently considered to be the most effective technology for population-based breast cancer screening. A mammogram is a type of X-ray imaging to examine the breasts. The benefits of mammography include early detection of breast cancer as it can identify problems before any symptoms (e.g. lumps) show up [6]. There have been randomized clinical trials indicating that mammography may reduce breast cancer mortality by at least 24% [7, 8].

However, mammography screening recommendations have been the subject of significant debate. The American Cancer Society (ACS) guideline is commonly adopted, which recommends women start annual screening at the age of 40. As suggested by Table 1.1, most other agencies follow this guideline with minor changes. In November of 2009, the U.S. Preventive Services Task Force (USPSTF) presented a new recommendation with an older screening start age (50 yrs), an earlier ending time (age 74), and less frequent screening (biennially). While mammography is voluntary in the U.S., screening guidelines in other countries where free exams are offered also vary. However, the international guidelines have the same starting age as the USPSTF recommendation.

Table 1.1 Summary of U.S. and International Mammography Screening Guidelines [9]

	Agency	Recommendation
U.S.	American Cancer Society	Annually for women >40
	U.S. Preventive Services Task Force (since 2009)	Every other year for women 50-74
	U.S. Preventive Services Task Force (before 2009)	Every 1-2 years for women 50-69
	American Academy of Family Physicians	Every 1-2 years for women 50-69, counsel women 40-49
	American College of Obstetricians and Gynecologists	Every 1-2 years starting at age 40, yearly after 50
	American Medical Association	Every 1-2 years for women 40-49, yearly beginning at 50
International	Canadian Task Force on Preventive Health Care	Every 1-2 years for women 50-69
	NHS Breast Screening Programme in UK	Every 3 years for women >50
	BreastScreen Australia	Every 2 years for women 50-69

Most of these current guidelines are for the average population, ignoring different personal backgrounds including demographic information, health history, family history, and other characteristics. However, there are various breast cancer risk factors that are known to affect breast cancer incidence. The most commonly used breast cancer risk prediction models are the models of Gail [10] and Barlow [11]. There are many other risk models acknowledged by National Cancer Institute (NCI) [12-19]. Most of these models consider the following factors: age, race, age at first birth, family history of breast cancer (mother, sister or daughter), number of past breast biopsies, etc. [20]. In practice, doctors may recommend different screening strategies according to individual characteristics and based on their own

clinical experiences, suggesting more personalized screening polices may be desirable and may improve the cost-effectiveness of early diagnosis.

1.2 Relevant Literature

Many of the previous operations research models for cancer screening have been summarized by Alagoz et al. [21] and Ivy [22]. Most of these studies adopt a Markov chain formulation to represent disease progression. Optimization was also commonly used to identify the optimal screening decisions in these stochastic models.

When characterizing the natural history of breast cancer, many studies considered progression from preclinical (i.e., asymptomatic disease) to clinical stages (i.e., symptomatic disease). The distribution of sojourn time (i.e., the time for cancer in the preclinical stage to progress to the clinical stage) was used to estimate the transition rates. Chen et al. [23] built a three-state Markov chain model to estimate transition rates from the preclinical to clinical stages and false negative rates simultaneously. They used the Swedish Two-County Trial of breast cancer screening with mammography data from the study and control groups to estimate the sojourn time using a quasi-likelihood method. Zelen and his colleagues have a series of studies on various problems in breast cancer screening. In one of the studies (Shen and Zelen [24]), they modeled the preclinical sojourn time in the preclinical state as a piecewise density function. They used generalized least squares and maximum likelihood methods to estimate the piecewise density function and they also showed the robustness for the estimation of the distribution function.

Other studies considered breast cancer progression through the in situ, local, regional and distant stages. Plevritis et al. [25] modeled the disease transition through tumor volume distribution estimation using a maximum likelihood method. They also compared the impact of symptom-prompted detection on the tumor size and stage of invasive breast cancer in a population not screened by mammography. The Wisconsin breast cancer epidemiology simulation model by Fryback et al. [26] used a systems engineering approach to replicate breast cancer progression in the U.S. population from 1975 to 2000. Women were individually simulated in the model by matching the case counts in the Wisconsin Cancer Reporting System (WCRS) state cancer registry, and they also adjusted the results to represent the U.S. population by calibrating parameters to Surveillance, Epidemiology, and End Results (SEER) data. Some decision models on breast cancer screening policies also used the parameters estimated from this model [21].

In contrast to earlier research which focused on the average population, this dissertation research addresses the complexity of breast cancer, including disease development and outcomes, particularly under conditions in which the information on the disease progression is not known with certainty. The goal of this dissertation research is to analyze disease risk at the individual level, and build a foundation for future research to develop more personalized screening polices. This dissertation has the potential to be extended to other chronic diseases including other types of cancer, and to other non-medical deteriorating systems which are concerned with maintenance planning decisions.

1.3 Dissertation Organization

The organization of the dissertation with brief description of each section is as follows. In Chapter 2, the impact of comorbidities on breast cancer patient outcomes, including length of stay, total charges and disposition, is explored. This study supports the need to consider comorbidity, in addition to demographic information, when characterizing breast cancer outcomes and for developing screening recommendations. Chapter 3 presents mortality modeling for breast cancer patients with selected breast cancer risk factors. Mortality is an important outcome measure for quantifying the effect of screening and treatment policies. This study explores the impact of several important risk factors on mortality for breast cancer patients. A novel model of breast cancer natural history is presented in Chapter 4 where spontaneous breast cancer regression from the in situ stage to the cancer-free stage is allowed. Partially observable Markov chain models with different treatment decision rules are formulated. The dissertation concludes with a summary and directions for future research in Chapter 5.

Chapter 2 Modeling the Comorbidity Impact on Breast Cancer Patient Outcomes

2.1 Introduction

There is little guidance for responding to the needs of medically complex patients with respect to breast cancer screening and treatment, although it is acknowledged that there is a significant need to individualize care, particularly when competing risk factors are present. The goal of this study is to improve the understanding of and model the impact of comorbid conditions on breast cancer patient outcomes using the AHRQ Nationwide Inpatient Sample (NIS) 2006 dataset for patients 18 years and older. The sampling frame for the 2006 NIS is a set of hospitals that comprises approximately 90% of all hospital discharges in the United States. Using the AHRQ NIS 2006 dataset for patients 18 years and older, this study seeks to understand the relationships among demographics (age, race, gender), comorbid conditions for women with breast cancer as a primary disease scenario and the patient outcome indicators. For this study a comorbidity is defined as a "clinical condition that exists before a patient's admission to the hospital, is not related to the principal reason for the hospitalization, and is likely to be a significant factor influencing mortality and resource use in the hospital" [27]. Patient outcome indicators modeled include patient disposition, total charges, and length of stay.

For the comorbid conditions, this study focuses on chronic diseases, as defined by the Centers for Disease Control and Prevention (CDC). *At A Glance: Chronic Diseases - The*

Power to Prevent, The Call to Control [28] defines these chronic medical conditions as "noncommunicable illnesses that are prolonged in duration, do not resolve spontaneously, and are rarely cured completely." Specifically in this study, hypertension, diabetes, mental disorder, and obesity are considered. These are common comorbidities and risk factors for aging patients and have prevalence rates of approximately 46%, 20%, 31%, and 6%, respectively, in NIS 2006 data. Obesity is also a risk factor for breast cancer. Although chronic diseases are controllable, seven in every ten Americans who die each year have at least one chronic disease [29]; they also account for billions dollars in healthcare costs annually. For example, the direct and indirect costs associated with diabetes are estimated at $174 billion each year. The healthcare costs of persons with chronic diseases account for more than 75% of the nation's $2 trillion medical care costs. The impacts of comorbidities along with the varied taxonomy to classify ICD-9 codes and diagnostic related groups (DRGs) raise clinical and health data management issues. For example, Suthummanon and Omachonu [30] investigated the feasibility of applying cost minimization analysis in determining length of stay for the primary DRGs with the highest volume from four payer classes: self-pay, Medicare, Medicaid, and commercial. It is hypothesized that patients with more severe illnesses tend to require more hospital resources than those with fewer conditions despite being admitted to the hospital for a similar reason. Further, it is contended that in the United States healthcare system, comorbidity carries considerable influence in determining a reasonable estimate of length of hospitalization under the DRG classification of diseases. Hence, the need to incorporate comorbid conditions in the modeling, prediction,

and/or estimation of healthcare outcomes is evident as observed by Starfield, Lemke, Bernhardt, Foldes, Forrest, and Weiner [31]:

> *In view of the high degree of comorbidity, even in a nonelderly population, single-disease management does not appear promising as a strategy to care for patients… New paradigms of care that acknowledge actual patterns of comorbidities as well as the need for close coordination between generalists and specialists require support.*

Modeling comorbid conditions can provide critical understanding of disease management, cost structures, and resource utilization. This is particularly true for understanding the needs of breast cancer patients. The presence of three or more comorbid conditions has been associated with a fourfold higher rate of all-cause mortality and a twenty-fold higher rate of mortality from causes other than breast cancer at three years (compared with women with primary breast cancer with no comorbid conditions) [32].

Breast cancer is a progressive disease in which malignant cancer cells form in the tissues of the breast. According to the United States Cancer Statistics Working Group: 1999-2005 Incidence and Mortality Web-based Report [33], breast cancer is the most common form of cancer in women behind non-melanoma skin cancer. Figure 2.1 shows cancer incidence rates among women in the U.S. from 1975 to 2006. It shows the high incidence rates of breast cancer compared to other noncutaneous cancers. For women aged 40 to 79, cancer is the leading cause of death, with breast cancer as the most common cancer in American women, accounting for (207,090) of all new cancer cases and 15% (39,840) of all cancer deaths among women in 2010 [34]. Figure 2.2 shows cancer death rates among

women in the U.S. from 1930 to 2006. It indicates that breast cancer is among the leading causes of cancer deaths among U.S. women.

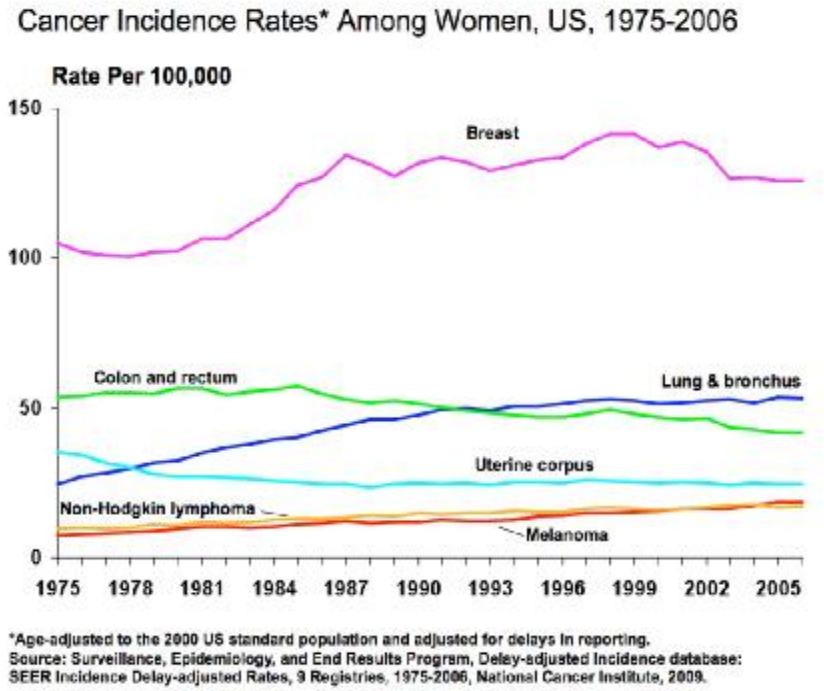


Figure 2.1 Cancer incidence rates among U.S. women [34]

Statistics show that breast cancer is the second-most-common cause of cancer across most ethnic groups, including Native American, African American, Caucasian and Asian women. Among Hispanic women, it is the leading cause of cancer deaths. Although African American women have a lower incidence of breast cancer (113.0 per 100,000 from 2002 to 2006 compared to 123.5 for Caucasians), they have a strikingly high mortality (33.0) in comparison to other groups (23.9 for Caucasians, 12.5 for Asians and 17.6 for Hispanics) [34]. The higher breast cancer mortality rate among African American women is related to

the fact that, relative to Caucasian women, a larger percentage of their breast cancers are diagnosed at a later, less-treatable stage. This is due in part to early incidence and possibly more aggressive cancers. In addition, those with comorbid conditions have increased chance of more severe disease and other medical complications. In fact, Tammemagi et al. [35] suggest that high incidence of comorbidity in African American women may play a role in the racial disparity among breast cancer patients. Their results show that more African American breast cancer patients die from competing causes than of breast cancer and suggest that effective control of comorbidity in African American breast cancer patients should help to improve life expectancy and lead to a reduction in survival disparities.

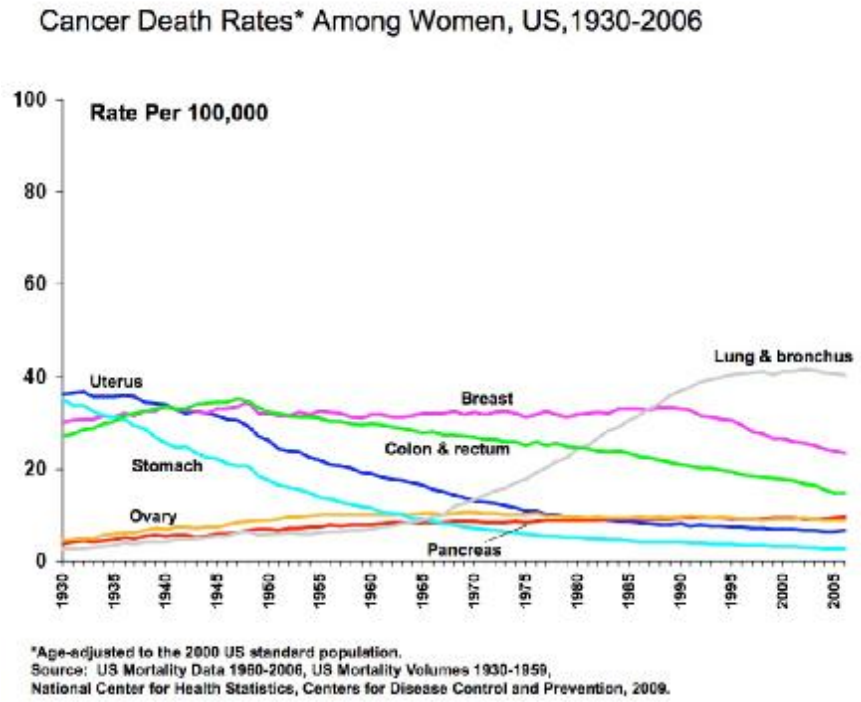


Figure 2.2 Cancer death rates among U.S. women [34]

However, much of the prior research tends to focus on cost-effectiveness analyses, policy optimization, even on clinical trials for disease in isolation. For example, cancer screening policies for the average patient often overlook the impact of competing illness, race, and age. Clinical trials purposely select participants with specific episodes of illness with emphasis on minimizing the number of complicating factors such as those created by patients with comorbid disease.

Hence, this research contributes to the field by articulating three critical points: 1) statistical modeling of chronic disease (breast cancer); 2) accounting for other comorbid conditions (diabetes, obesity, mental illness, and hypertension) in the modeling; and 3) accounting for patient age, race, length of stay, and number of procedures in the modeling. The goal is to develop models to address the challenge of the question:

Are we accurately representing the population in question or under consideration?

In this challenge, this study focuses on chronic diseases, particularly those with significant health disparities reflected in the prevalence and incidence estimates relative to breast cancer. Given the complexity of the biological, behavioral, and epidemiological factors, modeling-informed policy serves to better convey achievements, shortcomings, and challenges in the disease management [36]. For the field, the outcomes of these and other statistical modeling research can offer extensive insight into policy modeling and inform policymakers. This knowledge is critical to better understand disease management, prevention, and treatment, in general, but even more salient for chronic diseases and disparities, in particular.

The rest of this chapter is organized as follows. Section 2.2 is a literature review of previous related studies on patient outcomes. Section 2.3 introduces the methods used in this analysis including descriptive statistics, ordinary least square regressions, logistic regression, survival analysis, principal component analysis and cluster analysis. Section 2.4 contains the results for each model. The chapter concludes with a brief discussion in Section 2.5.

2. 2 Relevant Literature Review

The NIS data, which is produced by the HCUP-3 (Agency for Health Care Policy and Research, Rockville, Maryland), has been widely used in prior studies to assess disease conditions, cost estimates, patient demographics, principal procedures, and other points of interest. Zhao, Wong, and Arguelles [37] used NIS 1991 and 1992 to study the distribution of leimyoma relative to length of stay, mean costs of care, diagnoses, principal procedures performed and admission types among women aged between 15 and 64. Others [38] examined the impact of age and Medicare status on bariatric surgical outcomes. These researchers applied regression modeling to determine the effects of complicated diabetes mellitus, electrolyte disorder, anemia, and depression on bariatric surgery. Similarly, Meguid, Brooke, Chang et al. [39] applied multivariate logistic regression to lung cancer resections in the Nationwide Inpatient Sample (NIS) dataset from 1998 to 2004. While there is an extensive literature exploring the effect of age and comorbidity in postmenopausal breast cancer patients [40-42], this literature has just begun to explore the impact of comorbidity on outcomes for patients with breast cancer; it has not quantified the impact in terms of charges

and length of stay and has not attempted to identify specific comorbid conditions that most significantly impact breast cancer patient outcomes.

An improved understanding of comorbid conditions can assist in epidemiological and health services research. To this end, relationships among comorbidities can impact prognosis, detection, and disease outcomes [43]. Although a "gold standard" for measuring comorbidity does not exist, prior studies have used valid indices to predict or assess healthcare expenditures [44, 45], health services utilization among osteoarthritis patients [45], prognostic information in a hospital-based cancer registry [46], illness burden measures among breast cancer patients, just to name a few. In an extensive review of the literature to assess how to measure comorbidity, de Groot, Beckerman, Lankhorst, and Bouter [47] uncovered the use of a plethora of terms to describe the coexisting disease and multimorbility. In their search of Medline (from January 1966 to September 2000) and Embase (from January 1988 to September 2000), de Groot et al.[47] determined that the Charlson Index is the most extensively used metric for predicting mortality, although others, such as the Cumulative Illness Rating Scale (CIRS), Index of Coexistent Disease (ICED), and Kaplan Index, are valid and reliable.

To capture a more complete picture of the patient length of stay with censoring information (i.e., the actual hospitalization time is not observable or truncated), survival analysis is also used in the literature to study hazards or conditions that affect a patient's hospital time. Li [48] compared several survival models to predict the expected length of stay from medical complexity level and age. A goal of this dissertation study is to compare the length of stay based on the different conditions present. Sá et al. [49] compared survival

models and competing risk models with different distribution assumptions for the hazard function and showed parameter estimates for length of stay are sensitive to underlying assumptions.

In a recent study, Roehrig et al. [50] developed a model for estimating personal health expenditures (PHE) by medical condition, including multiple chronic conditions, which could be used to understand the sources of expenditure growth. A goal of this dissertation research is to understand and model for a specific condition (breast cancer), how patient outcomes, as defined by total charges, length of stay, and disposition, are affected by the presence of specific comorbid conditions (hypertension, diabetes, mental disorder, and obesity) that are prevalent in the population.

This research is unique in its goal to model the interaction effect of breast cancer and comorbid disease as it relates to inpatient outcomes. Rather than focusing on a single disease in isolation, this study develops integrated statistical models to explore the interrelationship between comorbid disease and the resulting outcomes.

2.3 Methods

2.3.1 Data Summary

The Nationwide Inpatient Sample (NIS) 2006 data by Agency for Healthcare Research and Quality (AHRQ) [51] is used for the analysis. It contains discharge data from 1,045 hospitals in 38 states, representing approximately a 20% stratified sample of US

community hospitals. There are 6,712,893 observations in the dataset representing individuals who are 18 years and older.

Both diagnosis and treatment ICD-9 codes are used to identify the breast cancer patient group. The diagnosis of malignancy in breast tissues is considered and the V codes for personal history of malignant neoplasm are included (ICD-9 codes: 174.0-174.9, 175.0, 175.9, 198.2, 172.5, 173.5, 232.5, 216.5, 233.0, V10.3). The procedure codes for breast cancer diagnosis and treatment are also considered, including biopsy (the vast majority, over 99% of these patients also have an ICD-9 code for breast cancer), lumpectomy, excision of lymph nodes, and mastectomy (ICD-9 codes: 85.11, 85.12, 85.21, 85.22, 85.23, 40.29, 40.23, 40.3 85.41, 85.42, 85.43, 85.44, 85.45, 85.46) [52].

In this analysis, for identifying breast cancer patients, we consider two breast cancer groups:

a) “Primary Breast Cancer Group”: Records in which breast cancer is the primary condition for hospitalization as defined by diagnosis code 1 and procedure code 1 (DX1 and PR1, respectively), i.e., if the primary diagnosis and primary procedure are breast cancer-related codes (specifically, in the analysis only DX1 is used to identify breast cancer).

b) “Breast Cancer Prevalence Group”: Records in which breast cancer is one of the patient diagnoses (primary or non-primary), i.e., in the analysis DX1 to DX15 are used to identify breast cancer.

Unlike other research on comorbidities, this study does not use the DRG group to identify comorbid conditions because the focus of DRG codes is resource utilization and hence they are less clinically relevant. Instead, to better identify comorbid conditions, this

study uses diagnosis information (ICD-9-CM) to identify disease. Comorbid conditions are defined as chronic diseases in addition to the index condition (i.e., the one to which a therapeutic chronic disease intervention is targeted). Specifically, additional chronic conditions diagnosed for a breast cancer patient (hypertension, diabetes, mental disorder, and obesity) are considered to be comorbidities. These conditions are the most common in aging patients; and in the NIS 2006 data, the percentage of occurrence among prevalence breast cancer patients are approximately 54%, 20%, 27%, and 6%, respectively. The ICD-9 codes selected for comorbidities are summarized in Table 2.1.

Table 2.1 Summary of ICD-9 Codes for Selected Comorbid Conditions [53,54]

Comorbid Condition	ICD-9-CM Codes
Hypertension	401-405, 430-438, 4258, 4290-4293, 4298-4299, 7962, 36211
Diabetes	250
Mental Disorder	290-319
Obesity	27800, 27801, 27802, V85.30-V85.4

2.3.2 Overview of Modeling Approach

Statistical models are developed and the results compared to explore the impact of comorbidities on breast cancer patient outcomes. Specifically, regression models using stepwise least squares are developed to identify the statistically significant relationships for the log transform of the dependent variables (length of stay and total charges). Because length of stay (LOS) and total charges both have skewed distributions [27], log transforms of the dependent variables are used. Logistic regression models are developed to identify the factors that significantly affect the chance of dying during hospitalization and those factors that significantly affect the chance of being transferred to another care facility at discharge.

Survival analyses with censoring information are performed, and both the parametric regression model and the Cox proportional hazards model are used to further quantify the factors affecting breast cancer patient outcomes. After identifying the general structure of the associations, mental disorder ICD-9 codes are decomposed using principal component analyses (PCA) and cluster analysis (CA). These analyses identify those mental disorders that are closely related to breast cancer, and mental conditions that are highly correlated with each other.

2.3.3 Descriptive Statistical Analysis

A descriptive statistical analysis (e.g., calculation of summary statistics such as the mean and standard deviation) is performed on important variables for characterizing the population as a function of the presence of chronic diseases. These variables include patient age (in years), length of stay (in days), total charges (in dollars); percentage of cohorts of the different genders, race (Caucasian, African American, Hispanic, Asian or Pacific Islander, Native American, and other); admission type (emergency, urgent, elective, newborn, trauma center, and other); percentage of death; total number of diagnoses; and total number of procedures.

The descriptive statistical analysis is performed on several patient population subgroups. Characteristics of the general population, the prevalence and primary breast cancer groups, and of the four comorbid disease groups (i.e., hypertension, diabetes, mental disorder, and obesity) are compared. In addition, the comorbid disease groups are compared with respect to the prevalence of breast cancer and the primary breast cancer groups.

2.3.4 Ordinary Least Square Regression

Ordinary least square (OLS) regression models are developed to predict LOS and total charges as a function of patient age, gender, race, admission type, total number of diagnosis and procedures, comorbid conditions, and disposition from the perspective of breast cancer patients. These models are used to quantify the impact of each of these factors on each breast cancer patient group (primary and prevalence). For categorical demographic characteristics, dichotomous variables are created to represent each category. Binary variables are also created for comorbid diseases (i.e., 1 for the presence of the disease and 0 otherwise). The records with a LOS of zero are assumed to correspond to a hospital stay shorter than 24 hours. A value of 0.5 is assigned to these records. This assumption translates to a hospital stay of approximately one-half day (or twelve hours).

Stepwise regression models with LOS and total charges as the dependent variables are developed for the general population. Explanatory variables include background variables summarized in the descriptive statistical analysis section. This study includes contributions of independent diseases and comorbidities (defined with an interaction term with breast cancer and another disease) of interest. Four additional explanatory variables that are investigated to determine their effects on LOS and total charges are NDX (number of diagnoses), NPR (number of procedures), DIED (indication of in-hospital death), and a binary variable to indicate transferring or not (transfer to short-term hospital or transfer to other facility). The regression model has the following form.

$$\log(LOS) = a_1 + a_2 DemoVar + a_3 NDX + a_4 NPR + a_5 DIED + a_6 Transferred$$
$$+ b_1 BC + b_2 H + b_3 D + b_4 MD + b_5 O + g_1 BC \times H + g_2 BC \times D + g_3 BC \times MD + g_4 BC \times O + e \qquad (2.1)$$

Note: *DemoVar* = (Age, Female, RaceGroup1-RaceGroup5, AdmissionType1-AdmissionType5)^T, a vector containing background variables characterizing the population. α_2 is a vector containing prediction coefficient corresponding to the variables in the *DemoVar* vector. BC=Breast Cancer, H=Hypertension, D=Diabetes, MD=Mental Disorder, O=Obesity. Regression log(total charges) is in the same form.

To justify the normal assumption after log transformation, univariate analysis is performed on the residuals from the regression on LOS and total charges for the primary breast cancer group. The statistical skewness after the transformation for LOS was reduced to 0.10 with kurtosis at 0.43, and similar for total charges with -0.08 for skewness and 0.70 for kurtosis. Quantile-Quantile (QQ) plots for testing normality are shown in Figure 2.3. From these statistics, it is acceptable (although the fit is not perfect) to assume normality after log transformation.

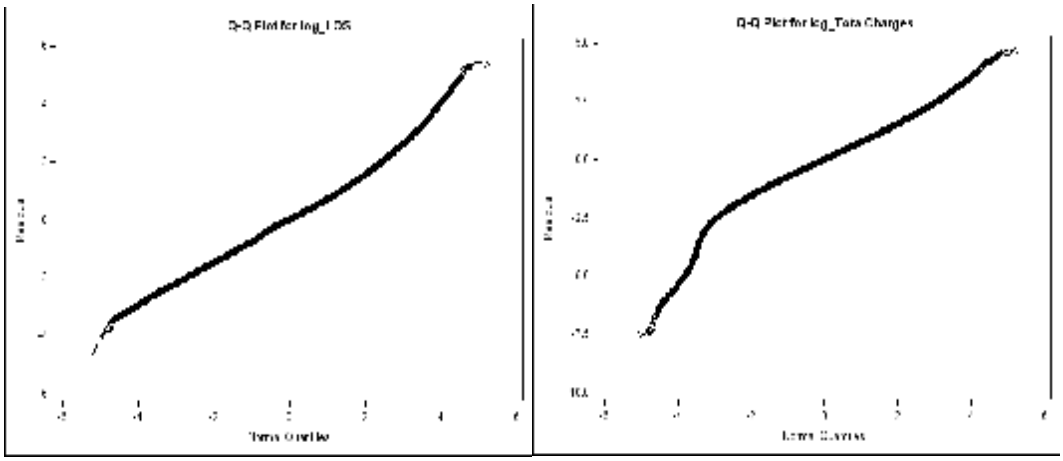


Figure 2.3 Quantile-Quantile Plots for Residuals after Log Transformation

To better observe and study the impact of age on the relationship between comorbidities and patient outcomes, separate regression models are developed on LOS and

total charges for different age groups. Although the general regression shows the effect of the independent variable “age” on patient outcomes, it does not quantify how the relationship may change between comorbidities and outcomes for different age groups. This type of age stratification highlights the role of age in understanding and characterizing the relationship between patient outcomes and comorbidity. The adult population is classified by eight subgroups at 10-year increments. The first group includes patients younger than 30 years old. The other groups correspond to 10 year patient age groups from 30 to 90, and the last group is for patients older than 90. For this analysis, only the general population with primary breast cancer is considered. For each age-based subgroup, a similar regression model is developed based on equation (2.1) considering comorbidities. This series of regressions characterize the impact of comorbidities on primary breast cancer patients among different age groups.

2.3.5 Weight Adjusted OLS Regression

NIS is a well-stratified dataset designed to represent U.S. community hospitals. Each inpatient stay record is associated with a discharge “weight” variable that describes the stratum. In order to quantify the weight impact, a weight-adjusted ordinary least squares regression model is proposed for the primary breast cancer patient group and compared with the results of an un-weighted model for the same group.

First, the discharge weight is normalized to a probability-like weight. For each stratum, the new normalized weight is

$$p_i = \frac{w_i n_i}{\sum_{l=1}^{c} w_l n_l} \tag{2.2}$$

where w_i is the original weight variable for the i^{th} weight group, n_i is the number of samples within the stratum and c is the number of strata in the sample.

When expressing a linear regression using conditional probabilities, for response variable Y, independent variables X_1 through X_k, group indicator U, the following model is assumed:

$$E[Y \mid X_1,...X_k, U = i] = b_{i0} + \sum_{j=1}^{k} b_{ij} X_j , \tag{2.3}$$

where b_{ij} is the corresponding regression coefficient for independent variable X_j and weight group i.

Taking expectation with respect to weight groups gives a regression estimate for the population. Assuming independence among stratums,

$$E[Y \mid X_1,...X_k] = \sum_{i=1}^{c} p_i b_{i0} + \sum_{j=1}^{k} (\sum_{i=1}^{c} p_i b_{ij}) X_j = b_0^* + \sum_{j=1}^{k} b_j^* X_j . \tag{2.4}$$

From the regression analysis for weight group i, independent variable X_j, the coefficient estimator $\hat{b}_{ij} \sim N(b_{ij}, s^2_{\hat{b}_{ij}})$, and the variance estimator $\hat{s}^2_{\hat{b}_{ij}}$ is χ^2 with g_{ij} degrees of freedom.

For the overall regression we have $\hat{b}_j^* = \sum_{i=1}^{c} p_i \hat{b}_{ij}$, with mean

$$E[\hat{b}_j^*] = \sum_{i=1}^{c} p_i E[\hat{b}_{ij}] = \sum_{i=1}^{c} p_i b_{ij} = b_j^* , \tag{2.5}$$

and variance

$$\hat{s}^2_{\hat{b}_j^*} \equiv Var(\hat{b}_j^*) = \sum_{i=1}^{c} p_i^2 Var(\hat{b}_{ij}) = \sum_{i=1}^{c} p_i^2 s^2_{\hat{b}_{ij}} . \tag{2.6}$$

The variance is estimated using the complex variance estimator $\hat{s}^2_{\hat{b}_j^*} = \sum_{i=1}^{c} p_i^2 \hat{s}^2_{\hat{b}_{ij}}$. From Satterthwaite's approximation [55], we have $\hat{s}^2_{\hat{b}_{ij}^*} \sim \frac{s^2_{\hat{b}_{ij}^*} c^2(g_j^*)}{g_j^*}$, where g_j^* is estimated using $g_j^* = \left\lfloor (\sum_{i=1}^{c} p_i^2 \hat{s}^2_{\hat{b}_{ij}})^2 \Big/ \sum_{i=1}^{c} \left(\frac{p_i^4 \hat{s}^4_{\hat{b}_{ij}}}{g_{ij}} \right) \right\rfloor$. Then, the overall p-value for $\hat{b}_j^*$ can be computed using $p_j^* = \Pr\left\{ | t_{g_j^*} | > \hat{b}_j^* \right\}$, where $t_{g_j^*}$ is a student t-random variable with g_j^* degrees of freedom.

Next, the results from the weight adjusted OLS regression model can be compared with the parameter estimate and significance level of the unweighted regression model.

In our analyses, when the results from the weight-adjusted regression are not different from the unweighted regression results, the stratified sample is used.

2.3.6 Logistic Regression

Mortality and transferring are two factors that have the potential to impact the measurement of patient outcomes. If a patient died or was transferred to another facility, her

(his) expected LOS and charges would be affected and thus would not be accurately reflected in the observed data. Comorbid conditions may change the mortality rate and the probability of being transferred and hence have an impact on the LOS and total charges.

Logistic regression models by age groups on LOS and total charges are developed to study the effect of comorbidities on patient disposition. Risk factors for the model include patient background variables, total number of diagnoses, total number of procedures, independent diseases, and interaction between breast cancer and comorbidities of interest. LOS and total charges are also included in the models. The models predict the probability of mortality ($\pi(Y)$) for the general population as a function of comorbidities and background characteristics (race, admission type, etc.).

$$\log\left(\frac{p(Y)}{1-p(Y)}\right) = a_1 + a_2 DemoVar + a_3 NDX + a_4 NPR + a_5 LOS + a_6 TOTCHG + b_1 BC + b_2 H + b_3 D + b_4 MD + b_5 O + g_1 BC \times H + g_2 BC \times D + g_3 BC \times MD + g_4 BC \times O \quad (2.7)$$

The same logistic regression analysis is performed for risk factors on transferred patients, where $\pi(T)$ is the probability a patient is transferred to another facility.

Reduced logistic models are also developed for the primary breast cancer group on mortality and transfer status.

$$\log\left(\frac{p(Y \mid BC)}{1-p(Y \mid BC)}\right) = a_1 + a_2 DemoVar + a_3 NDX + a_4 NPR + a_5 LOS + a_6 TOTCHG + b_1 BC + b_2 H + b_3 D + b_4 MD + b_5 O \quad (2.8)$$

Note: $\pi(Y|BC)$ = Probability the patient died during hospitalization given the patient is in the primary breast cancer group.

The predicted coefficient of the β's in equations (2.7) and (2.8) are compared. This quantifies the difference in risk factors between the general population and the primary breast cancer group. The ordinary least square regression and the logistic regression are used to analyze the impact of comorbidities on patient outcomes: LOS, total charges, and disposition.

2.3.7 Survival Models

Survival models are developed to explore the relationship between event time (hospitalization time) and comorbid conditions for breast cancer patients. Unlike OLS regressions, survival analysis considers censored information. For example, for patient LOS, censoring occurs when the real LOS was not observed and hence the variable LOS does not represent the true hospitalization time. Specifically, the following situations are defined as censoring: a patient stays in the hospital longer than 365 days (LOS>365) since the NIS data only includes records for one year; a patient is transferred to a short-term hospital (DISPUniform=2); and a patient is transferred to another facility including a skilled nursing facility, intermediate care facility, etc. (DISPUniform =5). These patients may have a longer LOS, but the recorded time is censored. Hence, the observed time in hospital $T = min\ (T', C)$, where T' is the recovery time for a patient and C is the censoring time.

The advantage of survival analysis is that it is possible to more accurately study censored information. This study uses survival models to study the impact of comorbidities on LOS. The same covariates as in the OLS regression models are considered. The survival function in the analysis is $S(t) = Prob(T>t)$, i.e., the probability a patient will stay in the hospital for time t or longer. A hazard function (or hazard rate, $\lambda(t)$) is the limit of an event

rate (for example, mortality rate or discharge rate) if the interval of time is taken to be small. It is the instantaneous rate of experiencing the event at time *t* given the individual is alive at time *t*. The hazard function is given in Equation 2.8. In this study, the hazard event is recovery (leaving the hospital).

$$l(t) = \lim_{h \to 0} \frac{P(t \le T \le t + h \mid T \ge t)}{h} \tag{2.9}$$

Two survival models are used to analyze the censored survival data. First, a parametric Accelerated Failure Time model [56] is used, assuming accelerated failure time, that is $S_1(t) = S_2(ct)$ for all t, where the constant $c>0$:

$$\log(T) = a_1 + a_2 DemoVar + a_3 NDX + a_4 NPR + b_1 BC + b_2 H + b_3 D + b_4 M + b_5 O + g_1 BC \times H + g_2 BC \times D + g_3 BC \times MD + g_4 BC \times O + se \tag{2.10}$$

Here $\alpha_1 \ldots \alpha_4$, $\beta_1 \ldots \beta_5$, $\gamma_1 \ldots \gamma_4$ are the regression coefficients of interest. The change of the coefficient *β's* and 's shows the effect of the binary variables being one (i.e., presence of the condition) on the Log of the hospitalization time while holding other coefficients constant. σ is a scale parameter and ε is the vector of the random disturbance terms.

The random disturbances are usually assumed to be independent identically distributed with a density function. Different distributions may be assumed. Thus the parametric models are developed based on the distribution of the error term. In this study, several distribution models for time *T* are compared: Weibull, Log Normal (error terms follow a normal distribution), and Log Logistic (error terms follow a logistic distribution). However since the dataset is large, it is difficult to fully visualize the distribution for the error

terms. Thus the maximum likelihood is compared to identify the best parametric distribution fit.

A Cox proportional hazard model, which is a semi-parametric model, is also used to analyze the comorbidity impact. The model has the following form:

$$l(t \mid x, y, z) = l_0(t)e^{x^T a + y^T b + z^T g} \tag{2.11}$$

Here x is a vector containing the background covariates: age, gender, race, admission type, number of diagnoses and procedures; y is a vector containing individual diseases and z is a vector containing interactions among breast cancer and comorbid conditions. The vectors of the coefficient estimates for the covariates α, β, and γ represent the change in hazard with the binary variables being one (or with an increase of one unit in the quantitative variables) while holding other covariates constant. The hazard function $\lambda(t|x,y,z)$ is the "hazard" of leaving the hospital given the covariates and λ_0 is the baseline case when $x,y,z=0$, i.e., when no comorbid condition exists and using the base case for demographic variables: hazard for a male patient having no breast cancer, hypertension, diabetes, mental disorder, or obesity, with race other and admission type other. We also study the maximum likelihood estimate of β and γ, i.e., the impact of comorbidities on LOS in hospital.

For both survival models, it is assumed that the event time T (i.e., time to leave the hospital) is independent of the censoring time C (i.e., time to be transferred or end of study period). This assumption is reasonable because the time a patient spends in the hospital does not depend on whether a patient gets transferred. However, there is another event in the survival analysis, i.e., patient death. In this study, a naïve method is adopted where the two

events are separated, patient leaving the hospital (time to leave the hospital) and patient mortality (time to death). Then the patients groups are separated into the survived group (DIED=0) and the mortality group (DIED=1). This study only analyzes the impact of comorbidities on patient LOS for the survived group.

The multivariate statistical models above characterize the impact of comorbidities on breast cancer patient outcomes. Based on these analyses, mental disorder was found to be an important and closely related comorbid condition and hence was selected for further study. The following analyses identify those specific mental disorders that are strongly associated with breast cancer, and characterize the relationship between the mental conditions.

2.3.8 Principal Component Analysis

In the regression analysis, 30 ICD-9-CM codes (290-319) are used to identify mental disorders. This is a broad category, and the goal of the PCA is to identify the most related of the 30 ICD-9 mental disorder codes for breast cancer patients. While least squares regression models explore the relationship between response variables and explanatory variables by minimizing the sum of squared residuals, PCA is a technique to reduce the dimensionality of variables while preserving the original variance and covariance structure. PCA provides a better understanding of the data by interpreting fewer principal components (PC) and removing unnecessary information. The first PC is a linear combination of the original variables (diagnosis codes) that accounts for the majority of the variation in the data. Each subsequent PC is uncorrelated with previous ones and accounts for a decreasing fraction of the remaining variation. Based on the PCA process, mental disorder ICD-9 codes can be

"ranked" and the most related conditions for the primary breast cancer patient group can be identified.

PCA is related to eigenvalue/eigenvector theory [57]. Specifically, it can be seen that $\lambda_i P_i = \sigma_i P_i$, where P_i is the i^{th} eigenvector of the covariance matrix X; σ_i is the variance of i^{th} PC, and λ_i is the corresponding eigenvalue. The total variance of the system is the sum of all eigenvalues. The i^{th} PC explains $\frac{l_i}{l_1 + l_2 + \mathbf{L}\ l_n}$ proportion of variance. Because n is large (n=30), the PCs are selected that explain 95% of the variance and thus reduces the dimensionality of the ICD-9 codes. After the principal components are selected, for each variable the weighted score is determined by multiplying each PC vector by the corresponding eigenvalue. The variables with higher scores are selected because they correspond to the most prevalent conditions for primary breast cancer patients. Given the rank for each variable, cluster analysis (CA) is performed on the variables to determine how many variables could be eliminated without changing the relationship.

2.3.9 Cluster Analysis

Cluster analysis (CA) is closely related to the PCA and also relies on the exploration of the eigenstructures in the dataset. It explores the correlation among variables and identifies the most significant condition within each cluster. This helps to determine the number of variables necessary for study and to identify the conditions that are highly correlated.

The CA analysis is performed in SAS according to the following algorithm: cluster components (similar to the principal components) are computed in each iteration, and each

variable is assigned to the component with which it has the highest squared correlation. It is then tested to see if the amount of variance explained increases if the variable is assigned to another cluster [58]. The most important variable (i.e., ICD-9 code for specific mental disorders) is selected from each cluster, i.e., the variable that has the greatest correlation with its own cluster and is the least correlated with other clusters. The ratio (1 – R-square) is used to identify these variables.

Together with PCA, the dimension of the variables is reduced, and the mental disorders that are most related are identified for the group of primary breast cancer patients.

2.4 Results

The results for each study are summarized in this section. Section 2.4.1 presents the results for the descriptive statistics analysis. In Section 2.4.2, the results for the ordinary least square regression models are discussed and compared with the weight adjusted regression model. Section 2.4.3 contains the results for logistic regression models, and in Section 2.4.4 the survival model results are summarized. Principal component and cluster analyses results are discussed in the last section.

2.4.1 Descriptive Statistics for Study Population

The results are discussed in the following two subsections. Section 2.4.1.1 focuses on the summary for the general population, and section 2.4.1.2 presents the findings for the breast cancer population.

2.4.1.1 Characterizing Patients with Diseases of Interest in Relationship to the General Population

Summary statistics for key characteristics of the population in the dataset are shown in Table 2.2. The first two columns are the means and standard deviations for background characteristics of the adult population. Similarly, these statistics are shown for the breast cancer patient group. Columns 3 and 4 are summary statistics for the prevalence breast cancer group and columns 5 and 6 are summary statistics for the primary breast cancer group. The prevalence characteristics for patients with selected diseases (hypertension, diabetes, mental disorder, and obesity) are summarized in columns 7 through 14. As shown in Table 2.2, breast cancer patients (both the primary and prevalence groups) have different characteristics than the general population and, in turn, have different patterns in LOS and total charges. For the primary breast cancer group, 65.63% are "elective" admissions (this admission type includes waiting list admission, booked admission, and planned admission). In comparison, only 25.51% of admissions in the general population are "elective." This difference may explain the shorter lengths of stay for primary breast cancer patients since they may stay in the hospital just for treatment. In general, primary breast cancer patients also have fewer diagnoses; however, they have more procedures on average than the general population. Interestingly, the total charges for breast cancer patients on average are significantly lower than the general population. This suggests that although the charges for procedures may be higher, the shorter LOS results in the lower total charges. It is also

important to note that not only women get breast cancer, but there are also about 8% males in the prevalence breast cancer group and 6% in the primary group.

Patients with hypertension, diabetes, and mental disorder have longer average LOS (5.14 to 5.38 days) than the general population (4.8). For total charges, only the category mental disorder has a lower mean than the general population; the other conditions have higher charges. In addition, the categories mental disorder and diabetes have fewer procedures performed (1.33 and 1.64) compared to general group (1.69). Further, those with the selected conditions have more diagnoses (8.09 to 9.3) than the general population (7.07). The mortality rate is lower in the mental disorder (1.88%) and obesity groups (0.88%) compared to the general population (2.38%).

Table 2.2 Summary of Characteristics of Study Population

		General Population (*n*=6,712,893)		Prevalence Breast Cancer (*n* =161,161)		Primary Breast Cancer (*n* =21,598)		Hypertension (*n* =3,079,009)		Diabetes (*n* =1,392,191)		Mental Disorder (*n* =2,109,840)		Obesity (*n* = 423,268)	
		Mean	S.D.[1]	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Age (years)		56.81	21.03	67.48	15.03	60.34	15.3	67.56	15.39	65.08	15.28	57.06	19.69	54.45	15.65
Length of Stay (days)		4.8	6.48	4.84	5.65	3.16	5.27	5.14	6.23	5.38	6.43	5.28	6.8	4.62	5.57
Total Charges ($)		26557	40304	28357	34781	***24337***	25210	29746	39589	29879	39950	***24566***	34288	28397	34429
Female (%)		60.57%	0.49	***91.22%***	0.28	***94.19%***	0.23	55.12%	0.5	53.39%	0.5	54.97%	0.5	64.57%	0.48
Race (%)	Caucasian	50.95%	0.5	60.12%	0.49	53.63%	0.5	53.16%	0.5	48.54%	0.5	54.64%	0.5	51.24%	0.5
	African American	10.09%	0.3	7.74%	0.27	10.12%	0.3	11.90%	0.32	13.32%	0.34	10.85%	0.31	13.50%	0.34
	Hispanic	9.03%	0.29	5.14%	0.22	6.70%	0.25	6.84%	0.25	9.74%	0.3	6.33%	0.24	8.33%	0.28
	Asian or Pacific Islander	1.53%	0.12	1.36%	0.12	2.02%	0.14	1.46%	0.12	1.65%	0.13	0.82%	0.09	0.52%	0.07
	Native American	0.48%	0.07	0.25%	0.05	0.28%	0.05	0.45%	0.07	0.66%	0.08	0.51%	0.07	0.46%	0.07
	Other	2.02%	0.14	1.57%	0.12	1.61%	0.13	1.67%	0.13	1.90%	0.14	1.66%	0.13	1.73%	0.13
Admission Type (%)	Emergency	45.97%	0.5	38.44%	0.49	11.13%	0.31	52.93%	0.5	54.57%	0.5	55.54%	0.5	44.99%	0.5
	Urgent	18.07%	0.38	14.18%	0.35	11.17%	0.32	15.66%	0.36	16.17%	0.37	16.34%	0.37	15.25%	0.36
	Elective	***25.51%***[2]	0.44	36.63%	0.48	***65.63%***	0.47	21.75%	0.41	19.00%	0.39	18.31%	0.39	27.77%	0.45
	Trauma Center	0.24%	0	0.07%	0.03	0.01%	0.01	0.11%	0.03	0.09%	0.03	0.27%	0.05	0.08%	0.03
	Other	0.10%	0.03	0.05%	0.02	0.00%	0	0.11%	0.03	0.10%	0.03	0.20%	0.05	0.19%	0.04
Died (%)		2.38%	0.15	2.83%	0.17	1.92%	0.14	2.57%	0.16	2.40%	0.15	***1.88%***	0.14	***0.88%***	0.09
Number of Diagnoses		7.07	3.92	7.94	3.79	***5.15***	3.29	8.69	3.72	9.3	3.79	8.09	3.74	8.8	3.66
Number of Procedures		1.69	2.05	***1.9***	2.07	***2.26***	1.52	1.69	2.22	***1.64***	2.22	***1.33***	1.95	1.69	2.11

[1] S.D. = Standard Deviation, and same in the following tables.

[2] Highlighted cells are the results specifically discussed in the text and worth attention

2.4.1.2 Understanding the Impact of Comorbid Conditions on Breast Cancer Patients in Comparison to the General Population

Tables 2.3 and 2.4 respectively, present the characteristics of the comorbidities of interest for the prevalence breast cancer and primary breast cancer patient groups. Some of the more salient differences between these patient populations are discussed below.

Table 2.3 Summary of Characteristics of the **Prevalence Breast Cancer** Group with Comorbidities

		Prevalence BC (*n*=161,163)		Hypertension (*n* = 86,526)		Diabetes (*n* =31,479)		Mental Disorder (*n* =43,028)		Obesity (*n* = 9,000)	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Age (years)		67.48	15.03	***72.38***	12.46	***70.61***	12.15	***68.04***	15.35	63.24	12.71
Length of Stay (days)		4.84	5.65	4.88	5.14	5.20	5.67	5.29	6.46	4.75	5.69
Total Charges ($)		28357	34781	***27949***	32414	28855	35119	***27083***	32667	29808	30850
Female (%)		91.22%	0.28	92.13%	0.27	93.09%	0.25	92.75%	0.26	93.02%	0.25
Race (%)	Caucasian	60.12%	0.49	60.59%	0.49	55.21%	0.50	63.67%	0.48	56.46%	0.50
	African American	7.74%	0.27	9.48%	0.29	12.38%	0.33	6.88%	0.25	11.61%	0.32
	Hispanic	5.14%	0.22	4.80%	0.21	7.40%	0.26	4.16%	0.20	6.71%	0.25
	Asian or Pacific Islander	1.36%	0.12	1.26%	0.11	1.58%	0.12	0.68%	0.08	0.50%	0.07
	Native American	0.25%	0.05	0.25%	0.05	0.40%	0.06	0.23%	0.05	0.28%	0.05
	Other	1.57%	0.12	1.49%	0.12	1.66%	0.13	1.42%	0.12	1.61%	0.13
Admission Type (%)	Emergency	38.44%	0.49	43.02%	0.50	46.35%	0.50	45.31%	0.50	32.08%	0.47
	Urgent	14.18%	0.35	13.78%	0.34	14.26%	0.35	14.18%	0.35	12.18%	0.33
	Elective	36.63%	0.48	32.45%	0.47	28.59%	0.45	29.65%	0.46	42.20%	0.49
	Trauma Center	0.07%	0.03	0.08%	0.03	0.05%	0.02	0.08%	0.03	0.02%	0.01
	Other	0.05%	0.02	0.08%	0.03	0.08%	0.03	0.08%	0.03	0.12%	0.03
Died (%)		***2.83%***	0.17	2.40%	0.15	2.61%	0.16	2.33%	0.15	1.10%	0.10
Number of Diagnoses		7.94	3.79	9.00	3.56	9.78	3.56	9.29	3.59	9.52	3.68
Number of Procedures		1.90	2.07	***1.79***	2.07	1.75	2.11	1.64	2.02	***2.16***	2.12

In comparison to the general population, the mean patient age with each comorbid condition is much higher in the prevalence breast cancer group. Similar to the general population, patients with comorbidities also have longer LOS in both breast cancer groups. For prevalence breast cancer group, only hypertension and mental disorder have lower

charges. However, in the primary breast cancer group, all comorbidities result in lower charges. This may relate to the differences in the types of admission between these groups. With respect to the admission type, as stated above, breast cancer patients are primarily admitted as "elective." Similarly, in the primary breast cancer group with comorbidities, each comorbid condition patient group has a high percentage of elective admissions, all over 60%. In comparison, for the general population only 25.51% of admissions are elective. This is also reflected in the number of procedures.

Table 2.4 Summary of Characteristics of the **Primary Breast Cancer** Group with Comorbidities

		Primary BC (*n*=21,598)		Hypertension (*n*=8,813)		Diabetes (*n*=3,157)		Mental Disorder (*n*=4,413)		Obesity (*n*= 1,272)	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Age (years)		60.34	15.30	67.57	12.97	66.17	12.65	59.38	15.35	59.18	20.00
Length of Stay (days)		***3.16***	5.27	3.23	5.05	3.82	6.75	3.71	7.52	3.28	3.99
Total Charges ($)		24337	25210	***23207***	24611	***25609***	31796	***25159***	24649	***25189***	24330
Female (%)		94.19%	0.23	93.98%	0.24	93.35%	0.25	93.40%	0.25	95.20%	0.21
Race (%)	Caucasian	53.63%	0.50	52.46%	0.50	46.85%	0.50	55.88%	0.50	50.39%	0.50
	African American	10.12%	0.30	13.68%	0.34	15.77%	0.36	9.47%	0.29	15.41%	0.36
	Hispanic	6.70%	0.25	5.90%	0.24	8.58%	0.28	4.80%	0.21	7.08%	0.26
	Asian or Pacific Islander	2.02%	0.14	1.80%	0.13	2.53%	0.16	0.79%	0.09	0.63%	0.08
	Native American	0.28%	0.05	0.30%	0.05	0.48%	0.07	0.29%	0.05	0.08%	0.03
	Other	1.61%	0.13	1.30%	0.11	1.20%	0.11	1.54%	0.12	1.26%	0.11
Admission Type (%)	Emergency	11.13%	0.31	11.61%	0.32	14.41%	0.35	14.43%	0.35	11.71%	0.32
	Urgent	11.17%	0.32	11.88%	0.32	12.86%	0.33	11.44%	0.32	9.98%	0.30
	Elective	65.63%	0.47	65.44%	0.48	61.55%	0.49	63.97%	0.48	62.74%	0.48
	Trauma Center	0.01%	0.01	0.00%	0.00	0.00%	0.00	0.00%	0.00	0.00%	0.00
	Other	0.00%	0.00	0.00%	0.00	0.00%	0.00	0.00%	0.00	0.00%	0.00
Died (%)		1.92%	0.14	1.21%	0.11	1.39%	0.12	1.61%	0.13	0.47%	0.07
Number of Diagnoses		5.15	3.29	6.53	3.19	7.47	3.44	6.84	3.34	7.62	3.36
Number of Procedures		***2.26***	1.52	2.10	1.43	2.07	1.47	2.26	1.54	2.23	1.57

With respect to the number of procedures, generally, breast cancer patients have more procedures than the general population. In the prevalence breast cancer group, those patients

with hypertension have fewer procedures compared to the general prevalence subgroup, and those who are obese have more (2.16). However, in the primary breast cancer group, all patients with comorbidities have fewer procedures (2.07 to 2.23) than the general subgroup (2.26). The primary breast cancer group on average has fewer diagnoses than the general population. Similar to the general population with comorbidities, those patients with the comorbid conditions also have more diagnoses, 9.00 to 9.78 compared to 7.94 for the prevalence group (6.53 to 7.47 compared to 5.15 for the primary group). All mortality rates for patients with comorbidities in the prevalence breast cancer group are lower (1.10% to 2.61%) than in the general subgroup (2.83%), and similarly for the primary breast cancer group.

Descriptive statistics show that breast cancer patients with selected chronic comorbidities behave differently than the general population, which results in different LOS and total charges. Next, regression models are used to study the detailed effect of comorbidities on patient outcomes adjusting for other effects.

2.4.2 Least Squares Regression on Patient Outcomes

2.4.2.1. Characterizing the Relationship between LOS (Total Charges) and Patient Background Characteristics, Comorbidities for General Population and Breast Cancer Patients

Using stepwise selection (with a 0.1 significance level) to identify significant relationships in the regression analysis, we first develop regression models for the general

population and compare the two cases where breast cancer is considered according to the primary breast cancer and prevalence breast cancer groups. The regression models in each case have R-squares of approximately 22% for the log(LOS) and 37% for the log(total charges). The transformed coefficient estimates (i.e., exponential transform of the log coefficient) are presented in Table 2.5. It should be noted that most of the interaction terms for breast cancer and a comorbid condition are significant, which indicates that comorbid conditions affect outcomes for breast cancer patients in terms of both charges and LOS.

Table 2.5 Ordinary Least Squares Regression for the General Population

		Length of Stay				Total Charges			
		Prevalence		Primary		Prevalence		Primary	
		Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
Age		1.0024	<.0001	1.0018	<.0001	1.0059	<.0001	1.0062	<.0001
Race	Caucasian	0.96	<.0001	0.9626	<.0001	0.9893	<.0001	0.9889	<.0001
	African American	1.1408	<.0001	1.1335	<.0001	1.05	<.0001	1.0519	<.0001
	Hispanic	1.0516	<.0001	1.0417	<.0001	1.0685	<.0001	1.0703	<.0001
	Asian or Pacific Islander	1.0683	<.0001	1.0516	<.0001	1.0535	<.0001	1.056	<.0001
	Native American	[1]				0.9208	<.0001	0.9194	<.0001
Admission Type	Emergency	1.1034	<.0001	1.1095	<.0001	0.6998	<.0001	0.6983	<.0001
	Urgent	1.11	<.0001	1.1075	<.0001	0.5867	<.0001	0.5866	<.0001
	Elective	1.0161	<.0001	1.0123	<.0001	0.6983	<.0001	0.7004	<.0001
	Newborn	1.1075	0.0015	1.0914	0.0049	0.5293	<.0001	0.53	<.0001
	Trauma Center	1.2371	<.0001	1.2567	<.0001	1.2922	<.0001	1.2866	<.0001
Died		***0.9638***	<.0001	0.9538	<.0001	***1.047***	<.0001	1.0418	<.0001
Transferred		***1.4534***	<.0001	1.4705	<.0001	***1.2452***	<.0001	1.238	<.0001
Number of Diagnoses		1.076	<.0001	1.0798	<.0001	1.04 22	<.0001	1.0418	<.0001
Number of Procedures		1.0776	<.0001	1.0746	<.0001	1.2453	<.0001	1.2467	<.0001
Breast Cancer		***0.9911***	0.0008	***0.7832***	<.0001	1.1903	<.0001	***1.2359***	<.0001
Hypertension	Independent	0.8676	<.0001	***0.8879***	<.0001	1.0243	<.0001	***1.0076***	<.0001
	Interaction			***0.9422***	<.0001	0.8661	<.0001	***0.8288***	<.0001
Diabetes	Independent	0.9579	<.0001	0.9333	<.0001	0.9868	<.0001	0.9975	0.0022
	Interaction	0.9895	0.0128	1.0554	0.0044	0.9655	<.0001		
Mental Disorder	Independent	1.0447	<.0001	0.8931	<.0001	0.9682	<.0001	1.0032	<.0001
	Interaction	0.9381	<.0001	0.8986	<.0001	0.9545	<.0001	0.9695	0.0185
Obesity	Independent	0.9348	<.0001	0.9437	<.0001	1.1267	<.0001	1.0864	<.0001
	Interaction			0.955	0.096	0.9165	<.0001	0.9178	0.0002
R-square		***22.43%***		***22.55%***		***36.74%***		***36.64%***	

[1] Blank cells represent non-significant results. The same holds in the following tables.

Consistent with the results of the descriptive statistical analysis, breast cancer patients have shorter LOS but higher total charges. When breast cancer is considered only at the primary diagnosis (or treatment) level, there is a decrease in the average LOS of 21.68% (i.e., 100%-78.32%). In contrast, there is only a 0.89% decrease for the prevalence breast cancer patients suggesting that the prevalence breast cancer patients are more similar to the general population as seen in the descriptive analysis.

Considering the independent disease (main effect), for the primary breast cancer group, hypertension, mental disorder, and obesity reduced LOS but increased charges, while patients with diabetes have both lower LOS and charges. In contrast, in the prevalence breast cancer group, mental disorder has higher LOS but lower charges. This suggests that when mental disorder is the primary or the secondary reason for hospitalization compared to when it is a secondary condition (primary being breast cancer), patients stayed in the hospital longer. In general, while patients diagnosed with breast cancer (primary or prevalence) and comorbid conditions have a shorter LOS, they have higher charges compared to patients without these conditions. For example, patients with primary breast cancer and secondary hypertension have a 34.48% (i.e., 100% - (78.32%*88.79%*94.22%)) shorter stay but spend 3.2% (i.e., (123.59%*100.76%*82.88%) – 100%) more on average, assuming all other factors are fixed.

When considering the comorbid effect on primary breast cancer patients, LOS decreases by 16% for hypertension (i.e., 100%-88.79%*94.22%), 1.5% for diabetes, 19.7% for mental disorder, and 9.9% for obesity compared to each corresponding non-comorbid

group. For total charges, hypertension decreases charges by 16.5%, 2.7% for mental disorder, and 0.3% for obesity, and there is no significant interaction for diabetes with breast cancer.

In addition, patient disposition, “died” or “transferred,” affected LOS and charges significantly. Consequently, when considering the effect of comorbid conditions on breast cancer patients, the effect of disposition also should be considered. Transferred patients generally have a longer LOS and higher total charges. For example, in the breast cancer prevalence group (Table 2.5), transferred patients have 45.34% longer LOS (assuming all other covariates are fixed) and a 24.52% increase in total charges compared to patients who are not transferred. Patients who died during hospitalization on average have a 3.62% shorter LOS with a 4.7% increase in total charges compared to those who did not die during hospitalization.

The results for comparing the weight-adjusted and unadjusted regression models for LOS for the primary breast cancer group are summarized in Table 2.6.

Table 2.6 Comparison for Weigh-Adjusted and Unadjusted Regression Models

	Weight-Adjusted		Unadjusted			Weight-Adjusted		Unadjusted	
Variable	Estimate	p-value	Estimate	p-value	Variable	Estimate	p-value	Estimate	p-value
Intercept	1.6988	<.0001	1.4384	<.0001	NDX	1.0760	<.0001	1.0797	<.0001
Age	1.0024	<.0001	1.0018	<.0001	NPR	1.0776	<.0001	1.0745	<.0001
Race1	0.9600	<.0001	0.9629	<.0001	Died	0.9638	<.0001	0.9987	<.0001
Race2	1.1408	<.0001	1.1292	<.0001	Transferred	1.4534	<.0001	1.4790	<.0001
Race3	1.0516	<.0001	1.0409	<.0001	BC[1]	0.9911	<.0001	0.9891	<.0001
Race4	1.0683	<.0001	1.0513	<.0001	H	0.8678	<.0001	0.8914	<.0001
Race5			1.0072	0.0902	D	0.9579	<.0001	0.9365	<.0001
Emergency	1.1034	<.0001	1.1097	<.0001	MD	1.0447	<.0001	1.0122	<.0001
Urgent	1.1100	0.0046	1.1086	<.0001	O	0.9348	<.0001	0.9460	<.0001
Elective	1.0161	<.0001	1.0119	<.0001	BC×H[2]				
Newborn	1.1075	0.0015	1.0929	0.0126	BC×D	0.9895	0.0128	0.8991	<.0001
Trauma Center	1.2371	<.0001	1.2598	<.0001	BC×MD	0.9381	<.0001	0.8871	<.0001
					BC×O				

[1] BC = breast cancer, H = hypertension, D = diabetes, MD = mental disorder, O = obesity

[2] BC×H (and same for the terms below) is the interaction term in the regression.

It can be seen from the results that although some of the parameter estimates for demographic variables like race are different, there is no change with respect to the comorbidity impact, and the significance levels are similar. These results suggest that while it is necessary to weight-adjust the NIS data to make national inferences based on the summary statistics, this is not necessary for the regression analysis as the sample is already well designed to represent disease impact. Hence, the rest of the studies are based on the sampled data only.

2.4.2.2 Understanding the Effect of Age on the Relationship between LOS (Total Charges) for Breast Cancer Patient with Comorbidities.

Different age groups have different distributions of LOS and total charges. Stratification of the population by age allows for a more detailed analysis. For simplicity, only the primary breast cancer group is considered for the age stratified regression analysis.

Tables 2.7 and 2.8 summarize the coefficients of the estimates for the sets of age-based regressions on LOS and total charges for the general population. For each comorbid condition, the first two columns have the coefficient and p-value for the independent disease, and the second two columns summarize for the coefficients and p-value for the interaction terms with breast cancer. R-squares are summarized in the last columns. It can be seen that in both regressions on LOS and total charges, R-square is higher, in general, for the older age groups (16.3% to 22.4% for LOS and 23.0% to 38.8% for total charges).

Table 2.7 Least Square Regression Results on **LOS** for General Population with Comorbidities by Age

	Died		Breast Cancer		Hypertension				Diabetes				Mental Disorder				Obesity				R^2 (%)
					Individual		Interaction		Individual		Interaction		Individual		Interaction		Individual		Interaction		
Age	Est. [1]	p [2]	Est.	p	Est.	P	Est.	p	Est.	p	Est.	p	Est.	p	Est.	p	Est.	p	Est.	p	
18 to 30	0.636	<.0001	***1.171***	<.0001	***0.979***	<.0001	***0.704***	0.019	1.016	0.001			0.800	<.0001	1.201	0.008	0.994				16.3
31 to 40	0.804	<.0001	0.907	<.0001	***0.902***	<.0001	***1.129***	0.053	0.970	<.0001			0.816	<.0001	1.168	0	0.948	<.0001	0.844	0.044	17.1
41 to 50	0.827	<.0001	0.784	<.0001	0.855	<.0001	1.129	0	0.933	<.0001	1.092	0.092	0.799	<.0001	1.155	<.0001	0.920	<.0001			18.9
51 to 60	0.941	<.0001	0.744	<.0001	***0.849***	<.0001	1.052	0.042	0.909	<.0001	1.065	0.07	0.857	<.0001			***0.911***	<.0001			20.1
61 to 70	***1.033***	<.0001	0.695	<.0001	0.870	<.0001			0.913	<.0001	1.094	0.003	0.919	<.0001	1.093	0.004	0.912	<.0001			21.4
71 to 80	***1.026***	<.0001	0.699	<.0001	0.887	<.0001	0.936	0.012	0.933	<.0001	1.078	0.015	0.966	<.0001			0.922	<.0001			22.4
81 to 90	0.949	<.0001	0.666	<.0001	0.907	<.0001			0.954	<.0001			***0.962***	<.0001	***1.088***	0.041	0.938	<.0001	0.841	0.095	21.4
Over 90	0.833	<.0001	0.660	<.0001	0.926	<.0001			***0.977***	<.0001	***1.249***	0.052	0.952	<.0001					***2.275***	0.045	18.8

[1] Est. = Parameter Estimate; [2] p = p-value

Table 2.8 Least Square Regression Results on Total Charges for General Population with Comorbidities by Age

	Died		Breast Cancer		Hypertension				Diabetes				Mental Disorder				Obesity				R^2 (%)
					Individual		Interaction		Individual		Interaction		Individual		Interaction		Individual		Interaction		
Age	Est. [1]	p [2]	Est.	p	Est.	p	Est.	p	Est.	p	Est.	p	Est.	p	Est.	p	Est.	p	Est.	p	
18 to 30	1.842	<.0001	***2.212***	<.0001	1.086	<.0001	0.650	0.009	1.072	<.0001			1.007	0.001	0.876	0.092	1.065	<.0001	0.744	0.066	23.0
31 to 40	1.520	<.0001	1.740	<.0001	1.06	<.0001	0.880	0.06	1.052	<.0001			1.100	<.0001	0.880	0.008	1.077	<.0001			25.8
41 to 50	1.162	<.0001	1.102	<.0001	***0.974***	<.0001	***1.078***	0.014					0.990	<.0001			1.011	0			35.2
51 to 60	1.145	<.0001	0.916	<.0001	0.949	<.0001			0.953	<.0001	1.057	0.084	0.950	<.0001			0.992	0.001			38.5
61 to 70	1.105	<.0001	0.848	<.0001	0.952	<.0001			0.943	<.0001	1.061	0.041	0.932	<.0001	1.065	0.036	0.990	0			38.8
71 to 80	1.074	<.0001	0.799	<.0001	0.955	<.0001			0.948	<.0001	1.065	0.036	0.915	<.0001	1.086	0.012	0.981	<.0001			36.7
81 to 90	1.011	0.003	0.895	<.0001	0.960	<.0001			0.964	<.0001	1.083	0.07	0.898	<.0001			0.959	<.0001			36.7
Over 90	***0.941***	<.0001			0.973	<.0001			0.981	0			0.923	<.0001			0.961	0.085			32.0

[1] Est. = Parameter Estimate; [2] p = p-value

Patients younger than 30 and older than 90 behave very differently than patients in other age groups. There are also a smaller number of samples in these two age groups. For patients younger than 50, the total charges are higher when the patients have breast cancer. This is especially true for patients under 30; for these breast cancer patients their total charges are about twice (221.2%-100%) as high as non-breast cancer patients and they also stay in the hospital for a 17.1% (i.e., 117.1%-100%) longer time. This difference decreases as the patient group gets older. In fact, for older patients the LOS and total charges are lower in the breast cancer group. For those older than 90, there is no significant difference in the total charges for the two cohorts.

For the general population, those patients with hypertension (independent or main effect) have a shorter LOS. The patient group aged 51 to 60 has the greatest difference with a 15.1% (i.e., 100%-84.9%) shorter time compared to the non-hypertensive patient. Total charges for patients having hypertension are higher for patients younger than 40 but lower for those older than 40. From Table 2.7 and 2.8, the coefficient estimate for hypertension follows a convex function for both regression models. That is, for patients between 50 and 70, the difference in LOS and charges for hypertension patients and non-hypertension patients are the largest.

For the primary breast cancer patients, having hypertension resulted in lower LOS and charges. However, for older patients, the differences in admission time and charges follow similar patterns to the general population. In contrast, for patients younger than 30, the difference is large; there is an approximately 31% (100%-97.9%*70.4%) decrease in LOS and 29.4% (100%-108.6%*65%) decrease in total charges. There are some exceptions:

For primary breast cancer patients age 31 to 40, the admission time increases about 2% for those patients with hypertension; and for patient ages 41 to 50, the total charges increase by 5% for those patients with hypertension as a comorbid condition to breast cancer.

In the general population, the effect of diabetes is similar to hypertension. The coefficient also follows a convex function, i.e., patients age 51 to 60 have the largest difference in LOS and patients age 61 to 70 have the largest difference in charges comparing those with diabetes to those without diabetes. However for primary breast cancer patients, there is almost no difference in LOS and charges for patients with diabetes, especially for patients between 41 and 80 (where the interaction terms adjusted (eliminated) the effect). For primary breast cancer patients older than 90, LOS increases by 22% (i.e., 97.7%*124.9% - 100%).

For patients with a mental disorder, the effects on LOS and charges have different patterns. LOS is shorter for patients with a mental disorder, but the magnitude of the difference decreases as the age group gets older. In contrast, charges for patients with a mental disorder in younger groups are higher than the non-mental disorder group; however, the magnitude of the difference decreases with age, and the charges are lower in older patient groups. For primary breast cancer patients, having a mental disorder also resulted in shorter admission time and lower total charges, except for patients age 81 to 90, for whom the admission time increased by about 5%.

Obesity patients have lower LOS and the magnitude of the difference also follows a convex function with the difference being largest for patients 50-70 years old with an almost 9% decrease. Charges for obese patients compared to non-obese ones are higher for younger

patients and lower for older patients. And the difference consistently decreases with age group. There is no significant difference in LOS for patients older than 90. Few interaction terms are significant for breast cancer patients who are obese. Notice that only for patients over 90, having obesity almost doubled LOS (i.e., 227.5% - 100%) for the primary breast cancer group.

The regression results also suggest that the disposition variable, DIED, has a significant impact on both LOS and total charges. The coefficients of estimates on LOS have a concave shape. For younger and older patients, mortality decreased the LOS. But for patient age 61 to 80, mortality increased the LOS. Total charges for patients who died during hospitalization are higher than for patients who survived, except for patients older than 90. The coefficients for total charges consistently decrease with patient age.

2.4.3 Logistic Regression on Patient Disposition: Characterizing the Probability of Mortality (or Transferring) During Hospitalization as a Function of Patient Background Variables and Comorbidities

The above analyses suggest that patient disposition has a significant effect on LOS and total charges for patients with comorbidities. Two separate logistic regression models are developed for patient mortality and transferring. The effect on the general population and the primary breast cancer group is compared. Since most of the interaction terms of breast cancer and comorbidities for the general population are not significant, the interaction terms in these regression models are not included. The odds ratios and p-values are summarized in Table 2.9 and Table 2.10 for patients who died or transferred. The odds ratio is a measure to

describe if an event is more or less likely to occur when the binary variable equals one, holding all other factors fixed [57]. For example, a value of 8.171 means having breast cancer increases the likelihood of death by 8.171 times compared to the patients without breast cancer.

Table 2.9 Logistic Regression Results for Patients who Died

	Breast Cancer		Hypertension		Diabetes		Mental Disorder		Obesity		R^2
Age	O.R. [1]	p	O.R.	p	O.R.	p	O.R.	p	O.R.	p	
18 to 30	***8.171***	<.0001					0.501	<.0001	0.474	<.0001	27.34%
31 to 40	***8.227***	<.0001	0.621	<.0001	0.718	<.0001	0.677	<.0001	0.606	<.0001	21.44%
41 to 50	***3.208***	<.0001	0.405	<.0001	0.664	<.0001	0.572	<.0001	0.426	<.0001	18.07%
51 to 60	1.957	<.0001	0.379	<.0001	0.627	<.0001	0.59	<.0001	0.411	<.0001	16.06%
61 to 70	1.380	0.0059	0.407	<.0001	0.659	<.0001	0.617	<.0001	0.422	<.0001	13.79%
71 to 80			0.441	<.0001	0.735	<.0001	0.713	<.0001	0.468	<.0001	11.05%
81 to 90	0.756	0.0485	0.481	<.0001	0.817	<.0001	0.788	<.0001	0.543	<.0001	7.43%
Over 90			0.569	<.0001	0.879	<.0001	0.838	<.0001	0.573	0.0002	5.05%

	Primary Breast Cancer Group										R^2
18 to 30											93.27%
31 to 40							0.05	0.0058			44.32%
41 to 50			0.212	0.0012			0.291	0.0010			33.15%
51 to 60			0.307	0.0001	0.404	0.0479	0.427	0.0067	0.126	0.044	27.53%
61 to 70			0.337	<.0001	0.262	0.0021			0.121	0.0404	25.42%
71 to 80			0.321	<.0001			0.476	0.0486			21.24%
81 to 90			0.492	0.0191							20.04%
Over 90											50.92%

[1] O.R. = odds ratio

Table 2.10 Logistic Regression Results for Patients who Transferred

	Breast Cancer		Hypertension		Diabetes		Mental Disorder		Obesity		R^2
Age	O.R.	p	O.R.	p	O.R.	p	O.R.	p	O.R.	p	
18 to 30			0.888	<.0001			***1.699***	<.0001	0.920	0.009	16.29%
31 to 40	0.537	0.0119	0.883	<.0001	***1.134***	<.0001	***1.466***	<.0001	0.863	<.0001	14.32%
41 to 50	0.372	<.0001	0.848	<.0001	***1.090***	<.0001	***1.148***	<.0001	0.897	<.0001	9.86%
51 to 60	0.368	<.0001	0.856	<.0001	***1.113***	<.0001	***1.197***	<.0001	0.978	0.0376	9.53%
61 to 70	0.334	<.0001	0.883	<.0001	***1.138***	<.0001	***1.397***	<.0001	1.042	<.0001	10.33%
71 to 80	0.373	<.0001	0.886	<.0001	***1.144***	<.0001	***1.834***	<.0001	1.024	0.0246	11.86%
81 to 90	0.447	<.0001	0.895	<.0001	***1.109***	<.0001	***2.087***	<.0001			11.97%
Over 90	0.509	<.0001	0.940	<.0001	***1.098***	<.0001	***1.922***	<.0001	1.134	0.0546	7.88%

Primary Breast Cancer Group

	Breast Cancer	Hypertension		Diabetes		Mental Disorder		Obesity	R^2
18 to 30									72.50%
31 to 40									26.41%
41 to 50									22.23%
51 to 60		0.452	0.0003	***2.378***	0.0001	***1.642***	0.0122		25.39%
61 to 70		0.508	<.0001						21.21%
71 to 80		0.731	0.0226			***1.904***	<.0001		25.81%
81 to 90		0.621	0.0005			***2.405***	<.0001		22.62%
Over 90						***2.201***	0.0205		16.55%

The last column in Tables 2.9 and 2.10 summarizes the Max-rescaled R-squares for each logistic regression. In general, the Hosmer-Lemeshow goodness-of-fit test is significant for the general population, primarily due to the large size of the dataset. This is a common phenomenon in the literature for the NIS data [27]. The test is not significant for the breast cancer group, which indicates a good fit. While a unit increase in total charges does not affect the mortality and transferring probabilities, a unit increase in LOS results in a lower mortality probability for the general population (odds ratio < 1) but a higher mortality probability for

the breast cancer group (odds ratio > 1). The odds ratios associated with LOS for patients who transferred are greater than 1 for both the general and breast cancer groups.

As shown in Table 2.9, patients with breast cancer have a higher probability of death for patients younger than 70 while older patients have lower probability of death compared to non-breast cancer patients. The probability of death for patients with comorbid conditions is lower for both the general population and the primary breast cancer subgroup. However, for the breast cancer subgroup shown, comorbidities generally do not significantly affect mortality for patients younger than 40 and older than 80.

As shown in Table 2.10, patients with either diabetes or mental disorder have a higher probability of being transferred. This is particularly true for primary breast cancer patients older than 50 where mental disorder increased the probability of being transferred. Perhaps this finding is related to these chronic diseases impacting breast cancer since both are long-term and often are in the later stages when diagnosed.

From the logistic regression results it can be seen that comorbidities affect patient disposition, especially mental disorder and diabetes for older patients. These comorbidities increase the chance of being transferred and thus actual LOS and total charges cannot be observed.

2.4.4 Survival Analysis on LOS: Incorporating Censoring on the Relationship Between LOS and Patient Background Characteristics and Comorbidities

For patients who do not die during hospitalization, i.e., the “survived” group, about 17.09% of the records are censored. These patients either stayed in the hospital for longer

than 365 days, or they were transferred to another facility. In the ordinary regression models, it is not possible to fully explore the effect of comorbidities on LOS and total charges for such patients because the LOS variable is biased as patients are transferred to other facilities. Treating transferred patients as a form of censoring makes it possible to more accurately study the impact of the comorbidities on patient outcomes, since discharge does not have to imply improved patient outcome, e.g., incorporating the censoring effect allows for the representation of patients whose discharge was not due to their condition improving. In fact, the results suggest some different patterns compared to ordinary regression results.

2.4.4.1. Accelerated Failure Time Model: A Parametric Model

Weibull, log normal, and log logistic distributions are fitted in the parametric regression models. The results for comorbidities are summarized in Table 2.11. The estimate for each coefficient shows the change in log(admission time) with the condition present as compared to not present while holding other covariates fixed. All coefficients for independent diseases are significant as well as the interaction terms except for obesity, which is consistent with earlier analysis.

Table 2.11 Results for Accelerated Failure Time Model

		Weibull		log Normal		log Logistic	
		Estimate	p-value	Estimate	p-value	Estimate	p-value
Breast Cancer		-0.3929	<.0001	-0.3373	<.0001	-0.3582	<.0001
Hypertension	Independent	-0.1999	<.0001	-0.1782	<.0001	-0.1737	<.0001
	Interaction	-0.0857	<.0001	-0.0831	<.0001	-0.0998	<.0001
Diabetes	Independent	-0.0841	<.0001	-0.0809	<.0001	-0.0774	<.0001
	Interaction	0.0733	<.0001	0.0533	0.0013	0.0487	0.003
Mental Disorder	Independent	***-0.0806***	<.0001	-0.0482	<.0001	-0.0488	<.0001
	Interaction	***0.0976***	<.0001	0.0393	0.0049	0.0385	0.005
Obesity	Independent	-0.1109	<.0001	-0.1033	<.0001	-0.0956	<.0001
	Interaction						
scale		0.8287		0.7928		0.4516	
shape		1.2067					
Maximum Likelihood		-7873796		***-7125916***		-7126686	

It is difficult to determine which parametric distribution fits the model best without analyzing the distribution of the error terms. For this very large dataset, it is challenging to plot the actual distribution. Hence, the values of the maximum likelihood estimates under each distribution assumption are compared to identify the best model. The model with the largest maximum likelihood estimate has the best fit. The log normal, which assumes that the error terms follow a normal distribution, has the largest maximum likelihood estimate.

Consistent with earlier results, comorbidities decreased LOS for primary breast cancer patients in most cases; however, the results are sensitive to the distribution assumption. Mental disorder increased LOS in the Weibull model but not under the other two distribution assumptions. This could be explained by the fact that mental disorder carried the most censored information and thus it is most sensitive to different distribution assumptions.

2.4.2.2 Proportional Hazards Model: Semi-parametric Assumption

The results for the proportional hazards model are summarized in Table 2.12. The first column summarizes the estimates for the coefficients α, β, and γ as in equation (2.10). The second column is the p-value corresponding to each estimate. The third column is the hazard ratio summary, which is the exponential of the coefficient estimate. The hazard ratio represents the hazards of the event when the binary covariate is one or one unit increase in the continuous variable while keeping other covariates fixed. For example, the hazard ratio of β_1 (the first element in the vector β) shows the effect on LOS if breast cancer is present. If it is greater than (less than) 1, then the hazard is higher (lower) for breast cancer patients. In this study, hazard refers to the "hazard" of leaving the hospital instead of dying, i.e., the conditional probability of leaving the hospital given that the patient has not left the hospital prior to time t.

In order to compare the results with the ordinary regression results, another stepwise regression is run on LOS for the same population (survived group) as the proportional hazards model and the results are summarized in Table 2.12. In this model the hazards associated with Hispanic, and Asian or Pacific Islander race categories are not significantly different from race group 6: other. In the ordinary regression, being Hispanic increases LOS by 3.5% (i.e., 1.035-1) compared to "other." Most interestingly, admission type 3: "Elective" increases the hazard of leaving the hospital by 4.1% (i.e., 1.041-1). In other words, the elective group stayed in the hospital for a shorter period of time compared to the "other" admission type, while in the ordinary regression results elective admission type did not have a significantly different impact on LOS as compared to "other".

Table 2.12 Comparison of Results for the Proportional Hazards Model and the OLS Regression Model

		Proportional Hazards			OLS Regression	
		Estimate	p-value	Hazard Ratio	Estimate	p-value
Age		-0.0108	<.0001	0.989	1.0039	<.0001
Race	Caucasian	0.0357	<.0001	1.036	0.9603	<.0001
	African American	-0.1471	<.0001	0.863	1.1338	<.0001
	Hispanic				***1.0349***	<.0001
	Asian or Pacific Islander				***1.0421***	<.0001
	Native American	-0.0106	0.0841	0.989	1.0071	0.0952
Admission	Emergency	-0.1224	<.0001	0.885	1.1159	<.0001
Type	Urgent	-0.0668	<.0001	0.935	1.1005	<.0001
	Elective	0.0399	<.0001	***1.041***		
	Newborn				1.0850	0.0233
	Trauma Center	-0.4226	<.0001	0.655	1.3576	<.0001
Total number of diagnoses		-0.1034	<.0001	0.902	1.0868	<.0001
Total number of procedures		-0.0555	<.0001	0.946	1.0718	<.0001
Breast Cancer		0.399	<.0001	1.49	0.7509	<.0001
Hypertension	Independent	0.1861	<.0001	1.204	0.8699	<.0001
	Interaction	0.1228	<.0001	1.131	0.9348	<.0001
Diabetes	Independent	0.0757	<.0001	1.079	0.9313	<.0001
	Interaction	-0.0622	0.0037	0.94	1.0517	0.0009
Mental Disorder	Independent	0.0457	<.0001	***1.047***	***0.9145***	<.0001
	Interaction	-0.0744	<.0001	***0.928***	***1.0712***	<.0001
Obesity	Independent	0.1011	<.0001	1.106	0.9303	<.0001
	Interaction					

The effect of individual disease on LOS is similar in the two models. Each independent disease increased the hazard of leaving the hospital and thus there was a shorter LOS. Breast cancer increased the hazard by 49% (i.e., 1.49-1); 20% for hypertension, 7.9% for diabetes, 4.7% for mental disorder, and 10.6% for obesity. As in the ordinary regression case, the LOS decreased by 25% for breast cancer patients, 13% for hypertension, 6.9% for diabetes, 8.6% for mental disorder, and 7% for obesity.

When looking at comorbidities, there is some difference between the two models. Coefficients for the interaction terms of breast cancer with obesity are not significant in

either model. Thus only the other three conditions are compared. Having hypertension for primary breast cancer patients increased the hazard of leaving hospital by 36% (i.e., 1.204*1.131-1) and in the OLS regression, LOS was shortened by 18.7% (i.e., 1-0.867*0.935). So these two results correspond. Similarly for diabetes, the survival model results in a 1.4% increase in hazard and ordinary regression has an approximately 2% decrease in LOS. However, for patients with a mental disorder the story is a different. In the proportional hazards model, for primary breast cancer patients comorbid with mental disorder the hazard decreased by 2.8% (i.e., 1-1.047*0.928). This implies a breast cancer patient with mental disorders stayed in the hospital for a longer time than a breast cancer patient without the comorbid condition. In contrast, in the ordinary regression models, the LOS associated with breast cancer patients with mental disorder decreased by 2% (i.e., 1-0.9145*1.0712). The logistic regression analysis for patients who transferred indicates that mental disorders increased the probability of being transferred. So it is reasonable that by considering censoring in the survival models, the results may be different from ordinary regression models. By incorporating the information from censoring, the survival models tell the story more completely.

Survival analysis has shown that censoring matters in the study of patient LOS, particularly for patients with a mental disorder who have a higher probability of being transferred, and thus have more censored information on LOS. Survival models help to study the true LOS by modeling the effect of the unobserved information. The above analyses have indeed shown that mental disorders have the greatest impact on breast cancer patients relative

to the other chronic disease studies, suggesting a closer study of the relationship between mental disorders and breast cancer is necessary.

2.4.5 Principal Component Analysis and Cluster Analysis

The 30 ICD-9 codes for mental disorders are decomposed into 30 variables, and principal component analysis is conducted on these variables. Variance structure is used because these variables are indicator variables and do not have scaling problems. The scree plot shown in Figure 2.4 indicates that the turning point is around component 7. After that, all remaining components have very small and close eigenvalues. As summarized in Table 2.13, the first seven principal components explained more than 90% of the total variance, and thus only these are used to explore the mental disorder codes related to breast cancer patients.

The seven variance proportions are used as weights for each variable, and the total variance for each variable is the eigenvector (the first 7 principal components) multiplied by the variance proportion as weights. The highest-weighted total variances (score > 0.01) are summarized in Table 2.14.

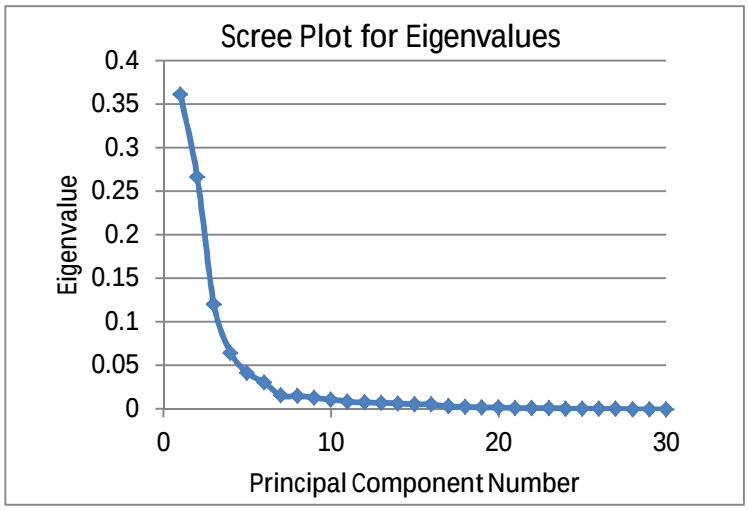


Figure 2.4 Scree Plot for Eigenvalues of PCA

Table 2.13 Summary of variance by principal components

	Total Variance Explained By the Components		
	Eigenvalue	% of Variance	% Cumulative
1	0.3114	36.15%	36.15%
2	0.2298	26.67%	62.82%
3	0.1034	12.01%	74.83%
4	0.0559	6.49%	81.32%
5	0.0362	4.21%	85.52%
6	0.0263	3.05%	88.57%
7	0.0137	1.59%	90.16%

Table 2.14 PCA rank for ICD-9 codes on mental disorder

ICD-9	Score	Description for ICD-9 Codes
311	0.4146	Depressive disorder not elsewhere classified
300	0.4015	Anxiety, dissociative and somatoform disorders
305	0.3901	Nondependent abuse of drugs
294	0.1518	Persistent mental disorders due to conditions classified
296	0.1100	Episodic mood disorders
295	0.0685	Schizophrenic disorders
290	0.0374	Dementias
309	0.0238	Adjustment reaction
293	0.0125	Transient mental disorders due to conditions classified
303	0.0123	Alcohol dependence syndrome
319	0.0119	Unspecified mental retardation

The PCA identified depressive disorders (ICD-9 code:311), anxiety (300) and nondependent abuse of drugs (311) as the mental disorders most related to breast cancer patients. Other related mental conditions include persistent mental disorders, episodic mood disorders and dementias. These are also known clinically to be assocaited with breast cancer [59].

The next question is to identify the maximum number of variables that can be deleted while leaving the relationship unchanged and extracting the most variance. The CA identified the most related variables (Table 2.15). There are five clusters with the corresponding ICD-9 codes shown in column 2. Based on the (1-R-square) ratio in column 3, the most important variables are highlighted in bold.

Within each cluster, the variables are correlated with each other. For example, nondependent abuse of drugs is correlated with alcohol-induced mental disorders. Consistent with the PCA result, the five variables selected from each cluster have the highest PCA score and are the most related for primary breast cancer patients.

Table 2.15 Summary for cluster analysis on mental disorder ICD codes

Cluster	ICD-9	R-squared with Own Cluster	R-squared with Next Cluster	1-R^2 Ratio	Codes Explanation
1	291	0.0008	0.0002	0.999	Alcohol-induced mental disorders
	305	1	0.1207	0	**Nondependent abuse of drugs**
	307	0.0033	0.0009	0.998	Special symptoms or syndromes not elsewhere classified
	312	0.0005	0.0002	1	Disturbance of conduct not elsewhere classified
2	290	0.0029	0.009	1.006	Dementias
	292	0.0024	0.0011	0.999	Drug-induced mental disorders
	293	0.0027	0.004	1.001	Transient mental disorders due to conditions classified
	299	0.0001	0.0002	1	Pervasive developmental disorders
	303	0.0006	0.0018	1.001	Alcohol dependence syndrome
	308	0.0002	0.0005	1	Acute reaction to stress
	309	0.0023	0.0052	1.003	Adjustment reaction
	311	1	0.1215	0	**Depressive disorder not elsewhere classified**
	315	0.0005	0.001	1.001	Specific delays in development
	318	0.0005	0.001	1.001	Other specified mental retardation
	319	0.0018	0.0047	1.003	Unspecified mental retardation
3	300	1	0.116	0	**Anxiety, dissociative and somatoform disorders**
	306	0.0007	0.0041	1.004	Physiological malfunction arising from mental factors
4	294	0.9999	0.0421	1E-04	**Persistent mental disorders due to conditions classified**
	297	0.0011	0.0006	1	Delusional disorders
	298	0.0047	0.0027	0.998	Other nonorganic psychoses
	310	0.0042	0.001	0.997	Specific nonpsychotic mental disorders due to brain damage
5	295	0.025	0.0062	0.981	Schizophrenic disorders
	296	0.9908	0.012	0.009	**Episodic mood disorders**
	301	0.011	0.0002	0.989	Personality disorders
	304	0.0028	0.0017	0.999	Drug dependence
	314	0.0012	0.0006	0.999	Hyperkinetic syndrome of childhood
	317	0.0021	0.0013	0.999	Mild mental retardation

To compare the original model with 30 ICD-9 codes used for mental disorders, a similar OLS regression model is run for the same population (primary breast cancer group) using only the ICD-9 codes identified by the CA, and the exact same results are achieved.

Thus, only these five ICD-9 codes (305, 311, 300, 294, 296) need to be included for future analyses on the relationship between mental disorders and breast cancer.

2.5 Discussion

Several statistical models are used to analyze the impact of comorbidities on breast cancer patient outcomes, including patient length of stay, total charges, and disposition. Descriptive statistical analysis has provided a general picture of the characteristics of the study population. It also has shown that comorbidities have significant impact on breast cancer patient outcomes.

Ordinary least square regression models provide details on these impacts adjusting for other effects. It has been shown that patients with hypertension generally have shorter lengths of stay with lower total charges. It is hypothesized that this is because the disease is often well-managed, and patients receive routine treatment which has lower charges. Although diabetes also decreases length of stay and total charges for the general population, there is no significant impact on total charges for primary breast cancer patients. This can be seen clearly when the data is stratified by different age groups. Obesity has limited impact on length of stay and total charges, and the effect is similar in the general population and the breast cancer patient group.

Patient disposition (i.e. mortality and transferring) also has been shown to affect a patient's length of stay and total charges. From the logistic regression results on patient disposition, mental disorder increases the probability of being transferred for primary breast

cancer patients; thus the effects on length of stay and total charges from the regression results do not show the complete picture.

When survival models are incorporated with “transferred” as the censored information different results are achieved. In the proportional hazards model mental disorder indeed increases length of stay for breast cancer patients. It is important to incorporate censoring into the regression models since comorbidity (especially mental disorder) affects patient disposition and thus it also affects the true observed admission time.

As seen in the analysis, comorbid conditions will affect length of stay and total charges for breast cancer patients, and this can help inform the policymaker and facilitate predicting hospitalization patterns for breast cancer patients with comorbidities. A summary of the findings of the analyses is presented in Table 2.16.

The aging of the United States population in conjunction with the ability to detect cancers at an earlier stage due to improved imaging and breast cancer screening has led to an interesting dilemma for clinicians involved in the care of breast cancer patients [60-62]. At what point does a person’s mortality risk from breast cancer exceed their mortality risk from their comorbid conditions, and what is the cost both monetarily and in quality of life of treating that breast cancer. Increasingly it will be important for clinicians to evaluate the cumulative effects of multiple comorbid conditions on patient outcome. For example, conditions such as obesity are associated with hypertension and diabetes. The ability to identify patients who have a longer length of stay in the hospital, or who are at risk of a discharge to a facility other then home, are areas for which further study is needed in order to

identify the causes of the increased length of stay or barriers to discharge and thus perhaps intervene and change those outcomes.

Table 2.16 Summary of Impact of Comorbidities for Breast Cancer Patient Outcomes [1]

Outcome	Groups	Hypertension	Diabetes	Mental Disorder	Obesity
Length of Stay	Prevalence BC	↓[2]	↓	↓	↓
	Primary BC	↓	↓	↓	↓
	18 to 30	↓Δ[3]	–	↓	–
	31 to 40	↑	–	↓	↓
	41 to 50	↓	↑	↓	–
	51 to 60	↓	↓	–Δ	–
	61 to 70	–	↑	↑	–
	71 to 80	↓	↑	–	–
	81 to 90	–	–	↑	↓
	Over 90	–	↑Δ	–	↑Δ
	Survival Cox Model	↓	↓	↑	↓
Total Charges	Prevalence BC	↓	↓	↓	↑
	Primary BC	↓	↓	↓	↓
	18 to 30	↓Δ	–Δ	↓Δ	↓Δ
	31 to 40	↓	–	↓	–
	41 to 50	↑	–	–	–
	51 to 60	–	↑	–	–
	61 to 70	–	↑	↓	–
	71 to 80	–	↑	↓	–
	81 to 90	–	↑	–	–
	Over 90	–	–	–	–
Mortality	18 to 30	–	–	–	–
	31 to 40	–	–	↓Δ	–
	41 to 50	↓Δ	–	↓	–
	51 to 60	↓	↓	↓	↓
	61 to 70	↓	↓Δ	–	↓Δ
	71 to 80	↓	–	↓	–
	81 to 90	↓	–	–	–
	Over 90	–	–	–	–

Table 2.16 Continued

Outcome	Groups	Hypertension	Diabetes	Mental Disorder	Obesity
Transfer	18 to 30	−	−	−	−
	31 to 40	−	−	−	−
	41 to 50	−	−	−	−
	51 to 60	↓Δ	↑Δ	↑	−
	61 to 70	↓	−	−	−
	71 to 80	↓	−	↑	−
	81 to 90	↓	−	↑Δ	−
	Over 90	−	−	↑	−

[1] The table summarized the impact on outcomes for breast cancer patients with comorbidities, not for general population.
[2] ↓= decrease in outcome; ↑ = increase in outcome; − = same effect as general population.
[3] Δ represents the biggest difference in outcome among all age groups.

There are some limitations in this analysis based on the dataset. Due to the de-identification of the data, it is not possible to determine if there are duplicate records for the same patient (i.e., if a patient is admitted more than once within the year) in the dataset. Possible duplicates are checked by identifying zip code, income range, and location, but duplication could not be conclusively identified. So it is assumed in this analysis there are no duplicate records. The data de-identification also prohibits a more longitudinal analysis. There is also no information regarding pre-existing conditions. Length of stay and total charges may be affected by pre-existing conditions that could not be identified in the data.

Another limitation as discussed above is that patients with zero length of stay are assumed to stay in the hospital a half day. However, this could introduce a bias in the analysis if patients actually stay shorter or longer than half of a day. There are 12,156 observations with both zero length of stay and disposition of death with mean charges of

$13,802; and 18,828 observations with both zero length of stay and the disposition of "transferred" with mean charges of $6,972.

Finally, a naïve survival analysis is used in this study that separates death and hospital stay as two events for the survival analysis. This helps to explain the partial impact of comorbidities on breast cancer patient outcomes.

In future research, we will extend the survival analysis on length of stay to incorporate terminating events which will capture the impact of death on patient hospital stay. Moreover, we will extend the model to combine a modified Charlson comorbidity index on patient outcomes. In addition, we will identify specific comorbid conditions in finer detail to indicate their impact on patient outcomes. In particular, we would like to further decompose the coding for mental disorder. We believe this modeling approach has application for studying many diseases, comorbidities, and various patient outcome measures.

Chapter 3 Mortality Analysis for Breast Cancer Patients

3.1 Introduction

As introduced in Chapter 2, breast cancer is a disease in which malignant cancer cells form in the tissues of the breast. It is the most common form of noncutaneous cancer in American women and the leading cause of cancer death for females. This study of breast cancer mortality not only quantifies the risk associated with specific patient characteristics, it also informs the development of more personalized screening policies.

Mortality estimation for breast cancer is a significant research question in the literature. While there are many mortality studies focused on predicting trends in breast cancer mortality [63-65], and quantifying the impact of treatment on breast cancer mortality [66,67], this dissertation research predicts breast cancer mortality and mortality from other causes as a function of breast cancer risk factors, tumor characteristics, patient demographics and time. Schairer et al. [68] used Surveillance, Epidemiology, and End Results (SEER) data to calculate the 5-year and 10.9-year cumulative death probabilities by age, race, breast cancer stage and estrogen receptor status. The study presented in this chapter extends their study to include confidence interval approximation for the probability estimate and incorporates additional breast cancer risk factors including breast density and family history. Rosenberg [69] created cohort life tables using the Berkeley Mortality Database and National Center for Health Statistics data to remove other causes from breast cancer mortality. While this study demonstrated removing breast cancer from other cause mortality is worthwhile

because breast cancer mortality can be very high at some ages, it did not distinguish mortality by race and other factors. Lee and Zelen [70] used SEER and BCSC data to develop a stochastic model for predicting changes in breast cancer mortality and compared the mortality in two models for individuals with and without screening history. However, they did not present the estimate of mortality probabilities from other causes. The Wisconsin simulation model by Fryback et al. [26] estimated age-specific breast cancer mortality and mortality from other causes from 1975 to 2000, and adjusted 10-year survival probability to annualized probability by assigning a constant cure fraction. Mortality from other causes was derived using Rosenberg's method [69]. This dissertation study derives mortality probabilities using community-based data, and compares the mortality estimates associated with different risk groups.

In a recent study using the Kopparberg randomized controlled trial in Sweden, Chiu et al. [71] found that dense breast tissue is significantly related to breast cancer incidence and mortality for Swedish women aged 45 to 59 using the Cox proportional hazard model. The research in this chapter calculates the mortality probabilities with confidence intervals using a community-based mammography screening registry in the U.S., and adjusts for other effects including race, cancer stage, and age.

The focus of this study is to estimate the effect of three under-studied breast cancer risk factors and tumor characteristics on mortality risks for breast cancer patients using competing risks analysis. Specifically, it aims to quantify the impact of breast density, estrogen receptor (ER) and progesterone receptor (PR) status, and family history of breast cancer on breast cancer mortality and mortality from other causes. The impact of breast

density on outcomes for American women who participate in screening is a significant open research question. The American Cancer Society has listed dense breast tissue as a breast cancer risk factor that one cannot change [72, 73], and dense tissue is known to make it difficult to detect problems on a mammogram [74-77]. Despite this, there have been limited studies on the effect of breast density on mortality particularly in American women [71]. This chapter quantifies the impact of breast density on mortality controlling for other factors including age, race and cancer stage.

In addition, this study estimates the impact of ER and PR status on breast cancer mortality. ER and PR are tumor markers used to predict response to hormone therapy and chemotherapy [78, 79]. While the effect of ER status on breast cancer mortality has been explored in some research [63, 67], this research also includes PR status and studies the impact of different combinations of receptor status on mortality for breast cancer patients. Family history of breast cancer, particularly among first-degree relatives, is another well-demonstrated breast cancer risk factor [72, 80] for which there has also been limited study regarding the association of family history and mortality. Chang et al. [81] used the Cox proportional hazards regression model to conclude that family history of breast cancer is not associated with all-cause mortality. This dissertation research quantifies the death probabilities for comparing breast cancer patients with at least a first degree relative with breast cancer to those with no family history, and separates breast cancer death and death from other causes to quantify the effect on both.

Moreover, this research develops different models to address “left censoring” when the true start time of breast cancer cannot be fully observed. This is important for estimating

mortality probabilities more accurately, as typically a survival period is only calculated from cancer diagnosis date. This may potentially create bias in the estimate. As an outline of the remainder of the chapter, methods are discussed in Section 3.2. Section 3.3 presents the results from the mortality analysis, and Section 3.4 concludes the chapter with a discussion.

3.2 Methods

3.2.1 Data

This study uses the community-based Carolina Mammography Registry (CMR) data for the survival analysis. The NCI-funded CMR is part of a collaborative research effort designed to study screening mammography in community practice [82] and is a member site of the national Breast Cancer Surveillance Consortium (BCSC) [83]. CMR has been collecting prospective data on breast imaging performed in community-based mammography practices across North Carolina since 1994. All mammography records are linked with breast pathology data from a file created at CMR and from the North Carolina Central Cancer Registry. The CMR data is also linked to the North Carolina Death Tapes for mortality information. There are more than 2 million visit records on approximate 663,000 women, among whom more than 20,000 were diagnosed with breast cancer. The registry is reviewed annually by the University of North Carolina, Chapel Hill, School of Medicine IRB, and this study has been approved by the IRB.

The following de-identified information was collected from CMR: patient date of birth, race (only Caucasian and African American women were included in this study as they

have the largest representation in the database and there are significant disparities in the breast cancer mortality), cancer diagnosis date, cancer stage at diagnosis, vital status, death date, cause of death (ICD code 174 in the 9th version [84] and C50 in the 10th version [85] are used to identify breast cancer death), risk factors including breast density and family history of breast cancer, and tumor characteristics including estrogen and progesterone receptor status. Patient age is calculated based on the time difference between diagnosis date and birth date. The end of study time is assumed to be January 1, 2008 the last time the registry was linked to the death tapes prior to this study. The records with unknown or missing information are excluded from the analysis.

For breast density, the BI-RADS coding of "heterogeneously dense" and "extremely dense" are grouped into the dense group, and the fatty group is defined as those coded with "almost entirely fat" and "scattered fibro-glandular densities". In order to achieve adequate sample size and control for other effects, the estimation is separated first by age and race and then by age and cancer stage. Two cases are considered for the study of the role of receptor status, the first by race and ER/PR status and second by cancer stage and ER/PR status, and four combinations of ER and PR status are considered. To study the role of family history, mortality probabilities are compared between the two groups (with and without a family history) controlling for age and cancer stage. Family history (FH) is defined as breast cancer in any of the first-degree relatives including mother, sister(s) or daughter(s).

3.2.2 Cumulative Incidence Function

Deaths from causes other than breast cancer are modeled as competing risks for breast cancer patients. The nonparametric cumulative incidence function, which has been proven to be an unbiased estimator for mortality probabilities when competing risks are present [86-89], is used to calculate the death probabilities. The cause-specific cumulative incidence at a given time is computed as

$$F_r(t) = \sum_{j:\, j \le t} \frac{d_{rj}}{n_j} S_{KM}(t_{j-1}) \tag{3.1}$$

where d_{rj} is the number of patients who die from cause r at time j, $S_{KM}(t_{j\text{-}1})$ is the Kaplan-Meier Estimate of the overall survival at the last time point j-1, and n_j is the number of patients at risk at the beginning of time j,

$$n_j = n - \sum_{k=1}^{j} (d_k + c_k) \tag{3.2}$$

where n is the number of patients initially at risk, c_j is the number of patients right censored (death is not observed) at time t_j, and $d_k = \sum_{r=1}^{R} d_{rk}$ is the total number of deaths from all causes that have occurred.

3.2.3 Confidence Interval Estimation

To calculate the variation of the estimate, Marubini and Valsecchi's variance estimator [90] is used. It is derived using the delta method as the mortality probability is estimated using a non-parametric method.

$$Var(\hat{F}_r(t_j)) = \sum_{k=1}^{j} \left\{ \left[\hat{F}_r(t_j) - \hat{F}_r(t_k) \right]^2 \frac{d_k}{n_{k-1}(n_{k-1} - d_k)} \right\} + \sum_{k=1}^{j} S_{KM}{}^2 \frac{(n_{k-1} - d_{rk})}{n_{k-1}} \cdot \frac{d_{rk}}{n_{k-1}^2} - 2\sum_{k=1}^{j} \left[\hat{F}_r(t_j) - \hat{F}_r(t_k) \right] \cdot S_{KM} \cdot \frac{d_{rk}}{n_{k-1}^2} \quad (3.3)$$

Then *ln*(*-ln*) transformed bounds are used to calculate the 95% confidence interval (CI) [91] associated with each estimate:

$$\hat{F}_r(t)^{\exp\left\{ \frac{\pm c_{a/2} \sqrt{Var_r(t)}}{\hat{F}_r(t) \ln(\hat{F}_r(t))} \right\}} \quad (3.4)$$

where $c_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. It has been shown that the *ln*(*-ln*) transformation provides a better coverage probability and is better for smaller sample sizes [91] compared to other forms of transformation.

In contrast to a proportional hazards model or an accelerated failure time model, which study the effects of covariates on the hazard rate, the cumulative incidence quantifies mortality probabilities over time and the estimation of confidence intervals for probabilities can be compared among different risk groups.

3.2.4 Left Censoring – Method 1: Using Mammography Screening Information

The model presented in Section 3.2.3 only considers right censoring, which occurs when the end of the event time (mortality in this study) is unknown. Specifically two situations are considered: when a patient leaves the study cohort before the end of the study period and no further information about her can be observed; or when the event (death) is not observed by the end of the study period. Right censoring is commonly incorporated in the survival analysis.

For this analysis, another important form of censoring is considered, left censoring. Left censoring is an indication of whether the event start time is observable or unknown. Since community-based screening registry data is used in this study, the breast cancer onset time is not observable, and only the cancer diagnosis time is known. This may affect the true probability of death for breast cancer patients when the survival period is considered to start at the cancer onset time.

Another related but different concept in survival analysis, which considers this type of unobservable property, is “lead time bias”. It is used to measure the time interval between diagnosis of asymptomatic cancer and the point when the disease surfaces clinically. While lead time bias focuses on the detectability of cancer, left censoring models the actual onset of the disease for a more accurate mortality estimate. Figure 3.1 illustrates and compares these three concepts.

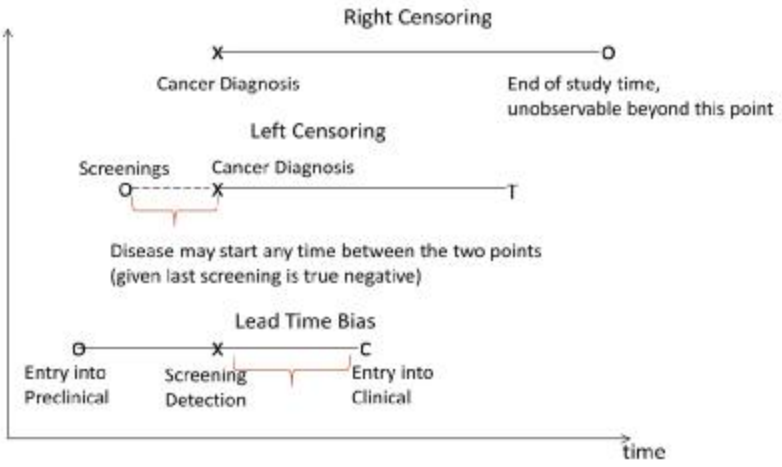


Figure 3.1 Representation of right censoring, left censoring and lead time bias

Two simulation models are developed for quantifying the left censoring time. The first algorithm is discussed below. There are two components for this model: assigning the left censoring status and estimating the censoring time. The second method is presented in Section 3.2.5.

1. Algorithm for assigning left censoring status

Step 1: Check the first record of screening for each patient and determine if the screening date is after the cancer diagnosis date but within six months. If the first screening date is more than six months after the cancer diagnosis, then the records are eliminated since this screening is assumed to be unrelated to the diagnosis. The screening may be a follow-up test after the cancer is found. Identify the patients that satisfy this requirement and proceed to step 2. For those patients with screening date before the diagnosis date, go to step 4.

Step 2: Check if there is a previous screening for this patient using a variable that records a previous screening date in the CMR data. (Note: If there is no previous screening

record for the patient in the CMR data, she may have screening elsewhere.) Then check the variable where the physician assigns reasons for visit: asymptomatic or symptomatic. For the left censoring criteria, go to step 3.

Step 3: For a symptomatic visit, if there is no previous screening history or the last screening was more than 6 months ago, then the record is considered to be left censored. If the last screening was less than 6 months ago, then no censoring is considered; i.e., the diagnosis date is considered as the true start time for cancer.

For an asymptomatic visit, if there is no previous screening history or the last screening was more than 12 months ago, then the record is considered to be left censored. If the last screening was less than 12 months ago, no censoring is considered.

Step 4: If diagnosis date is not related to the first screening in the registry, then locate the screening that corresponds to the diagnosis (excluding diagnostic screenings and continuing follow-up screenings). Check the physician assignment regarding the reason for the visit and go to step 5 to determine the censoring status.

Step 5: For a symptomatic visit, if the last screening is more than 6 months ago, then the record is considered as left censored. Otherwise, no censoring is considered. For an asymptomatic visit, if the last screening was more than 12 months ago, then the record is considered to be left censored. Otherwise, no censoring is considered.

The above assignment is based on the criteria that a patient with a clinical stage of breast cancer (symptomatic) will have a higher probability of being left censored, thus a shorter period of time is allowed to check the previous screening in order to identify the left censoring cases. Patients in a preclinical stage (asymptomatic) may have a cancer onset time

closer to the screening date, and thus a longer period of time is allowed for identifying the left censoring cases.

2. Estimating censoring time

After the left censoring status is assigned, then the censoring time is estimated and is added to the total event time for final mortality analysis. In this study, screening is assumed to be perfect, and this is reasonable as mortality probabilities are only estimated for those with confirmed cancer diagnosis. It is also assumed that cancer may start at any point in the censoring interval. The interval is calculated based on two quantities: (1) the time difference between the diagnosis date and the previous screening, and (2) the time difference between the cancer diagnosis date and the diagnosis date for the last patient in the noncensored group, with data ordered by patient age. For (1), the maximum period allowed is 36 months. If the difference calculated is greater than 36, then the interval is assigned to be 36 months. This is based on the assumption that breast cancer will be clinically detected within 3 years. The final censoring interval is estimated by the average of the two time intervals, to reduce possible over or under estimation.

For the mortality analysis, the event time is modeled as:

$$T = T_0 + C \tag{3.5}$$

where T_0 is the original event time, and C is the left censoring time.

The cancer is assumed to occur according to a uniform distribution on the time interval calculated above. One thousand replications are run for each interval to determine the left censoring period. The mean, minimum and maximum of the simulated time are compared. Results are discussed in Section 3.3.4.

3.2.5 Left Censoring – Method 2: Simulation Using Tumor Size Information

In the second simulation model for left censoring time, information on the tumor size in the CMR data is used, and breast cancer staging is also defined according to the tumor size [2]. There are many tumor growth models in the literature [92], and these tumor growth models assume different distributions for tumor growth [93, 94]. In this simulation model for left censoring, for simplicity an exponential tumor growth model is assumed, which is a variation of the Gompertz model in the literature [93], and it is in the following form:

$$TS(t) = TS(t_0) \cdot \exp(b \cdot (t - t_0)) \tag{3.6}$$

where $TS(t)$ is the tumor size at time t and b is unknown growth parameter. The heuristic for quantifying the left censoring time has the following steps.

Step 1: A patient is assumed to die from breast cancer when tumor size reaches 150mm in diameter [93]. The cancer diagnosis time t_0 is known from the data. Then for each record where death occurred, parameter b has the following inverse relationship:

$$b_i = 1/(t - t_0) \cdot \ln(150/TS(t_0)) \tag{3.7}$$

where t is the recorded death time. Tumor size at diagnosis is also identified from the CMR data.

Step 2: After collecting the data for the b_is, the Arena® Input Analyzer [95] is used to fit the distribution for the growth rate b. The best fit is then selected from the fitting result based on the smallest squared error.

Step 3: Input modeling is also used to fit the distribution for tumor sizes. Then random samples are generated from the distribution for those records with missing tumor

size. “Missing at random” is assumed in this study since no specific missing pattern could be identified, thus the same distribution could be applied to those missing data.

Step 4: A similar relationship to that shown in equation (3.6) is derived as the following:

$$t_{CS} = t_0 - (1/\hat{b}) \cdot \ln(TS(t_0)/1) \qquad (3.8)$$

where $\hat{b}$ is the fitted growth rate parameter from step 2, and t_{CS} is the true cancer onset time. In this study, it is assumed that cancer starts at a tumor size of 1 mm in diameter [94].

Step 5: After the starting point of left censoring is determined, the interval $t_0 - t_{CS}$ is then added to the total survival time, and then the mortality probabilities are adjusted.

3.3 Results

Confidence intervals are compared for the associated risk groups. The overlap in confidence intervals means there is insufficient evidence to say there is a statistically significant difference in mortality probabilities for comparison at the 5% level of significance; while no overlap means there is a statistically significant difference at the 5% level. The curves for the cumulative mortality probabilities for each comparison group are presented in Figures 3.2 to 3.6. For each graph, the estimated probabilities over time for breast cancer (pink) and other causes (black) death are plotted with solid lines while confidence intervals are plotted with the dashed lines. The counts for breast cancer patients in each group and the number of deaths are summarized in Tables 3.1 through 3.5. The estimates for mortality probabilities with the associated 95% CI at 5 years, 10 years and 13

years are also summarized in these tables. The mortality probabilities are plotted over a 160-month (about 13 years) horizon unless otherwise noted. The results for each of the three selected risk factors are discussed in Section 3.3.1 through Section 3.3.3. Abbreviations used in the following discussion are: BC for breast cancer, OC for other cause(s). The plus (+) and minus (-) signs are used to indicate the positive and negative groups, respectively.

3.3.1 Breast Density

There are a total of 15,243 BC patients in the registry with known age, race (Caucasian or African American) and breast density, 1,099 of whom died from breast cancer and 1,407 from other causes by the end of the study time (Table 3.1). From the table it can also be seen that women younger than 60 years old have more cases with dense breast compared to fatty, and for women older than 60 the opposite is true. This is consistent with the established knowledge [74].

Estimated mortality probabilities with confidence intervals for studying the role of breast cancer are plotted in Figure 3.2. The y-axis is the estimate of cumulative mortality probabilities, and x-axis is the time since diagnosis (in month). It is shown that the effect of age and race are consistent with earlier mortality estimates [33, 68, 72] that did not consider density. Specifically, African American women have a statistically higher probability of dying from BC than Caucasian women particularly for women diagnosed between the ages of 40 and 59 (for all three mortality estimates at 5, 10, and 13 years). The CI for 10-year mortality for African American women with dense breast tissue at age 40-49 (subplot 3.2-7) is (0.111, 0.202) compared to (0.049, 0.080) for Caucasian women (Figure 3.2-5). The

impact of breast density is most apparent for women age 70-79, for whom there is a significant difference in the BC mortality between African American and Caucasian women with fatty breast tissue (e.g., a 10-year mortality probability of (0.112, 0.223) in Figure 3.2-20 compared to (0.065, 0.097) in Figure 3.2-18), but not for women with dense breast tissue (e.g., (0.095, 0.243) in Figure 3.2-19 compared to (0.053, 0.098) in Figure 3.2-17). For the other age groups, although the mortality probability point estimate is always higher for African American women, no significant differences can be identified when comparing the 95% CIs. While BC mortality probabilities do not indicate significant differences among different age groups, younger patients are more likely to die from BC while older patients have significantly higher probabilities of dying from OC, regardless of breast density. Note this is only true for the Caucasian population. For African American women, from 60 years old there is not enough evidence to indicate significant differences in the mortality probabilities between BC and OC, and the two probability lines cross from time 0.

Interestingly, within the same racial and age groups, the BC mortality CIs overlap between the two density groups (e.g., (0.056, 0.090) in Figure 3.2-13 vs. (0.054, 0.083) in Figure 3.2-14 for 60-69 age group in Caucasian patients). The CIs have similar patterns with respect to the BC and OC mortality curves for the same racial and age groups, except for African American women over 80, where the 10-year mortality 95% CI for BC is (0.091, 0.349) and (0.213, 0.460) for OC in the dense group (Figure 3.2-23), while in the fatty group the 95% CIs are (0.078, 0.238) and (0.407, 0.594) respectively, which suggests a significant difference.

There are a total of 11,194 BC cases with known age, density and cancer stage at diagnosis, 839 cases of BC death, and 824 cases of OC death. Counts for breast cancer death and death from other causes are also summarized in Table 3.2. As shown in Figure 3.3, the more advanced stage at diagnosis, the higher the death probability from BC, this is particularly true for short-term mortality risk (consistent with other findings [34,64]). However for women with dense breasts the 95% CI for the distant BC stage ((0.184,0.747) in Figure 3.3-3 and (0.084, 0.861) in Figure 3.3-9) overlap with the regional stage ((0.159, 0.222) in Figure 3.3-2 and (0.138, 0.201) in Figure 3.3-8) while the CIs do not overlap for the fatty group. The two density groups behave similarly with respect to all other aspects, controlling for age and cancer stage.

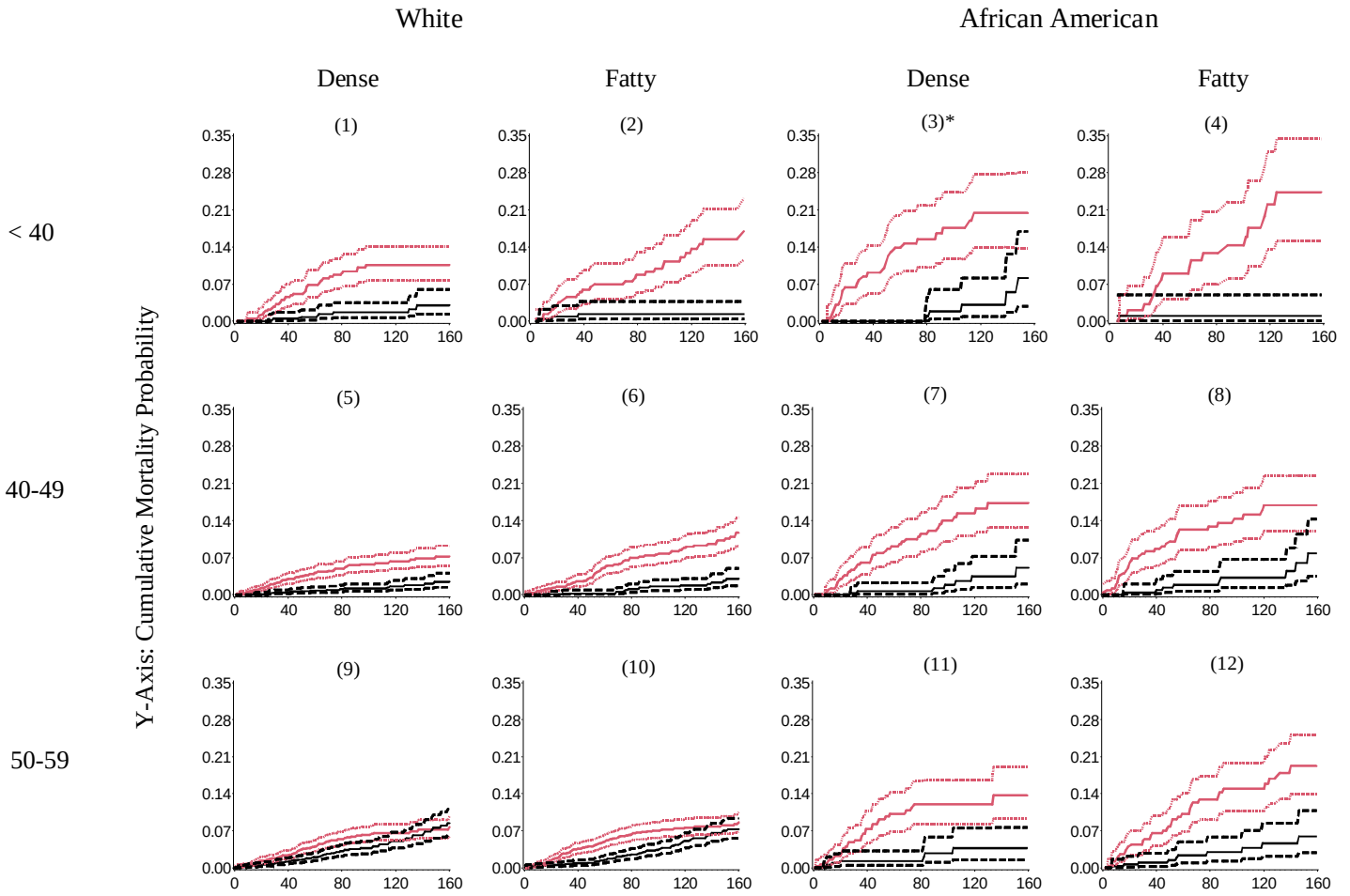
Table 3.1 Summary of study population and estimation of mortality probabilities with 95% confidence interval by age, race and density

Age Grp	Race	Density	Counts			Mortality at 5-year (60 months)		Mortality at 10-year (120 months)		Mortality at 13-year (156 months)	
			Case	BCD	OCD	BC Prob Est(CI)	OC Prob Est(CI)	BC Prob Est(CI)	OC Prob Est(CI)	BC Prob Est(CI)	OC Prob Est(CI)
<40	CAU	Dense	413	38	8	0.069(0.046-0.097)	0.008(0.002-0.022)	0.106(0.077-0.141)	0.017(0.007-0.035)	0.106(0.077-0.141)	0.031(0.013-0.060)
		Fatty	226	31	3	0.070(0.041-0.109)	0.014(0.004-0.037)	0.137(0.092-0.191)	0.014(0.004-0.037)	0.155(0.106-0.212)	0.014(0.004-0.037)
	AA	Dense	165	26	5	0.138(0.089-0.199)	0*	0.204(0.139-0.277)	0.032(0.009-0.082)	0.204(0.137-0.280)	0.081(0.029-0.169)
		Fatty	97	18	1	0.102(0.051-0.175)	0.010(0.001-0.050)	0.220(0.135-0.318)	0.010(0.001-0.050)	0.243(0.153-0.344)	0.010(0.001-0.050)
40-49	CAU	Dense	1406	71	21	0.041(0.031-0.054)	0.008(0.004-0.015)	0.063(0.049-0.080)	0.016(0.009-0.027)	0.072(0.055-0.092)	0.025(0.014-0.041)
		Fatty	915	79	16	0.051(0.037-0.068)	0.002(0.001-0.008)	0.088(0.068-0.111)	0.018(0.010-0.031)	0.019(0.085-0.137)	0.031(0.017-0.050)
	AA	Dense	356	40	8	0.091(0.062-0.127)	0.007(0.001-0.022)	0.154(0.111-0.202)	0.035(0.014-0.072)	0.173(0.126-0.227)	0.051(0.020-0.104)
		Fatty	258	34	10	0.123(0.084-0.168)	0.019(0.006-0.045)	0.169(0.120-0.224)	0.033(0.013-0.066)	0.169(0.119-0.225)	0.078(0.035-0.143)
50-59	CAU	Dense	1642	78	60	0.041(0.031-0.053)	0.020(0.013-0.029)	0.066(0.052-0.082)	0.051(0.037-0.068)	0.072(0.056-0.090)	0.080(0.059-0.105)
		Fatty	1479	96	70	0.049(0.038-0.062)	0.015(0.010-0.023)	0.076(0.061-0.092)	0.051(0.038-0.067)	0.080(0.064-0.097)	0.073(0.056-0.094)
	AA	Dense	340	30	7	0.102(0.069-0.143)	0.013(0.004-0.031)	0.120(0.083-0.165)	0.037(0.015-0.076)	0.137(0.092-0.190)	0.037(0.015-0.076)
		Fatty	324	40	11	0.101(0.069-0.141)	0.024(0.010-0.048)	0.149(0.107-0.198)	0.046(0.022-0.084)	0.192(0.139-0.252)	0.060(0.029-0.108)
60-69	CAU	Dense	1372	79	111	0.044(0.033-0.057)	0.027(0.019-0.039)	0.072(0.056-0.090)	0.101(0.081-0.124)	0.089(0.069-0.111)	0.163(0.134-0.194)
		Fatty	1818	103	173	0.031(0.023-0.041)	0.040(0.031-0.051)	0.068(0.054-0.083)	0.100(0.084-0.118)	0.083(0.067-0.101)	0.146(0.124-0.169)
	AA	Dense	177	20	17	0.079(0.043-0.128)	0.051(0.024-0.093)	0.144(0.086-0.215)	0.114(0.063-0.181)	0.195(0.119-0.285)	0.163(0.092-0.252)
		Fatty	346	27	38	0.064(0.040-0.095)	0.050(0.029-0.079)	0.096(0.064-0.137)	0.130(0.090-0.177)	0.107(0.070-0.152)	0.169(0.119-0.226)
70-79	CAU	Dense	887	48	160	0.041(0.029-0.058)	0.090(0.070-0.112)	0.074(0.053-0.098)	0.259(0.223-0.297)	0.097(0.070-0.128)	0.335(0.293-0.378)
		Fatty	1524	97	270	0.056(0.044-0.070)	0.067(0.054-0.081)	0.080(0.065-0.097)	0.221(0.195-0.248)	0.090(0.073-0.110)	0.327(0.296-0.358)
	AA	Dense	133	16	23	0.126(0.073-0.195)	0.107(0.058-0.174)	0.161(0.095-0.243)	0.269(0.182-0.364)	0.161(0.092-0.247)	0.327(0.232-0.425)
		Fatty	239	32	42	0.120(0.080-0.168)	0.119(0.079-0.167)	0.163(0.112-0.223)	0.229(0.168-0.296)	0.212(0.147-0.284)	0.285(0.211-0.363)
>=80	CAU	Dense	353	21	116	0.052(0.030-0.082)	0.211(0.167-0.259)	0.081(0.050-0.123)	0.506(0.445-0.563)	0.093(0.053-0.147)	0.643(0.523-0.739)
		Fatty	597	50	187	0.083(0.061-0.110)	0.247(0.210-0.285)	0.108(0.079-0.142)	0.490(0.444-0.534)	0.126(0.090-0.167)	0.561(0.492-0.624)
	AA	Dense	58	11	11	0.204(0.110-0.318)	0.192(0.100-0.305)	0.204(0.091-0.349)	0.334(0.213-0.460)	0.204(0.091-0.349)	0.334(0.213-0.460)
		Fatty	118	14	39	0.132(0.075-0.205)	0.286(0.205-0.373)	0.148(0.078-0.238)	0.504(0.407-0.594)	0.148(0.078-0.238)	0.504(0.407-0.594)

Abbreviations: Grp, group; CAU, Caucasian; AA, African American; BCD, breast cancer death; OCD other cause death; Prob Est, Probability Estimate; CI, confidence interval

* When there is no event observed, the probability estimate is zero with no confidence interval.

Figure 3.2 Mortality probability with confidence interval by age group, race and density



* Estimate will remain zero when there is no event happening until later time.

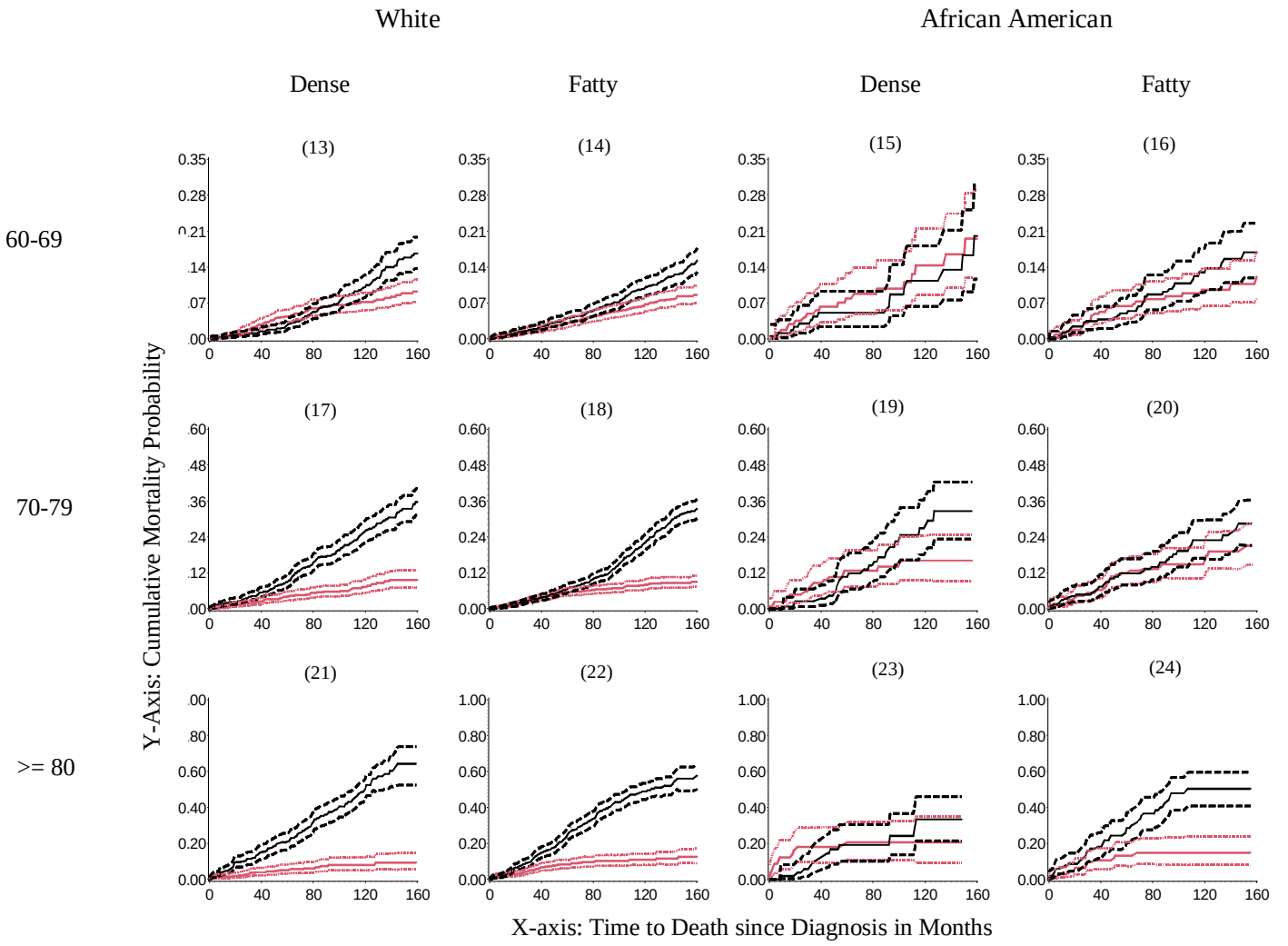
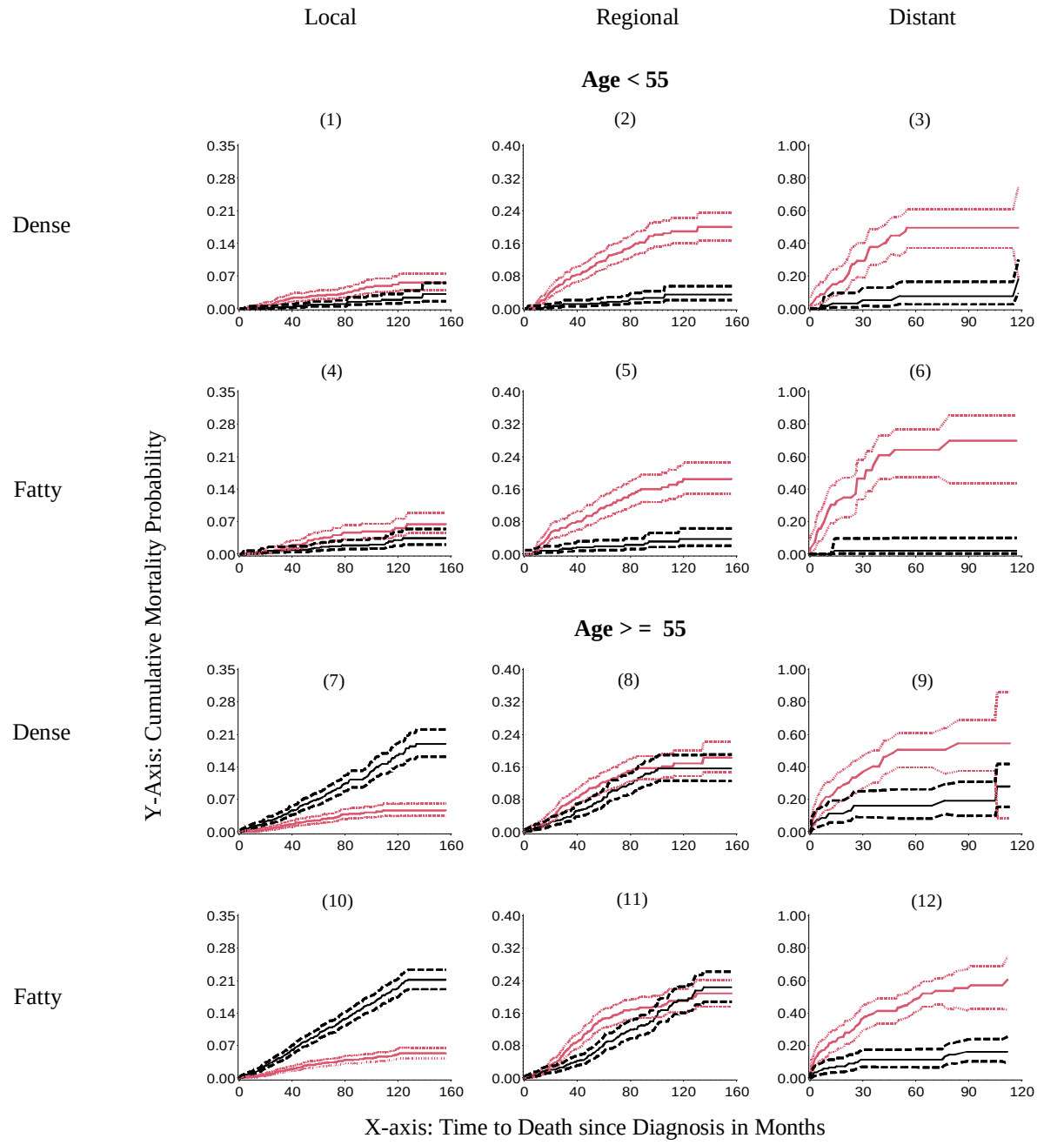

White
African American
Dense
Fatty
Dense
Fatty
60-69
70-79
>= 80
(13)
(14)
(15)
(16)
(17)
(18)
(19)
(20)
(21)
(22)
(23)
(24)
Y-Axis: Cumulative Mortality Probability
X-axis: Time to Death since Diagnosis in Months
Estimation for Mortality from Breast Cancer
Confidence Interval for Breast Cancer Deaths
Estimation for Mortality from Other Causes
Confidence Interval for Other Causes Deaths

Table 3.2 Summary of study population and estimation of mortality probabilities with 95% confidence interval by age, cancer stage and density

Age Grp	ST	DN	Counts			Mortality at 5-year (60 months)		Mortality at 10-year (120 months)		Mortality at 13-year (156 months)	
			Case	BCD	OCD	BC Prob Est(CI)	OC Prob Est(CI)	BC Prob Est(CI)	BC Prob Est(CI)	OC Prob Est(CI)	BC Prob Est(CI)
< 55	L	D	1487	47	17	0.028(0.019-0.038)	0.008(0.004-0.015)	0.051(0.037-0.069)	0.019(0.010-0.031)	0.055(0.040-0.074)	0.031(0.015-0.055)
		F	980	38	19	0.032(0.021-0.046)	0.013(0.007-0.023)	0.055(0.039-0.075)	0.030(0.018-0.048)	0.064(0.045-0.088)	0.034(0.020-0.054)
	R	D	946	119	18	0.117(0.096-0.141)	0.014(0.008-0.025)	0.190(0.159-0.222)	0.034(0.020-0.054)	0.200(0.166-0.236)	0.034(0.020-0.054)
		F	595	76	14	0.114(0.088-0.143)	0.019(0.010-0.034)	0.184(0.147-0.225)	0.037(0.020-0.063)	0.184(0.147-0.225)	0.037(0.020-0.063)
	DT	D	71	27	5	0.495(0.369-0.609)	0.075(0.025-0.163)	0.495(0.184-0.747)	0.183(0.092-0.297)	NA*	NA
		F	53	29	1	0.642(0.476-0.767)	0.022(0.002-0.098)	0.698(0.437-0.855)	0.022(0.002-0.100)	NA	NA
>= 55	L	D	1924	53	174	0.025(0.018-0.034)	0.070(0.058-0.084)	0.046(0.034-0.060)	0.165(0.141-0.191)	0.046(0.034-0.060)	0.190(0.162-0.220)
		F	2853	92	338	0.030(0.024-0.038)	0.094(0.082-0.106)	0.051(0.041-0.063)	0.194(0.175-0.214)	0.053(0.042-0.065)	0.212(0.191-0.234)
	R	D	924	103	83	0.122(0.099-0.148)	0.086(0.066-0.108)	0.168(0.138-0.201)	0.157(0.126-0.190)	0.183(0.147-0.221)	0.157(0.126-0.191)
		F	1133	153	122	0.146(0.124-0.170)	0.083(0.066-0.103)	0.190(0.162-0.220)	0.192(0.161-0.225)	0.208(0.176-0.243)	0.224(0.188-0.263)
	DT	D	83	35	14	0.507(0.397-0.607)	0.161(0.083-0.263)	0.545(0.084-0.861)	0.279(0.155-0.418)	NA	NA
		F	145	67	19	0.483(0.403-0.558)	0.114(0.065-0.176)	0.604(0.419-0.746)	0.163(0.092-0.253)	0.604(0.291-0.814)	0.202(0.127-0.289)

Abbreviations: ST, cancer stage; L, local; R, regional; DT, distant; DN, breast density; D, dense; F, fatty

* There is no event (mortality or censoring) observed beyond this time point.



* Estimate for Confidence Intervals will grow wider at later times when sample size is small.

Figure 3.3 Mortality probability with confidence interval by age group, cancer stage and density

3.3.2 Estrogen and Progesterone Receptor Status

From Table 3.3, there are a total of 10,429 cases with known race and receptor status (598 have died from BC and 618 from OC) and 10,218 cases with known cancer stage and receptor status (643 BC and 636 OC deaths, respectively). The majority of the patients had ER+/PR+ combination, with the fewest cases with ER-/PR+ tumor status.

As shown in Figure 3.4, for Caucasian BC patients, when PR status is positive, the probability of dying from breast cancer in the ER- group is significantly higher than in the ER+ group (e.g., 10-year CI is (0.082, 0.179) for ER- as shown in Figure 3.4-3 compared to (0.057, 0.076) for the ER+ group in Figure 3.4-1). For Caucasian PR- group, the ER+ and ER- groups do not have significantly different mortality probabilities (e.g., 10-year CI (0.063, 0.113) in Figure 3.4-2 vs. (0.107, 0.148) in Figure 3.4-4, respectively). For African American women, when PR is either negative or positive, there is not enough evidence to indicate significant differences in the BC death probabilities among the ER+ and ER- groups as the paired confidence intervals overlap ((0.101, 0.411) in Figure 3.4-7 vs. (0.086, 0.147) in Figure 3.4-5). When ER status is controlled for, there are almost no significant differences between the two PR groups except the 5-year BC mortality for African American ER+ women, the PR- group ((0.096,0.192) in Figure 3.4-6) is higher than the PR+ group((0.052, 0.091) in Figure 3.4-5).

The differences between the BC and OC mortality probability curves for different risk groups are also shown in Figure 3.4. For Caucasian women, the CIs for BC death and OC death do not overlap for the ER+/PR+ group, but overlap for the ER+/PR-group. Interestingly for African American women, it is the opposite. For the ER-/PR+ group, both

races have similar patterns in that the BC and OC CIs overlap. In the ER-/PR- group for Caucasian patients there are significant differences between the BC and OC mortality probabilities before the 10-year mortality probability after which they overlap (10-year CI: (0.107, 0.148) vs. (0.079, 0.123) in Figure 3.4-4), while for African American patients, the BC and OC mortality probabilities are always significantly different (Figure 3.4-8).

Controlling for cancer stage at diagnosis (results are shown in Table 3.4 and Figure 3.5), for local stage BC, when PR is positive, there is no significant difference in mortality from BC at 5-year between ER+ and ER- groups (CI: (0.011, 0.019) in Figure 3.5-1 and (0.024, 0.092) in Figure 3.5-7). However if PR is negative, the BC mortality probability in the ER- group (0.045, 0.073) in Figure 3.4-10 is higher than the ER+ group (0.015, 0.042) in Figure 3.5-7. The results are similar results for regional stage BC. However, there are no significant differences for patients detected with distant stage BC. For patients with local stage BC who are PR+, the difference of the 10-year BC mortality between the ER+ group and the ER- group is not statistically significant (CI: (0.027, 0.043) in Figure 3.5-1 vs. (0.035, 0.117) in Figure 3.5-7). However the difference is significant for women diagnosed with regional BC ((0.128, 0.174) in Figure 3.5-2 vs. (0.184, 0.399) in Figure 3.5-8). The results are similar for PR- patients; there is no significant difference shown in mortality for the PR groups when ER status is controlled for. As shown in Figure 3.4, the two mortality curves do not overlap for women diagnosed with local stage BC for both the ER+/PR+ and the ER+/PR- groups. Women with regional stage BC have a higher BC mortality probability. However, for ER negative patients diagnosed with local stage BC, the BC and OC mortality probabilities are not statistically different

Table 3.3 Summary of study population and estimation of mortality probabilities with 95% confidence interval by race and ER/PR status

Race	RS	Counts			Mortality at 5-year (60 months)		Mortality at 10-year (120 months)		Mortality at 13-year (156 months)	
		Case	BCD	OCD	BC Prob Est(CI)	OC Prob Est(CI)	BC Prob Est(CI)	OC Prob Est(CI)	BC Prob Est(CI)	OC Prob Est(CI)
Caucasian	ER+/PR+	5333	197	341	0.038(0.033-0.045)	0.055(0.048-0.063)	0.066(0.057-0.076)	0.121(0.108-0.134)	0.068(0.058-0.078)	0.138(0.123-0.154)
	ER+/PR-	1004	47	58	0.060(0.044-0.081)	0.064(0.047-0.084)	0.086(0.063-0.113)	0.104(0.078-0.134)	0.086(0.063-0.113)	0.114(0.084-0.149)
	ER-/PR+	217	23	13	0.110(0.070-0.160)	0.055(0.028-0.094)	0.126(0.082-0.179)	0.075(0.040-0.124)	0.150(0.095-0.207)	0.096(0.050-0.160)
	ER-/PR-	1603	141	83	0.112(0.095-0.131)	0.045(0.034-0.059)	0.127(0.107-0.148)	0.099(0.079-0.123)	0.127(0.107-0.148)	0.105(0.082-0.130)
African American	ER+/PR+	956	62	77	0.070(0.052-0.091)	0.077(0.058-0.099)	0.114(0.086-0.147)	0.168(0.132-0.207)	0.132(0.098-0.171)	0.168(0.132-0.208)
	ER+/PR-	244	32	10	0.140(0.096-0.192)	0.049(0.024-0.087)	0.188(0.133-0.250)	0.059(0.030-0.103)	NA *	NA
	ER-/PR+	84	10	7	0.086(0.036-0.164)	0.030(0.006-0.091)	0.195(0.101-0.311)	0.141(0.062-0.250)	0.195(0.101-0.311)	0.141(0.062-0.250)
	ER-/PR-	595	86	29	0.161(0.130-0.195)	0.055(0.036-0.079)	0.202(0.165-0.243)	0.073(0.049-0.103)	0.202(0.165-0.243)	0.073(0.049-0.103)

Abbreviations: RS, receptor status; ER, estrogen receptor; PR, progesterone receptor; +, positive; -, negative

* There is no event (mortality or censoring) observed beyond this time point.

Table 3.4 Summary of study population and estimation of mortality probabilities with 95% confidence interval by cancer stage and ER/PR status

Stage	RS	Counts			Mortality at 5-year (60 months)		Mortality at 10-year (120 months)		Mortality at 13-year (156 months)	
		Case	BCD	OCD	BC Prob Est(CI)	OC Prob Est(CI)	BC Prob Est(CI)	OC Prob Est(CI)	BC Prob Est(CI)	OC Prob Est(CI)
Local	ER+/PR+	4338	76	321	0.015(0.011-0.019)	0.063(0.055-0.072)	0.034(0.027-0.043)	0.136(0.121-0.151)	0.036(0.028-0.045)	0.152(0.135-0.170)
	ER+/PR-	822	19	48	0.026(0.015-0.042)	0.062(0.044-0.083)	0.043(0.026-0.066)	0.097(0.071-0.128)	0.043(0.026-0.066)	0.110(0.077-0.148)
	ER-/PR+	194	11	13	0.050(0.024-0.092)	0.062(0.032-0.106)	0.068(0.035-0.117)	0.077(0.042-0.125)	0.091(0.045-0.157)	0.098(0.052-0.161)
	ER-/PR-	1427	67	71	0.058(0.045-0.073)	0.043(0.031-0.056)	0.071(0.055-0.089)	0.093(0.072-0.117)	0.071(0.055-0.089)	0.097(0.075-0.122)
Regional	ER+/PR+	1906	159	101	0.089(0.074-0.105)	0.046(0.036-0.059)	0.150(0.128-0.174)	0.101(0.082-0.122)	0.160(0.135-0.186)	0.110(0.088-0.134)
	ER+/PR-	381	47	19	0.133(0.097-0.176)	0.049(0.028-0.080)	0.198(0.148-0.254)	0.090(0.054-0.138)	0.198(0.148-0.254)	0.090(0.054-0.138)
	ER-/PR+	90	18	5	0.164(0.094-0.251)	0.023(0.005-0.072)	0.287(0.184-0.399)	0.093(0.033-0.192)	NA	NA
	ER-/PR-	798	141	36	0.221(0.189-0.254)	0.043(0.028-0.061)	0.262(0.224-0.300)	0.076(0.053-0.105)	0.262(0.223-0.301)	0.089(0.059-0.127)
Distant	ER+/PR+	121	42	8	0.449(0.352-0.542)	0.084(0.037-0.155)	0.545(0.405-0.665)	0.113(0.054-0.196)	NA	NA
	ER+/PR-	41	14	3	0.509(0.349-0.649)	0.094(0.023-0.227)	0.509(0.349-0.649)	0.094(0.023-0.227)	NA	NA
	ER-/PR+	12	5	2	0.679(0.010-0.968)	0.083(0.002-0.375)	NA	NA	NA	NA
	ER-/PR-	88	44	9	0.645(0.464-0.778)	0.097(0.039-0.188)	0.645(0.252-0.869)	0.149(0.080-0.237)	NA	NA

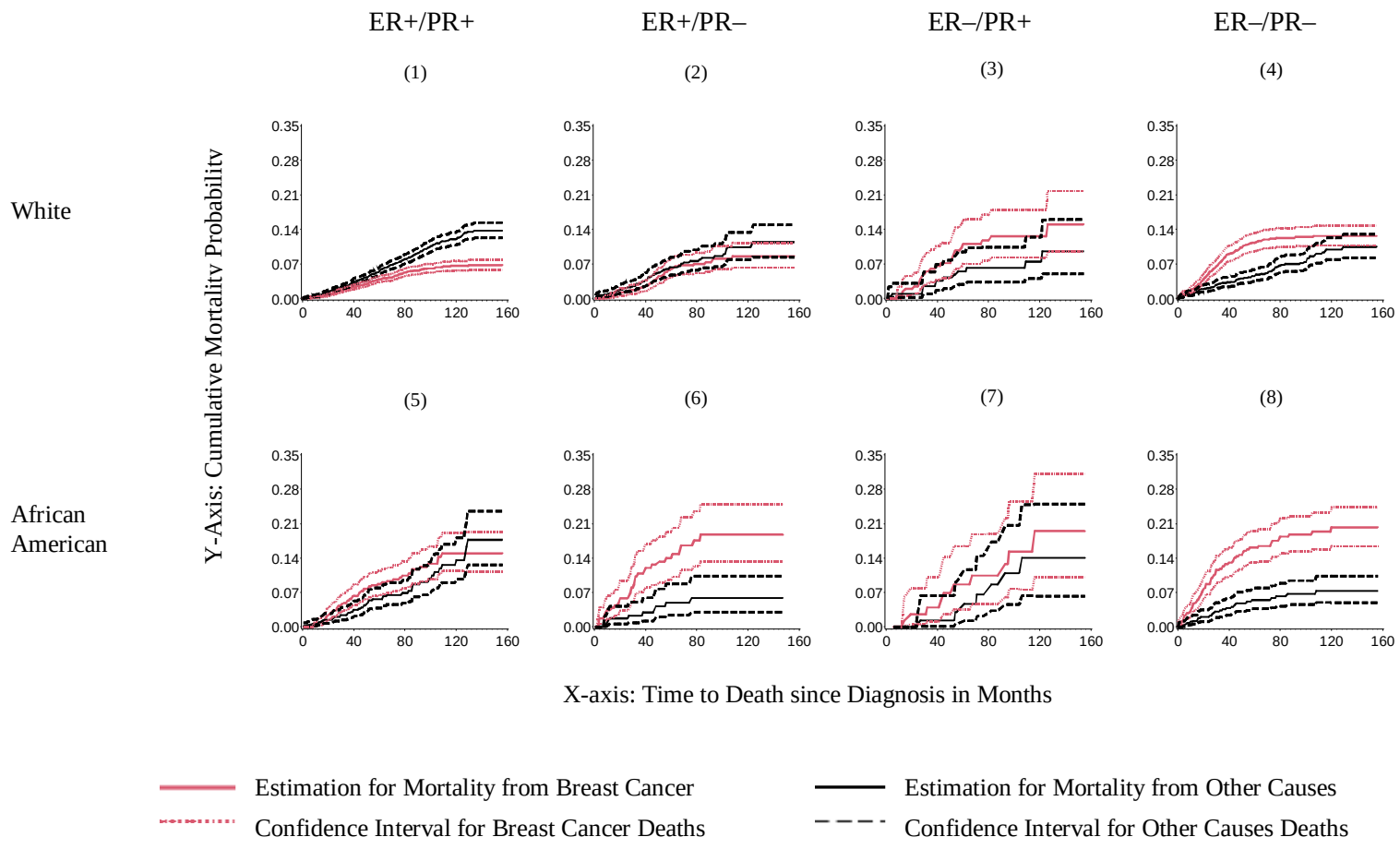
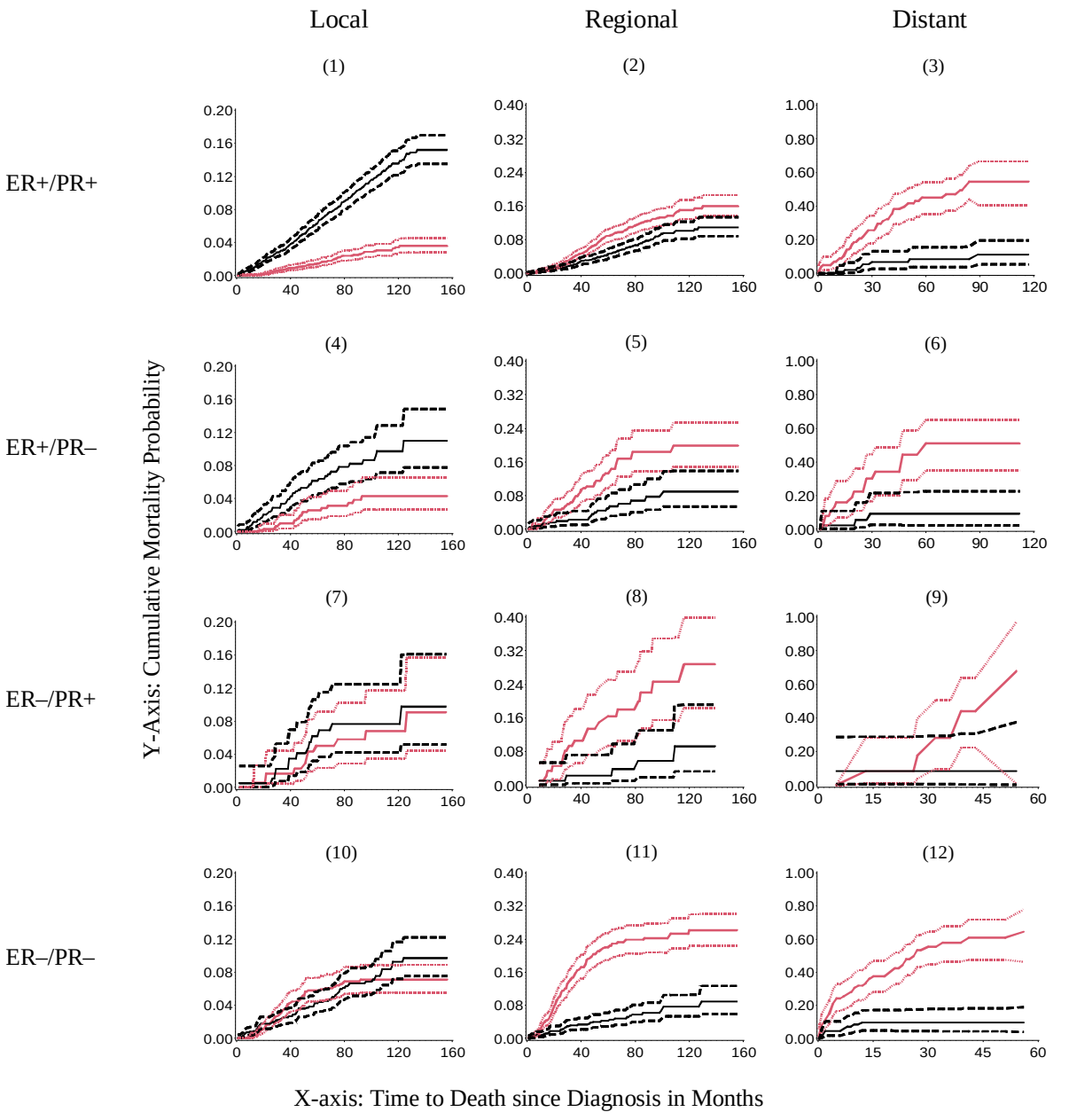


Figure 3.4 Mortality probability with confidence interval by race and ER/PR status



* Sample size in this group is very small.

Figure 3.5 Mortality probability with confidence interval by cancer stage and ER/PR status

3.3.3 Family History of Breast Cancer

There are 12,707 BC patients (899 BC and 916 OC mortality) with known race, cancer stage and family history and about 20% of them for each sub-group have a family history of BC, except for African American women diagnosed with distant BC, for whom only 10% have family history (FH). The results are summarized in Table 3.5 and Figure 3.6.

For Caucasian patients, those diagnosed at the local stage have similar behavior irrespective of FH in that the probabilities of dying from OC are significantly higher than from BC. For the FH+ group patients diagnosed at the regional stage, there are no significant differences (e.g., 10-year CI: (0.113, 0.191) vs. (0.097, 0.181) in Figure 3.6-2). However, there is significant difference for the FH- group (e.g., 10-year CI: (0.141, 0.178) vs. (0.097, 0.133) in Figure 3.6-5).

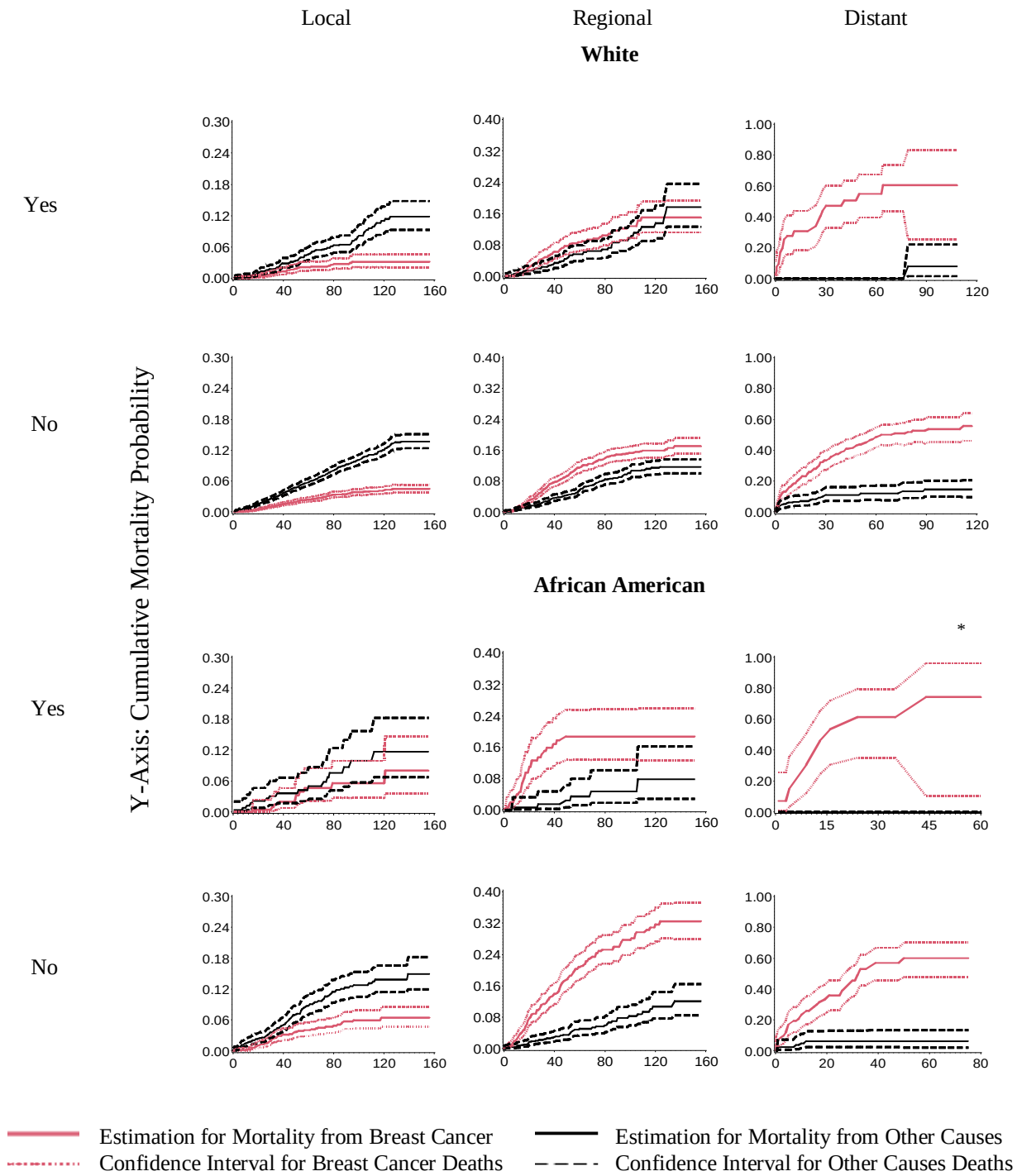
For African American patients diagnosed with local stage BC, the two confidence intervals overlap for the FH+ group (e.g., 5-year CI: (0.022,0.085) vs.(0.035,0.060) in Figure 3.6-7), but BC and OC CIs do not overlap for the FH- group after 40 months (e.g., 5-year CI: (0.029,0.055) vs. (0.070,0.108) in Figure 3.6-10). Also at regional stage, BC mortality probabilities are significantly higher than the OC mortality in both FH groups. Surprisingly, the BC mortality probabilities for the FH- group (e.g., 10-year CI: (0.274,0.360) in Figure 3.6-11) are significantly higher than for the FH+ group (10-year CI: (0.126, 0.258) in Figure 3.6-8), but there are a limited number of observations and no deaths occurred after 5 years for the FH+ group with regional BC. The death observations for groups with distant cancer stage also have a limited number of cases and hence the confidence intervals are wide in those groups.

Table 3.5 Summary of study population and estimation of mortality probabilities with 95% confidence interval by race, cancer stage and family history of breast cancer

Race	Cancer Stage	FH	Counts			Mortality at 5-year (60 months)		Mortality at 10-year (120 months)		Mortality at 13-year (156 months)	
			Case	BCD	OCD	BC Prob Est(CI)	OC Prob Est(CI)	BC Prob Est(CI)	OC Prob Est(CI)	BC Prob Est(CI)	OC Prob Est(CI)
CAU	Local	Yes *	1488	29	81	0.021(0.014-0.031)	0.046(0.035-0.060)	0.032(0.021-0.045)	0.110(0.086-0.137)	0.032(0.021-0.045)	0.119(0.093-0.148)
		No	5655	156	428	0.024(0.020-0.029)	0.058(0.051-0.065)	0.042(0.035-0.049)	0.121(0.110-0.133)	0.045(0.038-0.053)	0.137(0.124-0.151)
	Regional	Yes	646	57	46	0.087(0.065-0.114)	0.057(0.038-0.079)	0.150(0.113-0.191)	0.136(0.097-0.181)	0.150(0.112-0.193)	0.177(0.126-0.235)
		No	2379	262	161	0.115(0.101-0.130)	0.057(0.047-0.069)	0.159(0.141-0.178)	0.114(0.097-0.133)	0.170(0.150-0.192)	0.117(0.100-0.136)
	Distant	Yes	41	20	1	0.548(0.398-0.675)	0	0.604(0.254-0.831)	0.079(0.014-0.221)	NA	NA
		No	220	96	27	0.486(0.420-0.548)	0.118(0.077-0.170)	0.556(0.463-0.639)	0.146(0.096-0.206)	0.556(0.372-0.705)	0.184(0.129-0.246)
AA	Local	Yes	258	10	16	0.046(0.022-0.085)	0.050(0.025-0.087)	0.056(0.027-0.100)	0.117(0.067-0.182)	0.080(0.036-0.147)	0.117(0.067-0.183)
		No	1077	46	101	0.040(0.029-0.055)	0.088(0.070-0.108)	0.064(0.047-0.085)	0.139(0.114-0.166)	0.064(0.047-0.085)	0.149(0.120-0.182)
	Regional	Yes	160	25	6	0.187(0.127-0.256)	0.035(0.011-0.080)	0.187(0.126-0.258)	0.078(0.029-0.162)	NA	NA
		No	685	147	42	0.208(0.176-0.242)	0.051(0.035-0.072)	0.316(0.274-0.360)	0.108(0.077-0.144)	0.325(0.279-0.371)	0.121(0.085-0.164)
	Distant	Yes	14	9	0	0.742(0.103-0.962)	0	NA	NA	NA	NA
		No	84	42	7	0.600(0.478-0.702)	0.062(0.022-0.134)	NA	NA	NA	NA

Abbreviations: CAU, Caucasian; AA, African American; FH = family history

* Family history of breast cancer is present for the patient (No for negative case).



* No death from other causes is observed and the estimation will remain zero

Figure 3.6 Mortality probability with confidence interval by race, cancer stage and family history of cancer

3.3.4 Left Censoring

Using Method 1, for each time interval calculated, 1000 simulations are run to estimate the true cancer start time. Table 3.6 summarizes the breast cancer mortality probability after 5 years and 10 years for Caucasian and African American women, respectively. The results without incorporating left censoring are compared with those with left censoring. The average, minimum and maximum of the 1000 simulated left censoring times are selected for comparison.

Table 3.6 Breast cancer mortality probability at 5 and 10 year with left censoring method 1

BC Mortality	Race	No Left Censoring	Simulation1 (mean)	Simulation2 (max)	Simulation3 (min)
5-year probability	Caucasian	**0.066**	**0.059**	**0.059**	**0.059**
	African American	**0.136**	**0.124**	**0.123**	**0.126**
10-year probability	Caucasian	**0.098**	**0.097**	**0.097**	**0.097**
	African American	**0.191**	**0.186**	**0.185**	**0.189**

As shown in Table 3.6, incorporating left censoring does affect the survival time when the disease onset time is not observed. The choice of simulated estimate for the left censoring period (mean, max or min) has little effect on the mortality probabilities. For Caucasian patients, on the addition of the left censoring time has little impact on the long-term mortality probabilities. However, for African American women, the difference in mortality probabilities grows over time (e.g., the effect is larger for the 10-year mortality (0.191 vs. 0.186)). This suggests that African American women may have more left censored information compared to Caucasian women. Thus, the breast cancer mortality for African American patients may be over-estimated.

For Method 2, the fitted distributions corresponding to the growth rate *b*, and tumor size are shown in Figure 3.7.

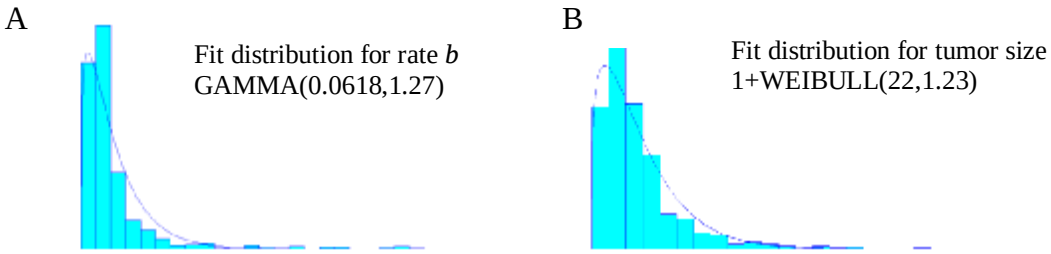


Figure 3.7 Input modeling for fitting distributions

A: corresponds to the Growth Rate and B: corresponds to the Tumor Size

The results comparing left censoring with the original model are summarized in Figure 3.8.

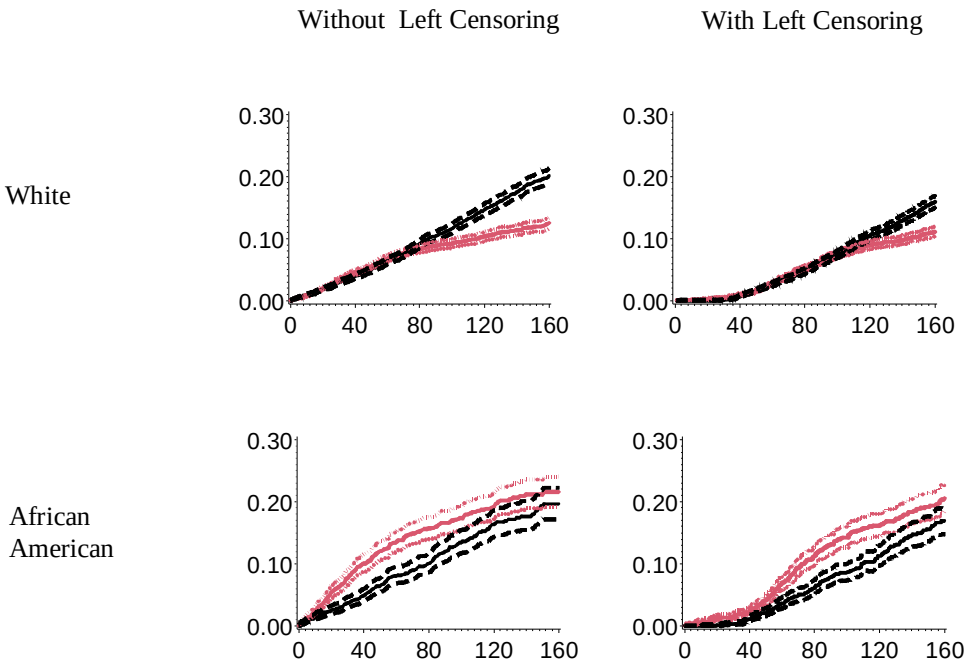


Figure 3.8 Mortality results comparison using left censoring method 2

Figure 3.8 suggests that adding left censoring adjusted the mortality probabilities downward to mitigate over-estimation. While there are no significant differences between the left censored and uncensored breast cancer mortality probabilities in the long run, the difference is significant over short time horizons, especially for African American patients. These results are consistent with Method 1.

3.4 Discussion

In this chapter, the cumulative incidence function with an approximate confidence interval is used to estimate mortality probabilities from breast cancer and other causes as a function of race, stage at detection, and age. Death probabilities are compared as a function of three under-studied breast cancer risk factors and tumor characteristics to quantify the impact of these factors on mortality for breast cancer patients.

The results suggest that although women with dense breast tissue are known to have a higher risk of developing breast cancer, density does not have a significant impact on mortality for breast cancer patients, except for women 70-79 with fatty breasts. Women with dense and fatty breast tissues have similar mortality probabilities after adjusting for age, race and cancer stage effects. This suggests that following diagnosis, patients with dense breast tissues do not have worse survival than women with fatty breast tissue.

Estrogen receptor status affects breast cancer mortality more than progesterone receptor status and race matters. Caucasian ER- patients have significantly higher breast cancer mortality probabilities than Caucasian ER+ patients. However, there is no significant

difference in BC mortality for African American women. The difference in mortality between the PR- and PR+ patients is not significant. There are differences in BC and OC mortality curves for different combinations of ER/PR. Receptor status significantly influences BC patient outcomes and this impact differs for African American and Caucasian women.

While family history of breast cancer is considered to be an important risk factor for breast cancer, it does not directly affect mortality. Although incidence may be higher for women with positive FH, the mortality risks are not significantly different from women without BC FH. However, between the FH groups the risks associated with breast cancer death and deaths from other causes are different for some race and cancer stage groups. While there is not enough evidence to indicate poorer survival for women with a family history of breast cancer, the mortality risk may be affected by other factors that reduce the difference in mortality from other causes and breast cancer, particularly in African American patients.

The method used for the analysis provides an estimate for the mortality probabilities so that risks may be compared among different groups. However, there are some limitations to this study. The first is the impact of small sample size on the approximation of confidence intervals. When the sample size is very small, confidence intervals tend to be wide; and conversely intervals will be small if there is a relatively large sample size. This means the point estimate should be not ignored when comparing probabilities among different groups. Another limitation is that lifetime follow-up for the patients is not available as the CMR started in 1994. Future research could be to estimate long-term mortality for breast cancer

patients with different risk information and to study the impact of additional risk factors on mortality for breast cancer patients.

In this dissertation, models to quantify the impact of left censoring on breast cancer mortality have been developed. These results suggest that it is important to incorporate left censoring for more accurate estimation. The methods presented in this dissertation use a simplified model to find the cancer onset time. However, this research has initiated a new direction for future research. Methods can be developed to address left censoring, and some assumptions in this dissertation can be relaxed. For example, additional distribution assumptions can be tested and different data can be used to validate the estimation.

Chapter 4 Decision Modeling with Disease Spontaneous Regression

4.1 Introduction

As discussed in previous chapters, breast cancer is typically modeled as a progressive disease, under the assumption that the cancer will not resolve in absence of treatment, and in absence of treatment the cancer will advance. For example, most Markov models introduced in Chapter 1 only allowed a transition to a more advanced cancer stage or to absorbing death states. This is a common assumption in most screening and treatment decision models. However, there has been medical evidence suggesting that breast cancer may actually spontaneously resolve without treatment. While this has initiated a lot of debate in the medical community, there have been limited analytical studies on the topic. This research seeks to quantify the impact of breast cancer spontaneous regression on patient outcomes with respect to different mammography screening and treatment policies. In this chapter, a literature review of several medical studies regarding breast cancer regression is discussed in Section 4.2. A partially observable Markov model is presented in Section 4.3, with results in Section 4.4. Section 4.5 concludes the chapter with a summary and a discussion of future work.

4.2 Medical Evidence

The medical exploration of the spontaneous regression phenomenon of breast cancer can be traced back as early as the beginning of last century [96]. In May 1974, Johns Hopkins Medical Institutions and the American Cancer Society organized the Conference on Spontaneous Regression of Cancer in Baltimore, Maryland. The conference was initiated with more recognition of dramatic but rare regression of cancer in the absence of treatment. Lewison [97] summarized some of the earlier findings in breast cancer regression, and presented three cases of spontaneous regression from his medical practices. He concluded that although the phenomenon is rare, there is ample clinical evidence to confirm that spontaneous regression of breast cancer does exist.

This has not only received attention from medical experts in the U.S., there have also been several international studies on breast cancer regression. Larsen and Rose [98] suggested that the natural history of breast cancer is variable, and there are a small number of patients who survive 10-15 years without treatment. They also found 32 cases of breast cancer spontaneous regression in the international literature. Although the phenomenon was considered to be rare, this suggests that the natural course for breast cancer is very variable, not as straight-forward as we generally model, and the possibility of regression should not be ignored.

In a recent case study, Burnside et al. [99] reported an actual example of breast cancer regression on imaging. They followed a 64-year-old asymptomatic female from an initial mammogram screening and recorded imaging from subsequent exams. Figure 4.1 shows two of the mammogram images selected from the report. Figure 4.1A is the spot compression

magnification view in the Craniocaudal (CC) projection from the woman's initial screening visit, which confirms a mass in the left inner breast. One year after the initial screening, the CC projection view as seen in Figure 4.1B has demonstrated that the mass has disappeared. Additional samples of imaging can be found in this article.

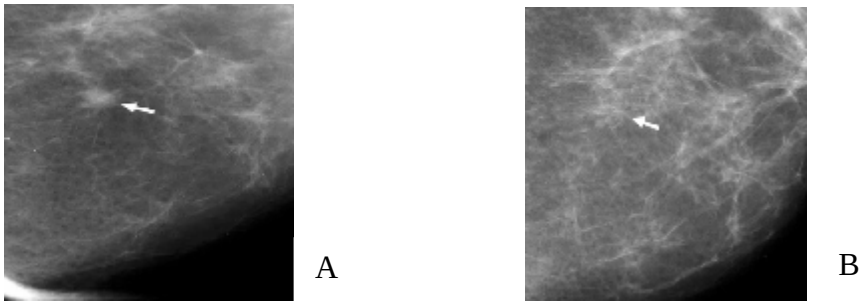


Figure 4.1 Mammogram screening images for a 64-year-old female [99]

A: Mass identified by arrow; B: image of same breast one year later without mass

Most of the examples shown above are medical reports on individual cases of breast cancer spontaneous regression. Because the current protocol is for women to seek treatment after diagnosis, it is difficult to observe the natural history of breast cancer progression and regression. Thus, it is not easy to calculate the probability of breast cancer regression.

Zahl et al. [100] presented a controversial argument that the natural course of some screening-detected invasive breast cancer may include spontaneously regression. This brought immediate attention from the media and the medical community. Different from previous medical reports, their findings were based on an observational study using the Norwegian Cancer Registry data, where cumulative incidence rates of breast cancer were

compared between two groups. The screening group included women between the ages of 50 and 64 who had three screenings over a six-year period between 1996 and 2001. The control group included women of the same age cohort who had no screenings between 1992 through 1997 when there was no screening program offered. These women were invited to a one-time prevalence screening at the end of the observation period. Other risk factors in these two groups were similar. When the number of biopsy-confirmed invasive cancers was compared between the two groups, it was found that the number in the control group was approximately 22% lower. Zahl et al.'s explanation for the difference was that some screen-detected breast cancer may spontaneously regress.

Although this phenomenon has been acknowledged for a long time in the medical literature, there have been limited analytical studies on this topic. The Wisconsin simulation model [26] actually incorporated regression in their stochastic model. This model replicated breast cancer incidence and mortality to fit real data, and they defined approximately 40% of the initiated breast cancer to be limited malignant potential (LMP) tumors. These tumors "progress to a maximum of approximately 1-cm in diameter, dwell at this size for 2 years, and then regress if undetected."

This dissertation research seeks to quantify the impact of breast cancer regression on patient outcomes with respect to different mammography screening and treatment polices.

4. 3 Model Formulation

4.3.1 Markov Chain Representation

A partially-observable, five-state, discrete-time Markov model is built to represent breast cancer development, as shown in Figure 4.2. Different from other Markov models for breast cancer, [23, 101-102] the transition from the in situ stage to a cancer-free state is now allowed to represent disease regression. In this research, only in situ cancer regression is considered. There are two reasons for this assumption. First, in situ cancer cells have not grown through the duct walls into the neighboring tissue. They are only detectable through mammography and are associated with low mortality risk. Thus, if in situ cancer is assumed to have a chance to regress, the efficacy of mammography may be explored, and various treatment protocols can be evaluated. Second, a patient who is diagnosed at the invasive cancer stage will probably seek treatment after diagnosis as invasive cancers have higher mortality risk.

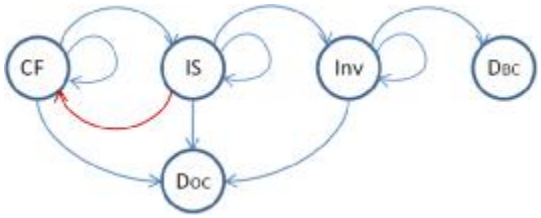


Figure 4.2 Markov representation for breast cancer progression and regression

CF = cancer free, IS = in situ, Inv = invasive, D_{BC} = Death from breast cancer D_{OC} = death from other causes

There are several other assumptions for the decision modeling. First, mammograms are assumed to be independent of each other. In other words, the results of the previous

mammogram screenings do not impact results of future screening. This is reasonable as a radiologist who is reading the mammogram may not refer to previous screening records. Second, if a screening result is abnormal, a biopsy will always be ordered, and the biopsy is assumed to be perfect, since it tends to have a high sensitivity [103]. This is used to identify if a screening is a false or true positive. If a mammogram result is found to be incorrect (i.e., a false positive result), the woman will continue mammography screenings as scheduled. Although the biopsy may impact the results of succeeding screenings, that effect is ignored in this study. It should be noted that the false negatives cannot be differentiated from the true negatives, and the woman will continue screening until either cancer is detected, she dies from breast cancer or another cause, or the ending age for screening is reached.

The underlying disease states (cancer free, in situ and invasive) are partially observable. The cancer states are only known with certainty after the cancer is detected during screening and confirmed through biopsy. In this model, death states are not partially observable but are absorbing states.

4.3.2 Policy Evaluation Framework

A policy evaluation framework is adopted to quantify the impact of the probability of breast cancer regression on screening policy and various treatment decision rules. This model extends the previous work by Maillart et al. [101] to include regression and to compare and evaluate different treatment rules.

The notation is introduced below:

$\underline{\pi}_n=[\pi_{n,1} \quad \pi_{n,2} \quad \pi_{n,3}]$ is the disease occupancy distribution at period n. These represent the belief states, i.e., with probability $\pi_{n,1}$ that a woman is considered to be in the cancer-free state, with probability $\pi_{n,2}$ she is believed to be in the in situ stage, and she is believed to be in the invasive cancer state with probability $\pi_{n,3}$.

$W_n(\underline{\pi}_n)$ is the lifetime probability of breast cancer mortality at period n. It is also the main outcome measure. It is a function of the belief states. The state transition probability from state i to state j for a woman at age a_n is $p_{ij}(a_n)$ for $i, j \in (1,2,3)$. $m_j(a_n)$ is the probability of an abnormal mammogram result given that a woman is at age a_n and she is in the state j, j = 1,2,3. $r_1(a_n)$ is the lifetime breast cancer mortality probability if the cancer is detected at the in situ stage and treatment is performed, and $r_2(a_n)$ is the corresponding lifetime breast cancer mortality probability if the cancer is detected at the invasive stage and treatment is performed.

In a given period n, there are two actions a woman can take: do nothing or have a mammogram. If the woman chooses to do nothing, the lifetime breast cancer mortality probability is calculated using equation (4.1).

$$\begin{aligned} W_n(\underline{p}_n) = DN_n(\underline{p}_n) \equiv\ & p_{n,3}p_{34}(a_n)\cdot 1 + (p_{n,1}+p_{n,2}+p_{n,3})p_5(a_n)\cdot 0 \\ & + [1-p_{n,3}p_{34}(a_n)-(p_{n,1}+p_{n,2}+p_{n,3})p_5(a_n)]\cdot W_{n+1}(\underline{p}_{n+1}) \\ =\ & p_{n,3}p_{34}(a_n)+[1-p_{n,3}p_{34}(a_n)-p_5(a_n)]\cdot W_{n+1}(\underline{p}_{n+1}) \end{aligned} \tag{4.1}$$

This estimation includes three parts: a woman has a lifetime breast cancer mortality probability of 1 if she dies from breast cancer in the next period; if she dies from other causes, then the probability she will die from breast cancer is zero; and if she survives the

next period, the problem starts again with W_{n+1}. The state occupancy distribution is updated using Bayes rule as follows,

$$\underline{p}_{n+1} = \frac{\underline{p}_n T(a_n)}{1 - p_{n,3} p_{34}(a_n) - p_5(a_n)} \tag{4.2}$$

where $T(a_n)$ is the transient state transition matrix.

The other action a woman can take is to have a screening mammogram. If the mammogram result is negative, then the woman will do nothing and continue screening at the next period. If an invasive cancer is confirmed, she will go on treatment immediately after diagnosis and leave the model. If an in situ cancer is confirmed, then there are three treatment decisions rule to choose from: (1) treat immediately; (2) do not treat immediately but continue screening until the cancer is detected at the invasive stage; and (3) do not treat immediately but continue regular screening and if the cancer is detected a second time then seek treatment. The formulation for calculating the lifetime breast cancer mortality probability (W_n) associated with each of these treatment rules is discussed next.

Treatment Rule 1: Always treat. Under this rule, the lifetime probability of breast cancer death at time n becomes:

$$\begin{aligned} W_n(\underline{p}_n) = M_n(\underline{p}_n) \equiv\ & p_{n,1} m_1(a_n) DN_n(\underline{e}_0) + p_{n,2} m_2(a_n) r_1(a_n) + p_{n,3} m_3(a_n) r_2(a_n) \\ & + [\textstyle\sum_{i=1}^{3} p_{n,i} (1 - m_i(a_n))] \cdot DN_n(\underline{p}_n'') \end{aligned} \tag{4.3}$$

where $\underline{e}_0 = [1 \ \ 0 \ \ 0]$.

The lifetime breast cancer mortality probabilities are known as a function of cancer stage at detection in situ (r_1) or invasive stage (r_2), and patient age. The first term in equation (4.3) corresponds to the case where the cancer is found to be a false positive after biopsy, then the woman is known to be in the cancer-free state for certainty, and she will do nothing for this period. The second, third and fourth terms calculate the impact of the mammogram results. If the screening result is negative (true or false), the woman will also do nothing, but the occupancy distribution is updated based on the negative finding according to the following.

$$p''_{n,j} = \frac{p_{n,j}(1-m_j(a_n))}{\sum_{i=1}^{3} p_{n,i}(1-m_i(a_n))}, \quad j = 1,2,3 \tag{4.4}$$

Treatment Rule 2: Wait until detection at the invasive stage. Under this rule the lifetime probability of breast cancer death at time *n* becomes:

$$W_n(\underline{p}_n) = M_n(\underline{p}_n) \equiv p_{n,1}m_1(a_n)DN_n(\underline{e}_0) + p_{n,2}m_2(a_n)DN_n(\underline{e}_1) + p_{n,3}m_3(a_n)r_2(a_n) + [\sum_{i=1}^{3} p_{n,i}(1-m_i(a_n))]\cdot DN_n(\underline{p}''_n) \tag{4.5}$$

where $\underline{e}_1 = [0 \;\; 1 \;\; 0]$.

The part of equation (4.5) that is different from equation (4.3) is highlighted in red. This represents the fact that a woman will do nothing if the cancer is diagnosed in the in situ stage. Since the cancer is confirmed, the occupancy distribution will have a probability of 1 in the in situ state.

Treatment Rule 3: Wait once after an in situ cancer is diagnosed. Under this rule the lifetime probability of breast cancer death at time *n* becomes:

$$W_n(\underline{p}_n) = M_n(\underline{p}_n) \equiv p_1 m_1(a_n) DN_n(\underline{e}_0) + p_2 m_2(a_n) DN_n(\underline{e}_1) \cdot \prod_{k=1}^{m-1}(1-m_2(a_k))$$
$$+ p_2 m_2(a_n) r_1(a_m) \cdot [1 - \prod_{k=1}^{m-1}(1-m_2(a_k))] + p_3 m_3(a_n) r_2(a_n)$$
$$+ [\sum_{i=1}^{3} p_i (1-m_i(a_n))] \cdot DN_n(\underline{p}''_n) \qquad (4.6)$$

where m is the number of mammograms so far.

Again, the parts of equation (4.5) that differ from equation 4.3 are highlighted in red. Under this treatment rule, a woman will only wait once at the in situ stage. If the cancer is detected for the second time, she will go to treatment and leave the model. Thus, the number of previous detections is incorporated in the formulation. If this is the first time an in situ cancer is confirmed, then the woman will do nothing. If this is not the first time, then she will go to treatment and the lifetime breast cancer mortality probability is calculated.

These formulations have built a recursive relationship for the lifetime breast cancer mortality. In order to calculate the probabilities for each period, boundary conditions are needed for the last period N. The equations (4.7) for the case of do nothing and (4.8) for the case of mammogram actions are based on the assumption that if a woman is in the cancer states (in situ or invasive) in period N, she will die from breast cancer. Otherwise, she will die from other causes. The mortality probability at each period can then be calculated through backwards recursion.

$$DN_N(\underline{p}) = p_2 + p_3 \qquad (4.7)$$

$$M_N(\underline{p}_n) = p_{n,2} m_2(a_n) r_1(a_n) + p_{n,3} m_3(a_n) r_2(a_n)$$
$$+ [\sum_{i=1}^{3} p_{n,i} (1-m_i(a_n))] \cdot (p''_{n,2} + p''_{n,3}) \qquad (4.8)$$

The formulation is based on the sample path behavior. A woman starting in a cancer-free state can decide if she will do nothing or have a mammogram for the next period. There are different screening results: false positive, true positive and negative. If Treatment Rule 1 (i.e., always treat) is selected, then all patients with a true positive mammogram will leave the model, so there are only two paths: false positive and negative. The complexity of the problem is 2^m (without the path for a true positive). However, since a false positive will reset the occupancy distribution to [1 0 0], the complexity is simplified to $2m + 1$(including the path for treatment), which is linear in m. Similarly for Treatment Rule 2 (i.e., wait until diagnosed at the invasive stage), there are three possible paths including the case of true positive since the woman will not leave the model from the in situ stage. The in situ true positive will also reset the occupancy distribution to [0 1 0], and thus the complexity of the problem will reduce from 3^m to $4m$. Numerical experiments using the methods described above are presented in the next section.

4.4 Numerical Experiments

The two existing screening guidelines are selected for the policy evaluation study: the American Cancer Society (i.e., annual screening from age 40), and the U.S. Preventive Services Task Force (i.e., biennial screening between age 50 and 74) guidelines. As discussed earlier in Chapter 1, in 2009 the USPSTF recommended less frequent screening in an attempt to address the concern regarding over-diagnosis. The goal of this policy

evaluation study is to quantify the impact of regression on patient outcomes under these two screening policies.

4.4.1 Data and assumptions

The numerical study begins with a 25 year old, cancer-free woman, and ends at age 100. The decision epochs are assumed to be every 6 months. The transition probability matrix updates every 5 years. In other words, the transition probability matrix is the same for a woman between the ages of 25 and 29, and a different matrix will be used for women aged 30 to 34.

The regression probability can be extracted from either the self-loop transition probability or the progression probability. In the transition probability matrix, the following relationship exits.

$$p_{21}(a_t) + p_{22}(a_t) + p_{23}(a_t) = 1 \tag{4.9}$$

Let u and v be the proportions to be extracted from the exiting probabilities. Then,

$$\begin{aligned} p'_{21}(a_t) &= u \cdot p_{22}(a_t) + v \cdot p_{23}(a_t) \\ p'_{22}(a_t) &= (1-u) \cdot p_{22}(a_t) \\ p'_{23}(a_t) &= (1-v) \cdot p_{23}(a_t) \\ 0 \le u, v &\le 1 \end{aligned} \tag{4.10}$$

The parameters used in the experiments are from Maillart et al. [100], except the treated lifetime breast cancer mortality probability by cancer stage at detection. The methods and results discussed in Chapter 3 are used to estimate these probabilities. The CMR data is used with 22,328 breast cancer cases with known age and cancer stage, 1,435 deaths from

breast cancer, and 1,890 deaths from other causes. A backward calculation is developed to estimate the lifetime mortality probabilities.

Specifically, patients over 85 years old are assumed to have the same lifetime breast cancer mortality probability. For this age group, the mortality probabilities at 15 years associated with detection in the in situ and invasive stages are calculated as described in Chapter 3, and they are considered to be the lifetime mortality probabilities. For each of the other age groups, a_n, the lifetime breast cancer mortality probability is estimated using the following recursive function.

$$r_i(a_n) = \text{P(5-year mortality)} + [1-\text{P(5-year mortality)}]^* \, r_i(a_{n+1}) \qquad (4.11)$$

where $i = 1$ indicates that cancer was diagnosed in the in situ stage and $i = 2$ indicates diagnosis in the invasive stage. The estimated five-year mortality probability is calculated using the method discussed in Chapter 3. The following table contains the estimates for the lifetime breast cancer mortality probability for each age group.

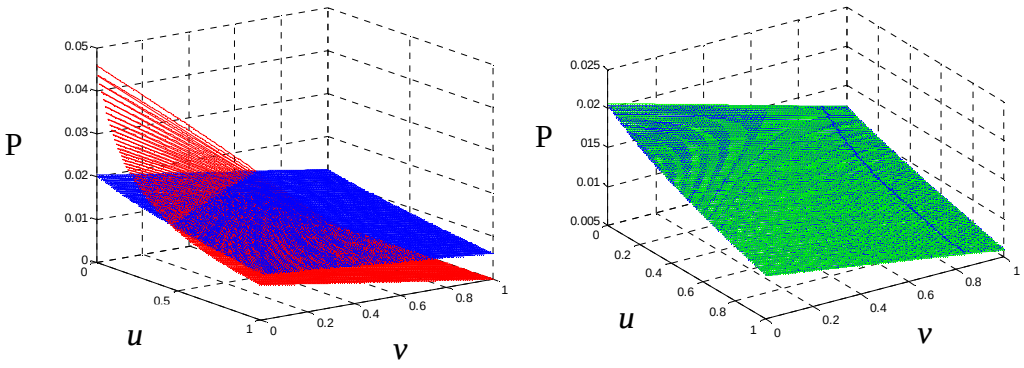
Table 4.1 Estimate for lifetime breast cancer mortality probability

Age Group	r_1	r_2
25-29	0.13657	0.68533
30-34	0.13657	0.63111
35-39	0.13657	0.58458
40-44	0.12520	0.53218
45-49	0.12169	0.49751
50-54	0.11102	0.45382
55-59	0.11102	0.41585
60-64	0.10475	0.36916
65-69	0.10475	0.32890
70-74	0.10116	0.29506
75-79	0.09845	0.24037
80-84	0.09210	0.18407
>= 85	0.08200	0.11600

The results for the numerical experiments are discussed in the next section.

4.4.2 Results

The relationship between the regression probability and lifetime breast cancer mortality probability is explored in this section. The results for the two screening policies are discussed separately. The mortality estimates for the three treatment decision rules are compared. For each policy, both 3-dimentional and 2-dimentional results are presented. While 2-D results only show the relationship between regression probability and breast cancer mortality, the 3-D plot shows the effect of the regression probability extraction (i.e., from self-loop or progression probability) on this relationship.



Note: Blue area: always treat; red area: wait till invasive stage; green area: wait once
P = Lifetime breast cancer mortality probability, *u* = proportion extracted from self-loop
v = proportion extracted from progression

Figure 4.3 3D results for the ACS Policy

The 3D results for the ACS policy are shown in Figure 4.3. The plot on the left compares the breast cancer mortality under Treatment Rule 1 (always treat) to that under

Treatment Rule 2 (wait until invasive). When both u and v are small, the "wait" decision is worse than the "treat" decision as the mortality probability is high. However, as the proportion gets larger, the mortality probability reduces and the two surfaces cross over so that the wait decision is better than the treat decision (this crossover occurs at extreme points when $u = 0.45$ and $v = 0.7$).

The plot on the right compares the breast cancer mortality between the "always treat" decision with "wait once". The two surfaces are very close to each other, which shows there is almost no difference between the two decision rules. This implies that waiting once in the in situ stage will not significantly affect the survival compared to the treat rule. Further, the cost associated with mammography screening is much lower than the treatment cost. Thus, if cost is considered in the decision, a woman may benefit more from waiting once at in situ stage.

The 2D result comparing all the three decision rules is shown in Figure 4.4.

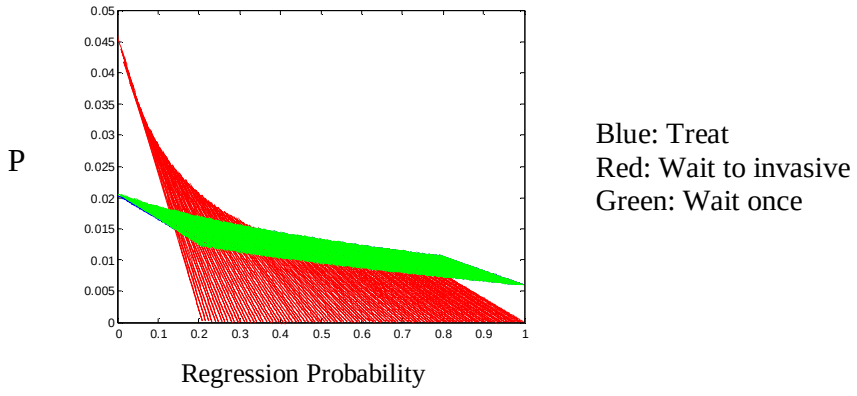
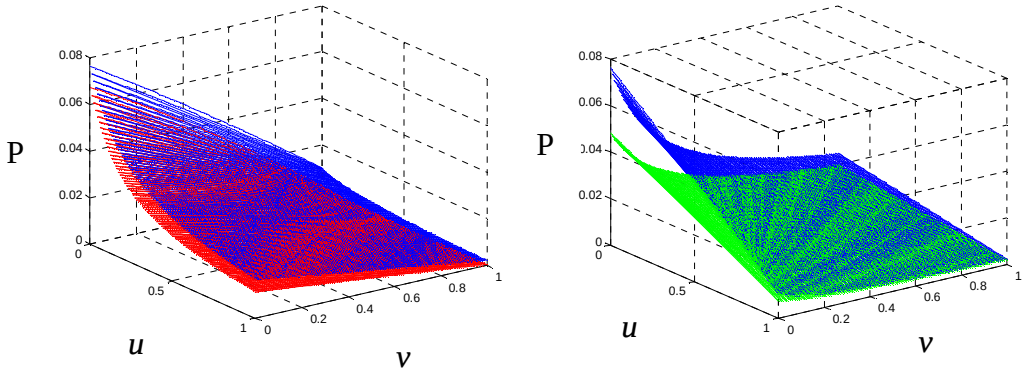


Figure 4.4 2D Result for ACS Policy

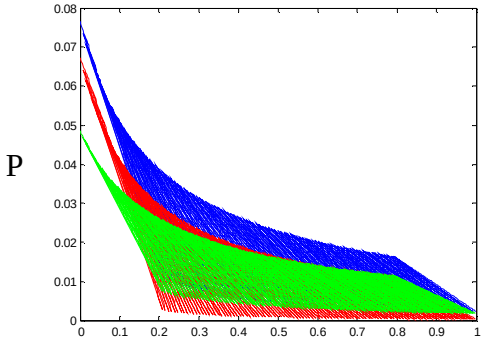
Consistent with the previous discussion, Figure 4.4 shows that the transition from the "treat" to the "wait" rule for the lifetime breast cancer mortality probability occurs when the regression probability is around 0.2. The "treat" rule is not significantly different from the "wait once" rule with respect to the mortality probability. It is worth noting that Zahl et al. [99] found an approximately 22% regression probability in their observational study.

The 3D and 2D results for the USPSTF policy are shown in Figure 4.5 and 4.6.



Note: Blue area: always treat; red area: wait till invasive stage; green area: wait once
P = Lifetime breast cancer mortality probability, *u* = proportion extracted from self-loop
v = proportion extracted from progression

Figure 4.5 3D results for USPSTF Policy



Blue: Treat
Red: Wait to invasive
Green: Wait once

Figure 4.6 2D Result for USPSTF Policy

In contrast to the ACS policy, in most cases, the "always treat" decision rule has a worse survival probability compared to the "wait" rules even when the regression probability is small. An explanation for this is that when the observation period is longer (i.e., when patients diagnosed with in situ cancers do not leave the model), more women may die from causes other than breast cancer so the lifetime breast cancer mortality probability is lower. Although the reduced mortality may not be completely associated with the regression probability, waiting rules may still suggest an improved survival under this screening policy.

4.5 Discussion

Understanding the natural history of breast cancer is an important open research question. Not only does it help to understand the disease, i.e., how it behaves, this understanding is critical in developing more accurate decision models for improved diagnosis and treatment. As the phenomenon of breast cancer regression has been debated in the medical community for many years, it has raised an important question regarding over-diagnosis and over-treatment. This study does not attempt to find a specific regression probability. Instead, this research seeks to determine whether and how regression may actually impact screening and treatment outcomes and decision making. In this decision model, the regression of breast cancer from the in situ stage to the cancer-free state is allowed, and the impact on breast cancer mortality is examined. Different screening and treatment rules are also incorporated for the analysis.

Regression is found to impact breast cancer mortality. Under a more frequent screening, policy (ACS), when in situ cancer can regress with a probability greater than 20%, it may be more beneficial to wait to treat; while for less frequent screening, waiting always results in lower lifetime breast cancer mortality. If the cost of treatment and screening are considered, these results suggest that it may be more beneficial to wait once at the in situ stage but continue screening until the cancer is detected second time. Waiting once does not increase breast cancer mortality, but may reduce unnecessary treatment.

There are some extensions to this research that can be made in the future. First, biopsy may not be necessary if cancer can actually regress. Currently a perfect biopsy is assumed to be ordered in order to identify if an abnormal mammogram reading is a true positive or false positive, and the cancer stage is determined by the biopsy result. However, biopsy may affect future screening, so the assumption regarding independence among mammograms may need to be studied. In the future, the biopsy decision could be incorporated into the modeling.

Secondly, in the current analysis, two existing policies are selected for comparison. There are more dynamic screening policies under which women can change their screening intervals. For example, when there is an abnormal screening result and the woman decides to wait for treatment, instead of waiting until the next scheduled screening, she may decide to screen more frequently. In such cases, the complexity of the problem increases greatly. At each decision epoch, there is more than one action a woman may take (in the current model, at each period n, only one action is selected). Then, the complexity of the problem is no

longer linear in the number of mammograms. An area for future research is to model dynamic screening strategies.

Lastly, in the current study many parameters are estimated in the literature, with the exception of the breast cancer mortality estimation which uses the model developed in this dissertation. The results may depend on the parameters, and estimation from different data sources may be a problem. Another area for future research is to develop a method for transition probability estimation using the same CMR data.

Chapter 5 Conclusions and Future Research Direction

Breast cancer is a complex disease that is associated with high mortality risks. This dissertation addresses three main issues of breast cancer when the disease risk is not the same for everyone. In the second chapter, the impact of comorbidity on breast cancer patient outcomes is studied. It is common to find more than one disease or disorder in women (this is particularly true in older women) where some comorbid conditions may be correlated with each other. In this study, hypertension, diabetes, mental disorder and obesity are studied in breast cancer patients. It is found that these comorbid conditions do impact patient outcomes. Thus, when decision models for breast cancer patients are developed, comorbidity should be considered as well.

The third chapter models breast cancer mortality when mortality risk is compared with respect to different breast cancer risk factors. It is generally known that the incidence rates of breast cancer for women with different backgrounds may be different, but there has been limited study on how these risk factors may affect breast cancer mortality. In most decision analyses, mortality is considered the same for women at the same age. However, age and race alone may not be enough to differentiate mortality for women with different characteristics. In order to develop more personalized screening policies, mortality probabilities for the complex population are necessary. Breast density, estrogen and progesterone receptor status, and family history are three under-studied breast cancer risk factors and tumor characteristics selected for the analysis. In addition to non-parametric point

estimates, confidence intervals are estimated to enable comparison between different groups. The methods developed may be extended to other applications.

The impact of breast cancer natural history on screening and treatment decision making is studied in the fourth chapter. This chapter studies the affect of allowing cancer to spontaneously regress. Decision models are developed incorporating different screening policies and treatment rules to study the impact on breast cancer mortality. Lifetime breast cancer mortality probabilities are estimated using the methods developed in Chapter 3. It is shown that when breast cancer can regress with a probability as low as 20%, it may be better to wait for treatment. When cost is considered in the decision, it may be worthwhile to screen more before a treatment decision is made.

The three studies have shown that breast cancer does not develop in the same way for everyone, and the risks associated with breast cancer are not the same for every woman. The current screening guidelines may not be sufficient for women with different backgrounds. A more personalized screening recommendation is needed to achieve individual maximum benefit. Comorbidities, breast cancer risk factors, breast cancer regression are issues a more individualized screening policy should consider.

Thus, this dissertation recommends the following directions for future research.

1. Modeling natural history of breast cancer at an individual level

The transition probabilities among cancer stages may not be the same for everyone. For example, breast cancer may progress slowly for women with other chronic conditions, as these conditions may be well managed, and the biology of breast cancer may be different under such an environment. Clinical studies for observing such natural progression may be

difficult to conduct, so an analytical study is needed to have a more reasonable estimate of the disease development.

Mortality estimates should continue to be explored with additional risk factors and patient demographics included. One limitation of this dissertation research is the lack of additional data for model validation. There could be more collaboration among different cancer sites in the future for validation.

2. Optimal treatment rules incorporating breast cancer regression

Following this dissertation, another area for future work is to develop a partially observable Markov decision model to identify the optimal treatment decisions for breast cancer patients. Cost needs to be incorporated to determine if continuing screening is more beneficial than starting treatment immediately when the cancer is detected at an early stage, and with a probability of regression. Sensitivity analysis should be conducted to test how large the regression probability must be in order to change the optimal decision.

When there is a biological test available to test if the cancer can regress or not, another research question is to determine the value of such test. Currently there are some tests to determine if a cancer is “good” or “bad”. In the future, there may be more tests available to test if the cancer may actually regress or not. Such tests may be expensive, and patient or societal willingness to pay for such tests could provide some insight for clinicians and decision makers.

3. Personalized optimal screening policy

This dissertation has addressed the importance of studying disease at an individual level. The goal of having more personalized optimal screening policies, i.e. the best screening policies based on a woman's own health history, serves as a motivating goal for this dissertation. Screening policies may not be uniform throughout the planning horizon. Screening decisions should be dynamic as the personal characteristics may be changing, and the screening result may also affect the future decisions.

References

1. American Cancer Society. The History of Cancer. Available via: http://www.cancer.org/Cancer/CancerBasics/TheHistoryofCancer/

2. National Cancer Institute. Cancer Staging. Available via: http://www.cancer.gov/cancertopics/factsheet/detection/staging

3. Smith RA, Cokkincides V, and Eyre HJ. American cancer society guidelines for the early detection of cancer. CA Cancer J Clin 2006; 56(1): 11-25.

4. American Cancer Society. Breast Cancer. Available via: http://www.cancer.org/acs/groups/content/@nho/documents/document/breastcancerpdf.pdf.

5. Edge SB, Byrd DR, Compton CC et al. AJCC Cancer Staging Manual. 7th Edition. Springer; 2009.

6. Radiology Society of North America. Mammography. Available via: http://www.radiologyinfo.org/En/Info.Cfm?Pg=Mammo.

7. Kerlikowske K, Grady D, Rubin SM, et al. Efficacy of screening mammography: A meta-analysis. JAMA 1995; 273(2):49-54.

8. Fracheboud J, Groenewoud JH, Boer R, et al. Sevety-five year is an appropriate upper age limit for population-based mammography screening. Int J Cancer 2006; 118(8):2020-2025.

9. Resnick B and McLeskey SW. Cancer screening across the aging continuum. Am J Manag Care 2008; 14(5):267-276.

10. Gail MH, Brinton LA, Byar DP et al. Prejecting individualized probabilities of developing breast cancer for white females who are being examine annually. J Natl Cancer Inst 1989; 81(24): 1879-1886.

11. Barlow WE, White, E, Ballard-Barbash R, et al. Prospective breast cancer risk predication model for women undergoing screening mammography. J Natl Cancer Inst 2006; 98(17): 1204-1214.

12. Tyer J, Duffy SW, and Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Stat Med 2004; 23(7): 1111-1130.

13. Colditz GA, and Rosner B. Cumulative risk of breast cancer to age 70 years according to risk factor status: data from the Nurses' Health Study. Am J Epidemiol 2000; 52(10): 950-064.

14. Lee SM, Park JH and Park HJ. Implications of systematic review for breast cancer prediction. Cancer Nurs 2008; 31(5):E40-46.

15. Tice JA, Cummings SR, Smith-Bindman R, et al. Using clinical factors and mammographic breast density to estimate breast cancer risk development and validation of a new predictive model. Ann Intern Med 2008; 48(5): 337-347.

16. Gail MH, Costantino JP, Pee D, et al. Projecting individualized absolute invasive breast cancer risk in African American women. J Natl Cancer 2007; 99(23): 1782-1792.

17. Chen J, Pee D, Ayyagari R, et al. Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. J Natl Cancer Inst 2006; 98(17): 1215-1226.

18. Claus EB, Risch N, and Thompson WD. Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction. Cancer 1994; 73(3): 643-651.

19. Taplin SH, Thompson RS, Schnitzer F, et al. Revisions in the risk-based breast cancer screening program at group health cooperative. Cancer 1990; 66(4): 812-818.

20. Susan G. Komen for the Cure. Breast Cancer Risk Assessment Tool. Available via: http://ww5.komen.org/BreastCancer/GailAssessmentModel.html.

21. Alagoz O, Ayer T, and Erenay FS. Operations Research Models for Cancer Screening. Encyclopedia for Operations Research and Management Science. Wiley; 2011.

22. Ivy JS. Can we do better? Optimization models for breast cancer screening. Handbook of Optimization in Medicine. Springer Optimization and Its Application; 2009; 26:1-28.

23. Chen HH, Duffy SW, and Tabar L. a Markov Chain method to estimate the tumor progression rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer screening. The Statistician 1996; 45(3): 307-317.

24. Shen Y, Zelen M, Robust modeling in screening studies: estimation of sensitivity and preclinical sojourn time distribution. Biostatistics 2005; 6(4): 604-614.

25. Plevritis SK, Salzman P, Sigal BM and Glynn PW. A natural history model of stage progression applied to breast cancer. Stat Med 2007; 26: 581-595.

26. Fryback DG, Stout NK, Rosenber MA, et al. The Wisconsin breast cancer epidemiology simulation model. J Natl Cancer Inst Monogr 2006; 36:37-47.

27. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Medical Care 1998; 36(1):8-27.

28. Chronic Disease Prevention and Health Promotion. Chronic disease - the power to prevent, the call to control: at a glance. National Center for Chronic Disease

Prevention and Health Promotion and Department of Health and Human Services, Editors: Atlanta; 2009.

29. Centers for Disease Control and Prevention. Chronic Disease Overview. 2009. Available via: http://www.cdc.gov/NCCdphp/overview.htm.

30. Suthummanon S, Omachonu VK. DRG-based cost minimization models: Applications in a hospital environment. Health Care Manag Sci 2004; 7:197-205.

31. Starfield B, Lemke KW, Bernhardt T, Foldes SS, Forrest CB, Weiner, JP. Comorbidity: Implications for the importance of primary care in 'case' management. Ann Fam Med 2003; 1(1):8-14

32. Satariano WA, Ragland DR. The effect of comorbidity on 3-year survival of women with primary breast cancer. Ann Intern Med 1994; 120:104-110.

33. CDC. United States Cancer Statistics Working Group: 1999-2005 Incidence and Mortality Web-based Report. Available from: http://www.cdc.gov/uscs.

34. American Cancer Society. Cancer Facts & Figures 2010. Atlanta: American Cancer Society; 2010.

35. Tammemagi CM, Nerenz D, Neslund-Dudas C, Feldkamp C, Nathanson D. Comorbidity and survival disparities among black and white patients with breast cancer. JAMA-J Am Med Assoc 2005; 294(14):1765-1772.

36. Brandeau ML, Zaric GS. Optimal investment in HIV prevention programs: More is not always better. Health Care Manag Sci 2009; 12:27-37.

37. Zhao SZ, Wong JM, and Arguelles LM. Hospitalization costs associated with leiomyoma. Clin Ther 1999; 21(3):563-575.

38. Livingston EH, Langert J. The impact of age and medicare status on bariatric surgical outcomes. Arch Surg 2006; 141(11):1115-1120.

39. Meguid RD, Brooke BS, Chang DC, Sherwood JT, Brock MV, Yang SC. Are surgical outcomes for lung cancer resections improved at teaching hospitals? Ann Thorac Surg 2008; 85:1015-1025.

40. Newschaffer CJ, Bush TL, Penberthy LE et al. Does comorbid disease interact with cancer? An epidemiologic analysis of mortality in a cohort of elderly breast cancer patients. J Gerontol A Biol Sci Med Sci 1998; 53(5):M372-378.

41. Yancik R, Wesley MN, Ries LA, Havlik RJ, Edwards BK, Yates JW. Effect of age and comorbidity in postmenopausal breast cancer patients aged 55 years and older. JAMA 2001; 285(7):885-892.

42. Lash TL, Fox MP, Buist DSM et al. Mammography surveillance and mortality in older breast cancer survivors. J Clin Oncol 2007; 25(21):3001-3006.

43. Farley JF, Harley CR, Devine JW. A comparison of comorbidity measurements to predict healthcare expenditures. Am J Manag Care 2006; 12:110-117.

44. Garis RI, Farmer K. Examining costs of chronic conditions in a Medicaid population. Manag Care 2002; 11:43-50.

45. Dominick KL, Dudley TK, Coffman CJ, Bosworth HB. Comparison of three comorbidity measures for predicting health service use in patients with osteoarthritis. Arthrit Care Res 2005; 53(5):666-672.

46. Piccirillo JF, Tierney RN, Costas I, Grove L, Spitznagel EL. Prognostic importance of comorbidity in a hospital-based cancer registry. JAMA 2004; 291:2441-2447.

47. De Groot V, Beckerman H, Lankhorst GJ, Bouter LM. How to measure comorbidity: A critical review of available methods. J Clin Epidemiol 2003; 56(3): 221-229.

48. Li J. An application of lifetime models in estimation of expected length of stay of patients in hospital with complexity and age adjustment. Stat Med 1999; 18(23):3337-3344.

49. Sá C, Dismuke CE, Guimaraes P. Survival analysis and competing risk models of hospital length of stay and discharge destination: The effect of distributional assumptions. Health Serv Outcomes Res Method 2007; 7:109-124.

50. Roehrig C, Miller G, Lake C, Bryant J. National health spending by medical condition, 1996–2005. Health Affair 2009; 28(2):358-367.

51. Agency for Healthcare Research and Quality. Healthcare Cost and Utilization Project (HCUP) Databases. Editor: Rockville, MD; 2008.

52. ICD-9 Codes for Breast Cancer. Available via: www.fortherecordmag.com/archives/ftr_021405p33.shtml.

53. ICD-9 Codes for Hypertension. Available via: http://www.fortherecordmag.com/archives/ftr_01232006p44.shtml.

54. ICD-9 Codes for Obesity. Available via: http://www.fortherecordmag.com/archives/ftr_012604p39.shtml.

55. Satterthwaite, FE. Synthesis of Variance. Psychometrika 1941; 16(5): 309-16.

56. Klein, JP, Moeschberger, M.L. Survival Analysis: Techniques for Censored and Truncated Data. 2nd Edition. Springer; 2003.

57. Tabachnick, BG, Fidell, LS. Using Multivariate Statistics. 5th Edition. Allyn and Bacon; 2006.

58. SAS/STAT 9.1 User's Guide, 2004: 4799-4805.

59. Barak Y, Lew T, Achiron A and Aizenberg D. Breast cancer in women suffering from serious mental illness. Schizophr Res 2008; 102(1-3):249-253.

60. Klemi PJ, Parvinen I, Pylkkänen L et al. Significant improvement in breast cancer survival through population-based mammography screening. Breast 2003; 12(5): 308-313.

61. Holmes CE, Muss HB. Diagnosis and treatment of breast cancer in the elderly. CA-Cancer J Clin 2003; 53(4):227-244.

62. Rovera F, Dionigi G, Riva C, Chiaravallib A, et al. Identifying factors contributing to reduced breast tumor size: a longitudinal study. Int J Surg 2008; 6 Suppl 1:S97-S100.

63. Kraus AS and Oppenheim A. Trend of mortality from cancer of the breast. JAMA 1965; 194(1):90-90.

64. Chu KC, Miller BA, Feuer FJ, et al. A method for partitioning cancer mortality trends by factors associated with diagnosis: An application to female breast cancer. J Clin Epidemiol 1994; 47:1451-1461.

65. Jatoi I, Chen BE, Anderson WF, and Rosenberg PS. Breast cancer mortality trends in the United States according to estrogen receptor status and age at diagnosis. J Clin Oncol 2007; 25:1683-1690.

66. Berry DA, Cronin KA, Plevritis SK, et al. Effect of screening and adjuvant therapy on mortality from breast cancer. N Engl J Med 2005; 353:1784-1792.

67. Berry DA, Inoue L, Shen Y, et al. Modeling the impact of treatment and screening on U.S. breast cancer mortality: a Bayesian approach. J Natl Cancer Inst Monogr 2006; 36:30-36.

68. Schairer C, Mink PJ, Carroll L and Devesa S. Probabilities of death from breast cancer and other causes among female breast cancer patients. J Natl Cancer Inst. 2004; 96(17):1311-1321.

69. Rosenberg MA. Competing risks to breast cancer mortality. J Natl Cancer Inst Monogr 2006; 36:15-19.

70. Lee S and Zelen M. A stochastic model for predicting the mortality of breast cancer. J Natl Cancer Inst Monogr 2006; 36:79-86.

71. Chiu SY, Duffy S, Yen AM, et al. Effect of baseline breast density on breast cancer incidence, stage, mortality, and screening parameters: 25-year follow-up of a Swedish mammographic screening. Cancer Epidemiol Biomarkers Prev 2010; 19(5):1219-1228.

72. McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. Cancer Epidemol Biomarkers Prev 2006; 5:1159-1169.

73. American Cancer Society. What are the risk factors for breast cancer? Available at: http://www.cancer.org/Cancer/BreastCancer/DetailedGuide/breast-cancer-risk-factors.

74. Carney PA, Miglioretti DL, Yankaskas BC, et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. Ann Intern Med 2003; 138:168-175.

75. Kerlikowske K, Ichikawa L, Miglioretti DL, et al. Longitudinal measurement of clinical mammographic breast density to improve estimation of breast cancer. J Natl cancer Inst 2007; 99:386-395.

76. Boyd NF, Guo H, Martin LJ, et al. Mammographic density and the risk and detection of breast cancer. N Engl J Med 2007; 356:227-236.

77. Olsen AH, Bihrmann K, Jensen MB, et al. Breast density and outcome of mammography screening: a cohort study. Br J Cancer 2009; 100:1205-1208.

78. Clark M, Collins R, Darby S, et al. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. Lancet 2005; 365:1687-1717.

79. Berry DA, Cirrincione C, Henderson IC, et al. Estrogen-receptor status and outcomes of modern chemotherapy for patients with node-positive breast cancer. JAMA 2006; 295:1658-1667.

80. Pharoah PD, Day NE, Duffy S, et al. Family history and risk of breast cancer: A systematic review and meta-analysis. Int J Cancer 1997; 71:800-809.

81. Change ET, Milne RL, Phillips KA, et al. Family history of breast cancer and all-cause mortality after breast cancer diagnosis in the Breast Cancer Family History. Breast Cancer Res Treat 2009; 117:167-176.

82. Carolina Mammography Registry. Available at: http://www.unc.edu/cmr.

83. Breast Cancer Surveillance Consortium. Available at: http://breastscreening.cancer.gov.

84. ICD-9-CM code for cause of death. Available at: http://www.ihs.gov/publicinfo/publications/trends96/96trind.pdf.

85. ICD-10-CM code for cause of death. Available at: http://www.cdc.gov/nchs/data/dvs/im9_2002.pdf.pdf.

86. Kalbfleisch, JD and Prentice RL. The Statistical Analysis of Failure Time Data. New York, NY: John Wiley; 1980.

87. Prentice RL and Kalbfleisch JD. The analysis of failure times in the presence of competing risks. Biometrics 1978; 34:541-554.

88. Pepe, MS and Mori M. Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data? Stat Med 1993; 12:737-751.

89. Gooley TA, Leisenring W, Crowley J and Storer BE. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. Stat Med 1999; 18:695-706.

90. Marubini E and Valsecchi MG. Analysing survival data from clinical trial and observational studies. Chichester, UK: John Wiley & Sons; 1995.

91. Choudhury JB. Non-parametric confidence interval estimation for competing risks analysis: application to contraceptive data. Stat Med 2002; 21:1129-1140.

92. Crooke PS. Mathematical modeling of tumor growth reference list. 2008. Available via: http://www.math.vanderbilt.edu/~pscrooke/CancerModeling.pdf.

93. Norton L. A gompertzian model of human breast cancer growth. Cancer Res 1988; 48: 7067-7071.

94. Fekjar HW, Lindqvist BH, Vatten LJ et al. Breast cancer tumor growth estimated through mammography screening data. Breast Cancer Res 2008; 10:R41.

95. Arena® Input Analyzer. Version 12. Rockwell Automation Technologies, Inc. 2006.

96. Osler W. The medical aspects of carcinoma of the breast, with a note on the spontaneous disappearance of secondary growth. Am Med 1901; 17-19.

97. Conference on spontaneous regression cancer. Natl Cancer Inst Monogr 1976; 44: 23-26.

98. Larsen S, and Rose C. [Spontaneous remission of bresat cancer: A literature review]. Ugeskr Laeger 1999; 161(26): 4001-4004.

99. Brunside SE, Trentham-Dietz A, Kelcz F., and Collins J. An example of breast cancer regression on imaging. Radiology Case Reports 2006; 1(2): 27-37.

100. Zahl PH, Mahlen J and Welch HB. The natural history of invasive breast cancer detected by screening mammography. Arch Intern Med 2008; 168(21): 2311-2316.

101. Maillart LM, Ivy JS, Ransom S, and Diehl K. Assessing dynamic breast cancer Screening policies. Oper Res, 56(6):1411-1427.

102. Weedon-Fekjaer H, Vatten LJ, Aalen OO et al. Estimating mean sojourn time and screening test sensitivity in breast cancer mammography screening: new results. J Med Screen 2005; 12(4): 172-178.

103. Baker JA, Kornguth PJ, Williford ME and Floyd CE. Breast cancer: Prediction with artificial neural network based on BI-RADS standardized lexicon. Radiology 1995; 817-822.