POST-DIAGNOSIS WEIGHT CHANGE, PHYSICAL ACTIVITY, AND SURVIVAL AMONG WOMEN WITH BREAST CANCER: A LONGITUDINAL STUDY WITH MISSING DATA

PATRICK T. BRADSHAW

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Epidemiology.

Chapel Hill
2009

Approved by
Marilie D. Gammon, Ph.D. (Advisor)
Joseph G. Ibrahim, Ph.D.
Robert C. Millikan, D.V.M., Ph.D.
June Stevens, Ph.D.
Jessie A. Satia, Ph.D.
ABSTRACT

Patrick T. Bradshaw: Post-diagnosis weight change, physical activity and survival among women with breast cancer: a longitudinal study with missing data
(Under the direction of Marilie D. Gammon)

Nearly forty thousand women per year are diagnosed with breast cancer and there are currently over 2 million female breast cancer survivors in the United States. Whether survival after breast cancer diagnosis is influenced by modifiable lifestyle factors, including post-diagnosis weight change and physical activity, is unclear. These associations were examined using data from a population-based follow-up study of 1,508 women diagnosed with first primary in situ or invasive breast cancer between August 1, 1996 and July 31, 1997 in Long Island, New York. During baseline and follow-up interviews, women self-reported their height, weight, recreational physical activity levels and other factors. Additional information on clinical factors was ascertained through medical records and the New York State Cancer Registry. Vital status was determined using the National Death Index; through the end of 2005, 308 women were deceased, with 164 due to breast cancer. Approximately one-third of the subjects did not complete the follow-up interview. To address the issue of potentially non-ignorably missing data, I developed a selection model for survival analysis with time-varying covariates. A sensitivity analysis using the data on post-diagnosis weight change illustrated that a standard analysis resulted in reduced statistical efficiency and differences in magnitude of effect when compared to the selection model. Mortality was positively associated with both post-diagnosis weight loss and weight gain, regardless of the
time since diagnosis or pre-diagnosis body size. More detailed analyses showed that previously reported associations of mortality with pre-diagnosis body mass index (BMI) and adult weight change were attenuated after accounting for post-diagnosis weight change, while associations with post-diagnosis weight change remained. Mortality was inversely associated with recreational physical activity, regardless of pre-diagnosis activity levels, timing of post-diagnosis activity, or pre-diagnosis BMI. Since weight gain and reduction in physical activity are common after breast cancer diagnosis, these findings that suggest weight maintenance and physical activity enhance survival among breast cancer survivors may be especially important.
To my parents and grandparents, who instilled in me the value of education, hard work and perseverance.

To my aunts, Antonetta “Nanny” Scribner and Janice Foster, who were taken long before their time.

To Robbie. My drive to succeed is only surpassed by your belief in me.
ACKNOWLEDGMENTS

So many people had a part in the completion of this project it is difficult to adequately express my gratitude succinctly. My committee’s careful eyes and thoughtful feedback has made this work something I am deeply proud of. I am especially grateful to my advisor, Marilie Gammon, whose enthusiasm for what she does and dedication to her students is inspiring. She has taught me how to think critically, express myself clearly and simply “get things done.” I like to think that the majority of my education here went beyond the coursework, and she is responsible for much of it.

The support and encouragement of my parents and grandparents throughout the years has played a tremendous role in this achievement. Growing up it was assumed that I would pursue higher education, but I don’t think anyone had an idea that it would take me on such a journey. The one constant throughout this process has been their pride in me. A significant thanks is also owed to my “dissertation support group”, Tania Desrosiers, Sarah Cohen and Sarah Nyante. They kept me on track when developing the proposal for this project, and offered understanding that only those going through the dissertation process can.

Last, and certainly not least, the completion of this work is due in no small part to Robbie Briggs and his boundless support and encouragement. Besides cheering me on when I need it most, he helps keep me focused on what is really important.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>ABSTRACT</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgments</td>
<td>vi</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>vii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xiii</td>
</tr>
</tbody>
</table>

## BACKGROUND

- Epidemiology of Breast Cancer Incidence ................................................... 1
- Etiology of Breast Cancer .............................................................................. 1
- Breast Cancer Risk Factors ........................................................................... 2
- Epidemiology of Breast Cancer Survival ....................................................... 4
  - Established Prognostic Factors ................................................................. 4
- Energy Balance and Breast Cancer: Biologic Mechanisms ............................. 7
  - Obesity and Carcinogenesis .................................................................... 7
  - Physical Activity and Carcinogenesis ....................................................... 10
- Epidemiology of Obesity and Breast Cancer ................................................ 12
  - Obesity and Breast Cancer Incidence ....................................................... 13
  - Obesity and Breast Cancer Survival ......................................................... 15
- Epidemiology of Physical Activity and Breast Cancer .................................. 19
- Physical Activity and Breast Cancer Incidence ........................................... 20
- Physical Activity and Breast Cancer Survival ............................................. 21
Summary ........................................................................................................................................... 25
References ............................................................................................................................................. 31

RESEARCH METHODS ......................................................................................................................... 52
Overview .................................................................................................................................................. 52
Parent Study ........................................................................................................................................... 53
Follow-up Study ...................................................................................................................................... 54
Study Population ..................................................................................................................................... 54
Description of Cohort of Women with Breast Cancer .............................................................................. 54
Outcome Assessment ............................................................................................................................... 55
Exposure Assessment .............................................................................................................................. 56
Results from Previous Analyses .............................................................................................................. 59
Data Analysis ........................................................................................................................................... 61
Variable Definitions for Outcome and Exposures ..................................................................................... 61
Missing Data: Background .......................................................................................................................... 62
Proportional Hazards Models Under Complete Data .............................................................................. 63
Selection Models for Survival Data with Missing Covariates ................................................................... 64
Confounding ............................................................................................................................................. 72
Effect Measure Modification .................................................................................................................... 74
Sensitivity Analysis ................................................................................................................................... 75
Study Power ............................................................................................................................................. 76
Strengths and Limitations .......................................................................................................................... 77
Summary ................................................................................................................................................. 80
References ............................................................................................................................................... 85
Introduction..................................................................................................................137
Methods......................................................................................................................137
Results.........................................................................................................................144
Discussion...................................................................................................................147
References...............................................................................................................158

DISCUSSION................................................................................................................162
Summary.....................................................................................................................162
Missing Data...............................................................................................................164
Study Design...............................................................................................................166
Exposure Assessment.................................................................................................167
Study Limitations.......................................................................................................168
Future Directions.......................................................................................................169
Conclusion..................................................................................................................171
References...............................................................................................................173

WinBUGS CODE.........................................................................................................177
References...............................................................................................................181
**LIST OF TABLES**

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Table 2.1.</strong> Selected characteristics of breast cancer cases at diagnosis (n=1,508), Long Island Breast Cancer Study Project, 1996-1997.</td>
<td>81</td>
</tr>
<tr>
<td><strong>Table 2.2.</strong> Power to detect varying hazard ratios for weight change, calculations based on the log-rank test statistic, assuming 5% significance level, complete data (n=1033 total; maintenance n=489; highest gain n=200) and full sample (n=1436 total; maintenance n=679; highest gain n=278).</td>
<td>84</td>
</tr>
<tr>
<td><strong>Table 2.3.</strong> Power to detect varying hazard ratios for post-diagnosis levels of physical activity categorized to three levels, calculations based on the log-rank test statistic, complete data (n=1033 total; inactive: 0 MET-hours/week n=171; highest activity: &gt;9.0 MET-hours/week n=416) and full sample (n=1423 total; inactive: 0 MET-hours/week n=236; highest activity: &gt;9.0 MET-hours/week n=573).</td>
<td>84</td>
</tr>
<tr>
<td><strong>Table 3.1.</strong> Coefficient estimates (posterior log-hazard ratios) and 95% credible intervals from piecewise exponential proportional hazards model for all-cause mortality in the Long Island Breast Cancer Study Project under different missing data assumptions.</td>
<td>107</td>
</tr>
<tr>
<td><strong>Table 3.2.</strong> Coefficient estimates (log-hazard ratios) and 95% credible intervals from piecewise exponential proportional hazards model for breast cancer mortality in the Long Island Breast Cancer Study Project under different missing data assumptions.</td>
<td>107</td>
</tr>
<tr>
<td><strong>Table 4.1.</strong> Characteristics of 1,436 women newly diagnosed with a first primary breast cancer between 1996-1997 on Long Island, NY, with follow-up assessments between 2002-2004.</td>
<td>128</td>
</tr>
<tr>
<td><strong>Table 4.2.</strong> Hazard ratios (and 95% credible intervals) for the association between post-diagnosis changes in body weight and all-cause and breast cancer-specific mortality among women newly diagnosed with a first primary breast cancer in 1996-1997 in Long Island, NY and followed through 2005.</td>
<td>130</td>
</tr>
<tr>
<td><strong>Table 4.3.</strong> Hazard ratios (and 95% credible intervals) for the association between post-diagnosis changes in body weight and all-cause and breast cancer-specific mortality, stratified by time, pre-diagnosis BMI and pre-diagnosis adult weight change among women newly diagnosed with a first primary breast cancer in 1996-1997 on Long Island, NY, and followed through 2005.</td>
<td>131</td>
</tr>
<tr>
<td><strong>Table 5.1.</strong> Characteristics of 1,436 women diagnosed with a first primary breast cancer in Long Island, NY, with follow-up assessments in 2002-2004.</td>
<td>152</td>
</tr>
<tr>
<td><strong>Table 5.2.</strong> Hazard ratios (and 95% credible intervals) for the association between all-cause and breast cancer-specific mortality, and average post-diagnosis physical activity levels</td>
<td></td>
</tr>
</tbody>
</table>
(MET-hrs/wk) over the entire follow-up period, among a cohort of women with a first primary breast cancer on Long Island, NY, and followed through December 31, 2005.

**Table 5.3.** Hazard ratios (and 95% credible intervals) for the association between all-cause and breast cancer specific mortality and average post-diagnosis physical activity levels (PA) over entire follow-up period, stratified by pre-diagnosis physical activity (PA) levels, among women diagnosed with a first primary breast cancer in 1996-1997 on Long Island, NY, and followed through December 1, 2005.

**Table 5.4.** Hazard ratios (and 95% credible intervals) for the association between all-cause and breast-cancer specific mortality and post-diagnosis physical activity levels (met-hrs/wk) assessed yearly over entire follow-up, among women diagnosed with a first primary breast cancer in 1996-1997 on Long Island NY, and followed through December 31, 2005.

**Table 5.5.** Hazard ratios (and 95% credible intervals) for the association between all-cause and breast cancer-specific mortality and yearly post-diagnosis physical activity (PA) levels, stratified by time since diagnosis and pre-diagnosis BMI, among women diagnosed with a first primary breast cancer in 1996-1997 on Long Island, NY, and followed through December 31, 2005.
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1. Hypothesized biological mechanisms for the relationship between adiposity and breast cancer survival among breast cancer patients.</td>
<td>29</td>
</tr>
<tr>
<td>Figure 1.2. Hypothesized biological mechanisms for the relationship between physical activity and breast cancer survival among breast cancer patients.</td>
<td>30</td>
</tr>
</tbody>
</table>
CHAPTER 1
BACKGROUND

Epidemiology of Breast Cancer Incidence

In 2005 the greatest number of new cancer cases among women was breast cancer with over 260,000 new cases, accounting for nearly a third of all cancers diagnosed that year (1). A sharp increase in yearly incidence rates of invasive breast cancer was observed between 1980 and 1987, due in part to improved use of mammographic screening and overall breast cancer awareness, but the yearly increases in incidence became much more stable between 1987 and 2002 (2). Age is the factor most strongly associated with breast cancer incidence with the median age at diagnosis between 1987 and 2002 being 61 years and most new breast cancer diagnoses occurring among women over 40 years. Breast cancer incidence rises across the entire age range reaching its apex by 75-79 years, however it increases with age more rapidly before menopause than after (3). This pattern is likely reflective of the hormonal mechanisms that are believed to play a role in breast carcinogenesis (4) as endogenous estrogen exposure is higher during the reproductive years. A positive association between breast cancer incidence and socioeconomic status has also been noted (5) as more affluent women have a higher risk of being diagnosed.

Etiology of Breast Cancer

The vast majority of breast cancers are adenocarcinomas which arise from the glandular tissue of the breast (6). In situ breast cancers are those malignancies that are confined to the ducts or lobules of the breast and are less common than invasive cancers, those that infiltrate the surrounding stroma. Hormones such as estrogen and progesterone are believed to play a
significant role in the etiology of breast cancer given their effect on cellular proliferation (7). Estrogen receptor (ER) positive or progesterone receptor (PR) positive tumors have relatively large numbers of the respective hormone receptors and are thought to generally have a better prognosis and specifically respond more favorably to hormone therapies compared to ER negative or PR negative tumors (8).

**Breast Cancer Risk Factors**

As estrogen plays a significant role in mammary cell carcinogenesis many established risk factors for breast cancer are related to reproductive characteristics, such as early menarche (5), later age of menopause (5, 7), age at first birth (5), parity (9) and lactation history (10). Early menarche or later age of menopause increases the length of time a woman’s body is exposed to endogenous estrogens and consequently increases her risk for breast cancer. In addition to their hormonal impact, pregnancy and lactation are thought to reduce risk by affecting differentiation of mammary cells, making them less susceptible to the DNA damage required for carcinogenesis (5). Use of exogenous hormones, such as oral contraceptives (11) or hormone replacement therapy (12) has also been associated with increased risk of breast cancer.

Many modifiable lifestyle factors have been evaluated for their association with breast cancer risk, notably alcohol consumption (13) and certain dietary behaviors (14). Alcohol intake has been consistently positively associated with breast cancer risk (6); the plausible mechanism is that alcohol increases in estrogen exposure (15) as well as folate antagonism (16), the latter thought to increase risk through impairment of DNA methylation and repair (17). The associations with dietary factors are less consistent (14, 18), however the evidence is intriguing for several exposures: intake of cooked and smoked meats has been
associated with an increase in the risk of breast cancer (14) while folate intake may be associated with a lower risk (17).

Factors related to energy balance, such as anthropometric measures of body size and physical activity levels, have been consistently associated with breast cancer risk and are emerging as important issues in cancer prevention. The relationship between body size and breast cancer risk varies by menopausal status with greater body size conveying a reduction in risk for premenopausal women while increasing risk for postmenopausal women (19, 20). Specifically, the reduction in relative risk for premenopausal women of heavier weight or higher body mass index (BMI, the ratio of weight in kg to squared height in meters) category is on the order of 40% while heavier or larger postmenopausal women have approximately 25% greater relative risk of breast cancer (21, 22). The association with adult weight gain has proven even more consistent than associations with BMI or body weight, with effects of similar magnitudes observed in cohort studies even when null associations for BMI or body weight were observed (20-22). The association of breast cancer risk with anthropometric measures of fat distribution, such as waist-to-hip ratio (WHR), have been less consistent among premenopausal women but generally support a positive association with risk among postmenopausal women (20). Physical activity has also shown fairly consistent associations with breast cancer risk, with relative risks ranging between 0.80 and 0.20 comparing the most physically active to least physically active in cohort and case-control studies (23). A plausible mechanism for these associations exists as energy balance is related to endogenous exposure to estrogen and insulin, both mitogenic hormones that are believed to increase the risk of breast cancer by functioning as tumor promoters (24, 25). The epidemiologic
evidence and specific biologic mechanisms surrounding the observed associations between body size, physical activity and breast cancer risk will be discussed in detail below.

**Epidemiology of Breast Cancer Survival**

Breast cancer claimed the lives of an estimated 40,410 women in 2005 accounting for 15% of the female cancer related deaths that year, second only to lung cancer. Survival after a diagnosis of breast cancer has improved over the last three decades with 88% of women surviving to 5 years by the year 2000 (1). Between 1990 and 2002 the death rate from breast cancer has decreased by 2.3% annually, which is believed to be due to earlier detection and improvements in treatment (2). As of 2004 there were over 2 million female breast cancer survivors in the United States (3) and there has been growing interest in identification of modifiable prognostic factors in an effort to improve morbidity and mortality.

**Established Prognostic Factors**

The most reliable predictors of breast cancer survival include age, socioeconomic status, race, pathological features of the tumor and clinical characteristics. Younger age is associated with poorer prognosis (26, 27), reflective of the fact that tumors of younger women often have histopathologic characteristics associated with poorer prognosis (28). Despite the fact that incidence rates for breast cancer are higher among white women, African-American women are less likely to survive, with approximately one-quarter dying within five years of diagnosis compared to only ten percent of Caucasian women (3). This association has been thought to be driven by socioeconomic status as women of lower socioeconomic status tend to have poorer outcomes after diagnosis of breast cancer regardless of race (29).

Pathological features such as tumor size, metastases to axillary lymph nodes and stage at diagnosis are strongly associated with breast cancer recurrence and survival, since
higher stage, greater tumor size and involvement of axillary lymph nodes generally signal more advanced or aggressive disease. Tumor size is one of the most important prognostic indicators of breast cancer survival (30) as studies have shown that women with tumors with a diameter of 2 centimeters (cm) or less have better survival compared to those with tumors greater than 2 cm (26). The metastatic involvement of axillary lymph nodes is also a significant predictor of survival (31) as involvement of a greater number of lymph node metastases indicates a more advanced disease that has spread to other areas of the body. Women with lower stage of disease have a markedly better prognosis, with 100% of those with in situ breast carcinomas surviving 5 years post-diagnosis, but only 88.6% of all women with invasive tumors surviving as long (3). Among women with invasive breast cancer, 5-year survival rates for those with localized tumors are high (98.0%) with 5-year survival progressively decreasing for those presenting with regional and distant carcinoma (83.5% and 26.7%, respectively).

Clinically, treatment regimen is associated with survival with surgery, radiation therapy, chemotherapy and hormone therapy among the standard options (32-36). Hormonal therapy, specifically the use of anti-estrogenic drugs such as Tamoxifen, is commonly used and has proven to provide a significant improvement in survival (37). A recent review of randomized clinical trials reports a 31% reduction in the annual death rate among women with ER+ tumors on tamoxifen (38), lending additional evidence to the effect of estrogen on both breast cancer risk and survival. Other reports note that response to treatment can be influenced by tumor characteristics, with hormone receptor negative tumors (ER- or PR-) being less responsive to the effects of endocrine therapy (39). Chemotherapy is also a very effective treatment as it has been shown to reduce the annual death rate by 38% for women
younger than 50 years and 20% for women older than 50 years of age (38). Radiation therapy is used in conjunction with surgery and chemotherapy, where it has proven effective at reducing local and regional recurrence and improving breast cancer related survival (40), however a recent review has reported an increase in vascular-related deaths (41). Tumor markers associated with treatment efficacy include expression of the p53 tumor suppressor gene (42) and HER-2/neu overexpression (43) but their association with survival has not been definitively established. Although clinical factors are important indicators of prognosis, they have been shown to explain only about 20% of the variability in breast cancer mortality (44).

Although established risk factors for breast cancer would seem to be plausible prognostic indicators, to date, few modifiable lifestyle factors have shown consistent associations with breast cancer recurrence and survival (45, 46). Estrogen exposure does appear to influence prognosis, as evidenced by the successful use of anti-estrogenic drugs for treatment (35); further some studies, although not all, have observed an increase in breast cancer mortality among recent users of oral contraceptives or hormone replacement therapy (47). Additional evidence for the deleterious effect of estrogen on survival is noted from the literature on reproductive history and survival, which been extensively studied given the hormonal changes associated with pregnancy and since these factors have been consistently associated with incident breast cancer. The reproductive factor most consistently associated with survival is recent birth, with studies finding that women who give birth within 1-5 years before diagnosis have a significantly reduced chance of survival (48-52). Several studies have also noted that parity is inversely associated with survival with an increase in mortality associated with a greater number of births (48, 50, 51, 53). The prognostic value of other
reproductive characteristics, such as gravidity, age at menarche, age at first birth, breastfeeding history and abortion history is less clear (51, 54). Alcohol consumption has not been associated with survival (55) and the results for smoking have been inconclusive (56). Studies of diet and breast cancer survival have not yielded consistent results, however a recent review of the available evidence suggests that increasing fat intake may be related to poorer survival while increasing fruit, vegetable, micronutrient and fiber intake may reduce mortality (55). Of the modifiable behaviors that are hypothesized to affect breast cancer survival, factors related to energy balance appear to hold great promise (45). Many of the factors associated with positive energy balance, particularly elevations in hormones that promote tumor growth and decrease apoptosis, are believed to influence survival by encouraging the development of existing tumors among patients diagnosed with breast cancer (57).

**Energy Balance and Breast Cancer: Biologic Mechanisms**

Energy balance represents the state of equilibrium between energy intake and expenditure and is determined by the interaction between diet, body size and composition, genetics and physical activity (58, 59). Positive energy balance indicates a condition in which energy intake exceeds energy expenditure and is reflected by an increase in adipose tissue; a high level of body fat is believed to be a significant cause of morbidity and mortality in the developed world (60-62).

**Obesity and Carcinogenesis**

Adipose tissue, especially visceral fat, is metabolically active (63) and has a number of physiological corollaries that are believed to influence the etiology of several cancers, including breast (figure 1). The association between obesity and sex hormones and related binding proteins, especially estradiol and sex hormone binding globulin (SHBG), are thought
to play a significant role in mammary carcinogenesis (64). Aromatization of androgens in adipose tissue yields estrone which is subsequently converted to estradiol, the most metabolically active estrogen (65). This pathway represents the primary source of estrogen for postmenopausal women, in contrast to premenopausal women where ovarian production of estradiol overshadows adipose-mediated formation (66). In addition, estrogen production in premenopausal women is tightly regulated through feedback loops, however adipose-mediated estrogen production is largely unregulated making it a significant source of estrogen exposure, particularly among postmenopausal women (66). The availability of estradiol to target tissues is primarily determined by the amount of circulating SHBG. Approximately half of the estradiol in the blood is bound to SHBG, the remainder bound to albumin or freely circulating (66). A common consequence of obesity-related hyperinsulinemia is a reduction in SHBG resulting in an increase in bioavailable estrogen allowing more free or albumin-bound estradiol to bind with estrogen receptors (64). The combined effect of unregulated estradiol exposure and reduction in SHBG results in a greater than two-fold increase in free estradiol among obese postmenopausal women compared to women of normal weight (66). Recently McTiernan and colleagues found that body mass index (BMI) was positively associated with estrone, estradiol, free estradiol, free testosterone and prolactin and negatively associated with SHBG among healthy postmenopausal women (67). As noted previously, sex hormones are powerful mitogens which stimulate cellular proliferation therefore increasing the likelihood of a DNA mutation during cell division and encouraging replication of aberrant cells (68); breast tissue is especially sensitive to the effects of these hormones (69, 70). There is also some evidence that estrogen metabolism
generates free radicals which may inflict DNA damage thereby initiating carcinogenesis (71, 72).

Obesity is also positively associated with increased levels of insulin and insulin like growth factors (73, 74) which can encourage proliferation of both normal and cancerous mammary cells (19, 24, 74). Insulin, a peptide hormone secreted by the beta cells of the pancreas, functions primarily to regulate blood glucose levels and inhibit breakdown of adipose tissue. An increase in fat mass is associated with elevated levels of serum free fatty acids through several mechanisms that encourage lipolysis: visceral adipose tissue is less sensitive to the antilipolytic effect of insulin and more sensitive to the lipolytic effects of catecholamine (75) and it is also known to secrete a variety of cytokines including the lipolysis stimulating tumor necrosis factor alpha (TNFα) (76). This increase in free fatty acids is thought to inhibit insulin’s effect on glucose uptake and oxidation (77) thereby resulting in a state of insulin resistance (78), and a subsequent compensatory increase in insulin secretion by the pancreas in an effort to maintain glucose homeostasis. This increase in insulin precipitates a decrease in insulin-like growth factor binding proteins (IGF-BPs) and a successive increase in bioavailable insulin-like growth factor I (IGF-I) (78). Both insulin and IGF-I, as well as TNFα, bind to membrane-bound receptors on cells that stimulate cellular proliferation and inhibit apoptosis, thereby providing a mechanism for tumor development (79). This biologic pathway is especially relevant to breast cancer (80, 81) as mammary cell carcinomas typically exhibit an overexpression of insulin receptors (82) and IGF-I receptors (83) making them very susceptible to the proliferative effects of these hormones. It is important to note that there are a number of genetic factors associated with obesity that may influence these relationships. Genetic variations associated with regulation
of food intake, glucose metabolism and lipid metabolism are potentially important factors in explaining the relationship between obesity and breast carcinogenesis (59).

Physical Activity and Carcinogenesis

The biologic pathways through which physical activity influences cancer risk are less understood than those for obesity, however several plausible mechanisms have been discussed in the recent literature (23, 84-87) and are illustrated in figure 2. Perhaps the most obvious mechanism through which physical activity may influence cancer risk is through its effect in reducing adipose tissue (58, 88), especially metabolically active visceral fat (23), thereby reducing the associated hormonal milieu that favors tumor initiation and promotion, discussed previously. In addition to this obesity-mediated pathway, physical activity has several independent effects on hormone and growth factor levels that are believed to influence carcinogenesis (89, 90). Independent of body size, higher levels of physical activity are associated with lower levels of estrogens and androgens and higher levels of SHBG among women (91-95). This impact of physical activity on sex hormone exposure may be due to a synergistic effect with adiposity as suggested by a recent clinical study showing that among women who lost weight, physical activity further increased SHBG and reduced total and free estradiol and estrone concentrations beyond those changes expected due to weight loss alone (95). Regular strenuous physical activity is also believed to reduce lifetime estrogen exposure through disruption of the normal menstrual cycle. Frequent intense physical activity can suppresses the secretion of lutenizing hormone by curbing secretion of gonadotropin-releasing hormone through several pathways (96). These exercise-induced alterations in the menstrual cycle are associated with delay of onset of menarche among adolescent females (97) as well as amenorrhea and menstrual irregularities for
premenopausal women in general (98), all of which are established protective factors for hormone-related cancers such as breast.

Physical activity is associated with an increase in insulin sensitivity by increasing expression of the insulin-stimulated GLUT-4 glucose transporter in the plasma membrane of skeletal muscle (99-103) and by reducing the level of free fatty acids, which have been linked to impaired insulin function (104). This increase in insulin sensitivity precipitates a decrease in insulin secretion, which is a possible mechanism for the observed increase in IGF-BPs (105) and decrease in IGF observed among physically active women (106). The ability of physical activity to mediate these metabolic hormones and growth factors (80, 81, 107) suggests another potential pathway for the observed protective effect of this exposure. This reduction in IGF may yield additional cancer protection as it may reduce sex hormone exposure by encouraging an increase in SHBG production by the liver (23).

While hormonal pathways offer the most convincing explanations for a protective effect of physical activity on carcinogenesis, other physiological effects of exercise may prove beneficial as well. Moderate physical activity may improve the immune response over both sedentary and exceptionally active individuals (108), possibly by promoting killer-cell, macrophage and cytokine activity (109, 110). Although it is possible for cancerous cells to be eliminated through an immune response, the specific effect of exercise on the anticarcinogenic actions of the immune system remains unclear (109). Moderate physical activity has also been shown to upregulate antioxidant enzyme activity (111), which may protect against DNA damage, although extreme levels of activity may increase the level of oxidative stress on the body (112). A moderate level of physical activity appears to be optimal for both immune and antioxidant systems, however the overall cancer protective
effect of exercise is likely due to an interaction between several different pathways that are affected by varying levels of activity (23), making specific dosing recommendations unclear.

**Epidemiology of Obesity and Breast Cancer**

Direct measures of body composition, such as percent body fat, are rarely measured directly in epidemiologic studies of cancer given the expense and logistical issues required for these assessments (113). Anthropometric measurements such as height, weight and body circumferences are easily gathered and reasonably accurate, even when self-reported (114) and can reflect elements of body composition relevant to disease etiology. BMI, the ratio of weight in kilograms to squared height in meters, is the currently accepted anthropometric measure of general adiposity (114). Correlations between BMI and densitometry-measured body fat have been noted to range from 0.55 to 0.85 (114), although studies have shown significant variations in body fat within levels of BMI, which appear to be related to body build, physical activity and race or ethnicity (115, 116). As a measure of general adiposity BMI also does not reflect the location of excess fat mass, which may be a more relevant factor in disease etiology and prognosis. Nevertheless, the ease of use, accuracy and general association with adiposity make it a useful tool in population-based epidemiologic research (113). Body circumference measurements and their index measures (such as waist-to-hip ratio (WHR), the ratio of the waist circumference to hip circumference) are similarly easy to obtain and do reflect fat patterning (114), as the WHR is a common measure of central adiposity, with higher values indicating a greater deposit of metabolically active visceral fat tissue (117). Unfortunately these measures are not always practical, and accuracy can depend upon the skill of trained personnel to consistently identify proper locations for measurement (114). An alternative anthropometric measure of adiposity is adult weight gain, which may be more relevant than BMI as weight gain after adolescence is primarily
associated with accumulation of adipose tissue so this measure may be more associated with the etiologically relevant exposure than BMI (118). Anthropometric measures of body composition such as these have shown consistent associations with breast cancer incidence and survival, likely reflecting the effect of energy balance on the carcinogenic process.

**Obesity and Breast Cancer Incidence**

Body size seems to have differential effects on breast cancer risk based on menopausal status, with obesity appearing to decrease risk among premenopausal women while increasing risk for postmenopausal women (21). These effects reflect the fact that excess adipose tissue has the potential to disrupt endogenous estrogen production among younger women yet represent a significant source of sex hormones after menopause, through mechanisms discussed previously. A pooled analysis of seven prospective cohort studies reported that premenopausal women with a BMI greater than 31 kg/m² had a relative risk (RR) of breast cancer of 0.54 (95% confidence interval (CI): 0.34-0.85) compared to women with a BMI below 21 kg/m² (22). This analysis also reported a positive association between BMI and risk of breast cancer among postmenopausal women: postmenopausal women with BMI above 28 kg/m² had a RR of 1.26 (95% CI: 1.09-1.46) compared to women less than 21 kg/m². Similar associations were found in this study when weight was used as the measure of body size. One recent meta-analysis of thirteen case-control and cohort studies across the world reported a RR of 1.12 comparing normal-weight (BMI less than 25 kg/m²) to overweight (BMI between 25 kg/m² and 30 kg/m²) postmenopausal women (119). A separate meta-analysis of eight prospective studies of postmenopausal breast cancer found a RR of 1.19 (95% CI: 1.05-1.34) for a 5 kg/m² increase in BMI (120); this association was significantly attenuated (RR: 1.02, 95% CI: 0.89 to 1.17) after adjustment for serum sex
hormones, further corroborating the hypothesis that adiposity’s effect is at least partially through the estrogen pathway.

The potential metabolic consequences of central obesity have lead researchers to examine breast cancer risk associated with anthropometric measures that reflect the pattern of visceral adiposity, such as waist circumference or WHR. Waist circumference and WHR have been inconsistently associated with an increased risk of breast cancer among premenopausal women in case-control studies (20), however most cohort studies do support a positive association (21). An analysis using data from the Nurses’ Health Study I showed an elevated risk associated with highest quintile of WHR compared to the lowest for all women (RR: 1.22, 95% CI: 1.12-2.99), while the association was stronger for women who never used hormone replacement therapy (RR: 1.85, 95% CI: 1.25-2.74). Data from the Iowa Women’s Health Study suggests that this association is strongest among women with a family history of breast or ovarian cancer (121, 122) although a recent update in this population shows this interaction is not as strong as previously reported (123).

Changes in body size over time, especially in adulthood, also appear to play a significant role in the association between body composition and breast cancer risk as results from several studies show (21). In general, an increase in body size has been associated with a slight decrease or null effect on premenopausal breast cancer risk (124-128), similar to the associations observed for BMI. Weight change appears to have a more profound effect on postmenopausal breast cancer as it has been consistently associated with increased risk, and weight loss with reduced risk (20). A case-control report by Eng and colleagues (129) examined the effect of changes in body mass over the life course on risk of incident breast cancer among postmenopausal women on Long Island, New York. The authors found that
postmenopausal women who gained more than 15 kg since age 20 years had a significant increase in risk of breast cancer (odds ratio (OR): 1.60, 95% CI: 1.11-2.26) compared to women who maintained their weight from age 20. This study also showed that women who reduced weight over their lifetime had a decrease in the estimated risk of breast cancer (OR: 0.55, 95% CI: 0.32-0.96) suggesting that weight loss is associated with decreased risk.

Huang and colleagues’ findings from the Nurses’ Health Study suggest that an increase in risk for postmenopausal breast cancer associated with weight gain may be stronger among women who never took hormone therapy compared to those who have (126), an interaction that has been corroborated by other reports (130, 131). These findings indicate that never users of HRT may be more susceptible to adipose-driven perturbations in estrogen levels than women with a history of HRT use. Recently a cohort analysis by Harvie and colleagues from the Iowa Women’s Health Study reported that compared to women who consistently gained weight over their adult lives, women who maintain their weight from age 18 to 30 and lost weight between 30 and menopause had a reduced risk of postmenopausal breast cancer (RR: 0.36, 95% CI: 0.22-0.60) as did those who maintained or lost weight from age 30 to menopause but lost weight after menopause (RR: 0.48; 95% CI: 0.22-0.65) (132). These results indicate that even recent weight change can have a significant impact on breast carcinogenesis, which may hint at the possible prognostic implications of weight change post-diagnosis.

**Obesity and Breast Cancer Survival**

**Obesity Near Diagnosis and Survival.** Obesity at diagnosis and weight gain pre-diagnosis have been established as indicators of poorer prognosis as discussed by several recent reviews (57, 133-135). Of the 26 studies reviewed by Rock and Demark-Wahnefried (134), 17 reported an association between greater body size and decreased survival or
increased rates of recurrence while 36 of the 51 discussed in an updated review by Goodwin found similar associations (57). Elevations in risk have generally been moderate with RRs and HRs comparing extremes of body size ranging from approximately 1.5 to 2.5 (57), however extremes as high as nearly sixfold have been noted (136). A 1995 meta analysis of 30 studies by Goodwin and colleagues reported a HR for death over 10 years comparing overweight to normal weight women of 1.56 (95% CI: 1.38-1.76) while the HR for recurrence over 5 years was 1.91 (95% CI: 1.52-2.40) (137). Although poorer prognosis with increasing body size is seen all women, Goodwin’s 2005 review reports that the findings in the literature support a somewhat greater increase in risk among premenopausal women and among women with hormone receptor positive cancers (57). Additionally, there is evidence that the relationship between BMI and prognosis is nonlinear, with underweight and overweight women having greater risk than women of normal body size (138).

More recently using data from a study of 1,254 young women (age 20 to 54) from Atlanta and New Jersey who were diagnosed with invasive breast cancer, Abrahamson and colleagues found an moderate increase in risk of all-cause mortality for women who were obese at diagnosis (BMI 30 kg/m$^2$ or above) compared to normal weight women (BMI 18.5-24.9 kg/m$^2$), reporting a hazard ratio (HR) of 1.48 (95% CI: 1.09-2.01) (139). The association with WHR was similar (HR comparing highest to lowest quartiles: 1.52, 95% CI: 1.05-2.19), suggesting the effect may be driven by central adiposity. Most recently, an analysis of 1,508 women diagnosed with breast cancer between 1996 and 1997 on Long Island, New York examined prognosis in relation to BMI in the year prior to diagnosis as well as changes in BMI during adulthood up to diagnosis (140). Among premenopausal women, comparing obese (BMI 30 kg/m$^2$ or above) to normal weight women (BMI 18.5
kg/m^2 -24.9 kg/m^2) the authors report a marked increase in risk of breast cancer specific
death (HR: 2.85, 95% CI: 1.30-6.24), with a similar magnitude of association for all-cause
mortality. The increased risk for postmenopausal women was more moderate, yet still
significant, with obese women having nearly twice the risk of breast cancer related death
compared to normal weight women (HR: 1.88, 95% CI: 1.04-3.34). This data also suggested
that significant adult weight gain (>16 kg since age 20) confers an increased risk of breast
cancer related death among premenopausal women compared to stable-weight women (HR:
2.09, 95% CI: 0.80-5.48) as does a greater than 12.7 kg gain from age 50 years to diagnosis
for postmenopausal women (HR: 2.95, 95% CI: 1.36-6.43).

**Changes in Body Size After Diagnosis of Breast Cancer.** Weight gain after
diagnosis in breast cancer patients is well documented, with the amount ranging from
approximately 1 kg to more than 10 kg within the first two years (57). Several reviews (57,
141) as well as recent reports (142-145) have identified younger age, lower pre-diagnosis
BMI, premenopausal status, later disease stage, use of systemic chemotherapy, and time
since diagnosis as major factors associated with increases in post-diagnosis bodyweight.
Behaviorally, alterations of dietary behavior due to psychological distress or physical
discomfort appear to be associated with weight gain (146, 147), however associations with
low physical activity levels have been inconsistent (144, 148-151). Demark-Wahnefried and
colleagues also report that women who gained weight while on chemotherapy showed a loss
of lean body mass (148), which has potential implications for reduction in basal metabolic
rate. Unfortunately, this finding is consistent with the fact that women who gain weight after
breast cancer diagnosis rarely return to their pre-diagnosis body size (152, 153).
Post-diagnosis Changes in Body Size and Breast Cancer Survival. The relationship of breast cancer prognosis with post-diagnosis changes in body size is far less studied than the prognostic value of pre-diagnostic body size. Only a handful of studies, most published over a decade ago have directly examined the relationship between weight gain post-diagnosis and survival with inconsistent results (152, 154-160). Most have reported no association of prognosis with relatively small increases in weight (155-157), however several have noted an association when examining larger changes in weight or BMI, often in subgroup analyses (152, 154, 158). Camoriano and colleagues report a negative prognostic effect of weight gain among premenopausal women (HR comparing gain of 5.9 kg to less than 5.9 kg: 1.62, 95% CI: 1.01-2.62), but no association among postmenopausal women (p=0.25, no HR reported).

Of the most recent reports, summarized in table 1, Kroenke and colleagues recently reported data from the Nurses Health Study cohort that showed that among never smokers, weight gain post-diagnosis was related to increased risk of death and recurrence of cancer (158). Among never smokers, those gaining between 0.5 and 2.0 kg/m² over the follow-up period had 35% greater risk compared to those who did not gain (HR: 1.35, 95% CI: 0.93-1.95) while those gaining more than 2.0 kg/m² had even greater risk (HR: 1.64, 95% CI: 1.07-2.51). The most recent study by Nichols et al. found sizeable increases in risk among women with weight gains of 10 kg or more compared to those who maintained 2 kg of their pre-diagnosis weight: (all-cause mortality HR: 1.70, 95% CI: 1.21-2.41; breast cancer specific mortality HR: 1.78, 95% CI: 1.01-3.14; cardiovascular disease mortality HR: 1.73, 95% CI: 0.83-3.62). A null report was published by Caan et al. who found modest, and statistically insignificant associations between women who gained >10% of their pre-
diagnosis weight compared to those who maintained within 5% for both all-cause mortality (HR: 0.7, 95% CI: 0.4-1.2) and breast cancer recurrence (HR: 0.8, 95% CI: 0.5-1.2).

Studies with null findings were limited by small sample size (155-157) and short follow-up (155, 157), or likely biased samples and poor timing of exposure assessment (159) possibly explaining the lack of associations. Given the limited evidence to date, additional investigation of the effect of weight change on prognosis is needed before weight management guidelines for breast cancer survivors can be defined.

**Epidemiology of Physical Activity and Breast Cancer**

Assessment of physical activity in epidemiologic studies of cancer is challenging as direct measures such as accelerometers, pedometers and calorimeters are not applicable to case-control designs, fail to measure specific activities or fail to capture activity at an etiologically relevant time period, and are often too expensive for cohort studies (161). Use of questionnaire-based physical activity assessments is standard practice in observational epidemiology given the relatively low cost and ease of administration. Physical activity questionnaires seek to assess some combination of the type, frequency, intensity and duration of activity levels throughout various domains of an individual’s life (occupational, leisure, domestic duties, transportation) (162). Although questionnaire data may yield only crude categorizations of physical activity levels, data from detailed assessments can be summarized into an intensity score representing the multiple of energy expended while at rest, the metabolic equivalent score (MET) (162, 163). The MET is only an estimate of relative energy expenditure and the true value for an individual would vary by a number of factors including intensity of activity, body composition and other physiological variables. Although qualitative differences in assessment tools and assessment at an etiologically relevant time frame of exposure make it difficult to compare results across studies the
literature overwhelmingly supports a protective effect for physical activity on incident breast cancer (163).

**Physical Activity and Breast Cancer Incidence**

The relationship between physical activity and incident breast cancer has been studied extensively and the overall findings have been discussed in several reviews (163-166). The most recent was a systematic review by Monninkhof and colleagues (167) that evaluated the evidence from 19 cohort and 29 case-control studies. Among the case-control studies examined, 4 of the 6 that assessed total physical activity (168-173) reported protective effects (which the authors define as relative effect measures below 0.80) while 14 of the 28 studies reporting on leisure time activity (165, 168-170, 172-195) found reduced risk associated with greater physical activity. Results for cohort analyses were more equivocal than for the case-control studies: the 3 studies reporting associations for total activity were in complete disagreement (one protective (196), one null (197), one harmful (198)) and only 8 of the 17 reports of leisure time activity (187, 197, 199-213) found a protective effect of greater activity levels. A larger proportion of studies report a protective effect of increased leisure time activity among postmenopausal women than premenopausal women, and several reports have indicated that initiation of exercise after menopause is associated with a reduction in risk (200, 205, 208). The authors observed that most studies examining effect modification by body size failed to find differences suggesting that physical activity has independent effects on carcinogenesis, a conclusion also reached in the review by Patel and Bernstein (163). After consideration of the variation in physical activity questionnaires, as well as formal statistical tests of heterogeneity of effect across study design (case/control) and menopausal status (premenopausal/postmenopausal), Monninkhof and colleagues felt it prudent to avoid calculation of a summary effect measure, although they noted that among
studies deemed “protective,” reductions in risk ranged from 21% to 52% among the case-control studies and 21% to 80% among the cohort studies. Ultimately these authors conclude that the current weight of evidence of a protective effect strong for postmenopausal women and inconclusive for premenopausal women.

**Physical Activity and Breast Cancer Survival**

The psychosocial and physiological effects of physical activity among female breast cancer survivors have been mentioned in recent reviews of exercise interventions in cancer patients (214, 215). During breast cancer treatment, physical activity has been associated with improvements in treatment-related side effects such as fatigue (216-219), nausea (220) and quality of life (219), as well as body size (149, 221). Studies of post-treatment physical activity have also noted improvements in immune function (222, 223); these improvements in body size and immune function due to exercise may be especially relevant to prognosis. Although these findings are encouraging, important limitations with these trials include low statistical power due to small sample size as well as lack of common and universally meaningful outcome measures (215, 224, 225). Perhaps the greatest issue with current intervention studies of survival is that outcomes focusing on short term measures such as physiological and psychological parameters (224) instead of survival or recurrence.

There are only a handful of studies examining the effect of physical activity, either pre- or post-diagnosis, on breast cancer recurrence and survival, as outlined in table 2. Two separate studies (226, 227) of 412 and 603 female breast cancer patients failed to find associations with pre-diagnosis physical activity and breast cancer prognosis. The smaller study by Rohan and colleagues reported a HR of 0.98 (95% CI: 0.50-1.94) for breast cancer related death comparing those expending more than 4000 kcal/week in recreational physical activity to those expending 0 kcal/week (227). Borugian and colleagues examine
associations with breast cancer mortality across levels of seven different types of recreational physical activity (climbing stairs, walking, sports, exercise, jogging, swimming, gardening), with five levels each (none, few times a year, few times a month, once a week, more than once a week) with most associations at or near unity (228). Abrahamson and colleagues report on all-cause mortality in a follow-up study of 1,264 women diagnosed with invasive breast cancer (229). The authors find a reduction in risk of death across extreme quartiles of physical activity expressed as MET hours per week (HR: 0.78, 95% CI: 0.56-1.08), an effect that appears to be limited to women who were overweight or obese near diagnosis (HR: 0.70, 95% CI: 0.49-0.99).

Recently, several studies have been published examining the effect of post-diagnosis levels of physical activity and survival. Holmes and colleagues report a beneficial effect of leisure-time physical activity on breast-cancer prognosis using data from the Nurses’ Health Study (230). Comparing those women with 24 or more MET-hours per week of physical activity two years after diagnosis to those with fewer than 3 MET-hours per week, reduction in risk of overall death (HR: 0.65, 95% CI: 0.48-0.88), breast cancer related death (HR: 0.60, 95% CI: 0.40-0.89) and recurrence (HR: 0.74, 95% CI: 0.53-1.04) were noted with significant trends for overall and breast cancer survival. There was no finding of effect modification for dichotomized physical activity across levels of menopausal status or stage, however results suggested the protective effect among women with hormone-receptor positive tumors was stronger (HR: 0.50, 95% CI: 0.34-0.74) compared to women with hormone-receptor negative tumors (HR: 0.91, 95% CI: 0.43-1.96, p for interaction: 0.08). Pierce et al. (231) report that women with high post-diagnosis physical activity and high fruit and vegetable consumption have longer survival compared to those with in the lowest
categories of both exposures (HR: 0.53, 95% CI: 0.31-0.98). The analysis by Holick et al. (232), in the largest study to date, reported effects for both all-cause mortality (HR: 0.44, 95% CI: 0.32-0.68) and breast cancer-specific mortality (HR: 0.51, 95% CI: 0.29-0.89) comparing women expending >21.0 MET-hours/week to those expending < 2.7 MET-hours/week. The strongest effect was noted by Irwin et al. (233) who found that women expending >9.0 MET-hours/week had 33% the risk of death from any cause (HR: 0.33, 95% CI: 0.15-0.73) compared to inactive women (0 MET-hours/week), although the effect for breast cancer-related deaths was not as strong (HR: 0.65, 95% CI: 0.23-1.87). Although several of these studies report inverse associations between post-diagnosis physical activity and survival, the most recent study by Sternfeld et al. (234) reported largely null associations between post-diagnosis physical activity and both recurrence and survival after breast cancer diagnosis. Of note is the fact that to date, all of the studies of post-diagnosis physical activity and survival have been limited to women who were enrolled at least 2 years post-diagnosis. Failure to include women who died within the first several years of diagnosis (or who may be otherwise unable to enroll) is certain to result in a biased sample. Additionally, exposure assessment occurred only once, at enrollment, and therefore only measured activity levels well after treatment was concluded, and did not capture the longitudinal patterns of physical activity that may be important in the survival experience.

The limited evidence and variance in methodology between these studies make it difficult to draw a general conclusion regarding the effect of physical activity on breast cancer prognosis. Notably, the two studies that found the strongest effects were large population-based epidemiologic analyses while the null results came from much smaller studies. Assessment of physical activity at a biologically relevant point is clearly important
for answering the question most relevant for survivors who must make decisions regarding lifestyle modifications after diagnosis and treatment. Although the level of pre-diagnosis physical activity is thought to be correlated with post-diagnosis level after treatment (235, 236) the use of pre-diagnosis assessments in two null reports (226, 227) could be a factor in their findings. The results from the recent reports of post-diagnosis activity are encouraging, but these studies did not evaluate the longitudinal pattern. Much work remains in the evaluation of the effect of physical activity on breast cancer survival before a consensus can be reached.
Summary

Given the weight of the evidence, the association between energy balance and breast cancer development is convincing. Biologically, excess adipose tissue and physical inactivity are associated with a hormonal environment that encourages cellular proliferation and possibly DNA damage required for carciogenesis. Epidemiologic studies have established obesity and weight change as risk factors for postmenopausal breast cancer, while they appear to be somewhat protective for premenopausal women. Increased levels of physical activity also appear to be associated with a decrease in risk of breast cancer for postmenopausal women, yet the evidence is less clear for breast cancer risk before menopause.

These facets of energy balance have also shown promise as potential prognostic indicators after diagnosis of breast cancer. Factors associated with positive energy balance, especially elevations in the hormones estrogen, insulin and IGF-1, are believed to negatively influence survival after breast cancer diagnosis by encouraging promotion of cancerous cells through similar mechanisms believed to increase breast cancer risk. Studies examining pre-diagnostic weight and breast cancer outcomes show that heavier women have less favorable prognosis, even among women with recent weight loss. Of special interest to breast cancer survivors is the effect of post-diagnostic lifestyle changes, such as weight modification and increasing physical activity, on survival. Recent work on the effect of post-diagnostic changes in body weight provides evidence that weight loss reduces mortality, perhaps more so among subgroups. Reports of the effect of physical activity on survival have recently increased, but the evidence has been somewhat conflicting. Overall, the evidence for a protective effect of post-diagnostic alterations in factors associated with energy balance is alluring, but much work remains. Further examination of these relationships, especially in
large epidemiologic studies, may yield information with significant clinical and public health impact.
Table 1.1. Recent studies of the effects of post-diagnosis changes in body size and breast cancer prognosis.

<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Sample Size</th>
<th>Median Length of Follow-up</th>
<th>Outcome Measure</th>
<th>Measure of Change in Body Size</th>
<th>Summary of Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camoriano et al., 1990 (152)</td>
<td>545</td>
<td>6.6 years</td>
<td>R, OS</td>
<td>weight change above/below median</td>
<td>premenopausal: HR recurrence: 1.5 (p=0.17), HR all-cause mortality: 1.62 (p=0.04)</td>
</tr>
<tr>
<td>Costa et al., 2002 (155)</td>
<td>106</td>
<td>4.9 months (mean)</td>
<td>R, OS</td>
<td>any percent body weight increase</td>
<td>log-rank test p=0.08</td>
</tr>
<tr>
<td>Kroenke et al., 2005 (158)</td>
<td>5,204</td>
<td>9 years</td>
<td>R, OS, BCM</td>
<td>change in BMI: 0-0.5 kg/m², 0.5-2.0 kg/m², &gt;2.0 kg/m²</td>
<td>Among never smoking women, comparing &gt;2.0 kg/m² to 0-0.5 kg/m² increase: OS HR: 1.59 (95% CI: 1.12-2.27)</td>
</tr>
<tr>
<td>Caan et al., 2008 (159)</td>
<td>1,689 (in final analysis)</td>
<td>83.9 months</td>
<td>R, OS, BCM (pre-diagnosis BMI only for BCM)</td>
<td>Weight change: &gt;10% loss/gain, 5-10% loss/gain, maintain ±5% (ref)</td>
<td>For &gt;10% weight gain: OS HR: 0.7 (0.4-1.2)</td>
</tr>
<tr>
<td>Nichols et al., 2009 (160)</td>
<td>3,993</td>
<td>6.3 years</td>
<td>OS, BCM, CVDM</td>
<td>Weight change: &gt;10kg loss, 10-2kg loss, maintain ±2kg (ref), 2-6 kg gain, 6-10 kg gain, &gt;10 kg gain</td>
<td>For &gt;10kg increase: OS HR: 1.70 (1.21-2.41)</td>
</tr>
</tbody>
</table>

Table 1.2. Recent studies of the effects of post-diagnosis physical activity and breast cancer prognosis.

<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Sample Size</th>
<th>Median Length of Follow-up</th>
<th>Outcome Measure</th>
<th>Physical Activity Measure</th>
<th>Summary of Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holmes et al., 2005 (230)</td>
<td>1,264</td>
<td>96 months (BCM)</td>
<td>OS, BCM, R</td>
<td>MET-hours per week for leisure-time physical activity, 2 years post-diagnosis, categorized as &lt;3, 3-8.9, 9-14.9, 15-23.9, &gt;=24</td>
<td>Comparing &gt;= 24 MET-h/wk to &lt;3 MET-h/wk: OS HR: 0.65 (95% CI: 0.48-0.88) BCM HR: 0.60 (95% CI: 0.40-0.89) R HR: 0.74 (95% CI:0.53-1.04)</td>
</tr>
<tr>
<td>Pierce et al., 2007 (231)</td>
<td>1,490</td>
<td>Average 6.7 years</td>
<td>OS</td>
<td>Various categorizations of MET-minutes/week measured once at follow-up (avg. 2 years, max 4 years post-diagnosis)</td>
<td>Compared to low fruit and veg, and low physical activity, women with high fruit and veg intake and high physical activity, HR: 0.53 (95% CI: 0.31-0.98).</td>
</tr>
<tr>
<td>Holick et al., 2008 (232)</td>
<td>4,482</td>
<td>5.5 years</td>
<td>OS, BCM</td>
<td>MET-hours week at follow-up (average ~6 years post-diagnosis): &lt;2.7 (ref), 2.8-7.9, 2.0-20.9, &gt;=21.0</td>
<td>Compared to &lt;2.7 MET-hours/week, those with &gt; 21.0 MET-hours/week: OS HR: 0.44 (0.32-0.68) BCM HR: 0.51 (0.29-0.89)</td>
</tr>
<tr>
<td>Irwin et al., 2008 (233)</td>
<td>933</td>
<td>6 years</td>
<td>OS, BCM</td>
<td>MET-hours 2 years after diagnosis: 0 (ref), 0.1-9.0, &gt;9.0</td>
<td>Compared to 0 MET-hours/week, women with &gt;9 MET-hours/week: OS HR: 0.33 (0.15-0.73) BCM HR: 0.65 (0.23-1.87)</td>
</tr>
<tr>
<td>Sternfeld et al., 2008 (234)</td>
<td>1,970</td>
<td>87 months</td>
<td>OS, BCM, R</td>
<td>Quartiles MET-hours/week: overall, mod/vig activity, moderate only, vigorous only</td>
<td>Total MET-hours/week, quartile 4 vs. quartile 1: OS HR: 0.76 (0.48-1.19) BCM HR: 0.87 (0.48-1.59) R HR: 0.91 (0.61-1.36)</td>
</tr>
</tbody>
</table>

Abbreviations: BCM: breast cancer mortality, OS: overall survival, R: recurrence of breast cancer, MET: metabolic energy expenditure, HR: hazard ratio, CI: confidence interval.
Figure 1.1. Hypothesized biological mechanisms for the relationship between adiposity and breast cancer survival among breast cancer patients.
Figure 1.2. Hypothesized biological mechanisms for the relationship between physical activity and breast cancer survival among breast cancer patients.
References


CHAPTER 2
RESEARCH METHODS

Overview

To evaluate the effects of post-diagnosis changes in body size and physical activity levels on survival after breast cancer diagnosis I used data from the follow-up study of the Long Island Breast Cancer Study Project (LIBCSP). The parent LIBCSP study was a population-based case-control study of newly diagnosed cases of breast cancer in Nassau and Suffolk counties on Long Island, New York. This follow-up study includes data from 1,508 women diagnosed with *in situ* or invasive breast cancer between August 1, 1996 and July 31, 1997 in Nassau and Suffolk counties in Long Island, New York. Baseline data from the original case control study as well as the follow-up interviews conducted approximately five years post-baseline included anthropometric and lifestyle characteristics as well as sociodemographic and medical factors relevant to breast cancer prognosis. Breast cancer specific mortality and all-cause mortality through December 31, 2005 were determined from data from the National Death Index. Self-reported physical activity and anthropometric data from the follow-up interview were related to the outcomes in a time-to-event analysis. Of the 1,508 breast cancer cases who participated in the baseline case-control interview, follow-up interview data was available for only 1,033 women, yielding a large percentage of missing data. To address this issue I applied a selection model from the biostatistics literature that allows for examination of the effects of varying the assumptions underlying the missing data on the observed relationship between changes in body size and physical activity levels and mortality.
Parent Study

This study uses data from the follow-up to the Long Island Breast Cancer Study Project (LIBCSP), a population-based case-control study of breast cancer in Nassau and Suffolk counties in New York. The LIBCSP has been described in detail in the recent literature (1) but relevant details will be reviewed here. Potential cases for this study were English-speaking women with in situ or invasive breast cancer who were diagnosed between August 1, 1996 and July 31, 1997. These potential cases were identified through pathology departments of participating hospitals and their physicians were contacted to confirm the diagnosis and obtain permission to contact the patients for participation in the study. Initially 2,271 women were identified as potentially eligible cases with physician consent ultimately obtained for 1,837 (80.9%) of these women. Potential controls were English-speaking women selected from the population of current residents of Nassau and Suffolk counties with no history of diagnosis of breast cancer.

Written informed consent was obtained from all study subjects prior to participation in the case-control interview. Questionnaires were administered at baseline by trained study personnel and included assessments of reproductive, occupational, residential and medical histories, as well as sociodemographic, anthropometric, lifestyle, and environmental exposures (2). A total of 1,508 (82.1%) of the eligible case women completed the main study questionnaire taking an average of 101 minutes to complete the interview. After the main interview study subjects were asked to self-administer a modified Block-style food frequency questionnaire, with 1,481 (98.2%) cases complying. Additionally, case subjects were asked to sign a medical record release to allow study personnel to gather data regarding clinical characteristics relevant to the breast cancer diagnosis, such as stage, tumor size and
hormone receptor status. A total of 1,493 (97.7%) of the case women signed the release forms and subsequently 1,402 records were abstracted.

Follow-up Study

Study Population
During the parent case-control study 94 women refused to be contacted at a later date for participation in the follow-up study. For those who agreed to participate in the follow-up, the case subjects or their proxy were contacted by mail approximately 5 years after initial diagnosis of breast cancer and informed consent was obtained via telephone follow-up calls. Of the 1,414 women who initially agreed to participate in the follow-up study, 60 subsequently refused at the initial mail contact, 65 refused when contacted by telephone, 18 refused due to illness, 22 were unable to complete the interview, 55 were lost to follow-up and 96 were deceased with no identifiable proxy to complete the interview. Of the remaining 1,098 subjects who agreed to the follow-up interview, 65 were only able to provide information on their first course of treatment for their original breast cancer diagnosis. Ultimately 1,033 case subjects (68.5% of the original 1,508 women) actually completed the follow-up interview (3). The follow-up interview ascertained information similar to that gathered in the baseline questionnaire but relevant to the time period since diagnosis including treatment, reproductive history, smoking and alcohol use, and as well as body size and physical activity.

Description of Cohort of Women with Breast Cancer
Descriptive statistics at diagnosis for the 1,508 case women have been previously published (1) and are summarized briefly in table 1. The median age at time of diagnosis for the case women was 58 years with a range from 25-98 years. The majority of case women were postmenopausal at time of diagnosis, white and of non-Hispanic ethnicity. Nearly all women
reported having been married and most had at least one child. Most had no history of a first-degree family relative with breast cancer. Less than one-third reported ever having used hormone replacement therapy. For those reporting height and weight (n=1,491) mean BMI at baseline was 26.5. Among all breast cancer cases in this study the average MET-hours per week of recreational physical activity from age at menarche to reference date was 27.7 (4). Most women were diagnosed with invasive breast cancer with small tumors (less than or equal to 2 cm), with no nodal involvement and with either ER- or PR-positive tumors. Over 60% of the cases reported having radiation therapy or hormone therapy during their treatment regimen, while less than half reported that they received chemotherapy after diagnosis.

**Outcome Assessment**

The outcome for this study, date and cause of death, was determined by the National Death Index (NDI) (5), a centralized database of death records maintained by the National Center for Health Statistics which is considered the standard source of mortality data for epidemiologic research (6). A text-formatted data file including first and last name, city, state, date of birth, social security number, gender, and marital status along with the unique LIBCSP subject identifier was sent to the NDI offices to be matched with their records to obtain the corresponding date and cause of death for each subject. Subjects without a match returned were assumed to be alive on the final date available in the NDI database. For women who were determined to be deceased from the NDI records I constructed two event indicators: (1) breast cancer related death (breast cancer listed as primary or secondary cause of death) and (2) death from any cause which, combined with time of death, represent be the two outcomes for this analyses. An initial match of the LIBCSP case records to the NDI was conducted in 2004 (7) and an update was completed in early 2008, yielding a total of 308 deaths as of December 31, 2005 with 164 of these attributed specifically to breast cancer.
Exposure Assessment

**Body Size.** Body mass index (BMI) is calculated as the ratio of weight in kilograms to squared height in meters. At baseline the initial questionnaire (2) collected self-reported data on height in inches and body weight in pounds at age 20 years and at 1 year prior to the reference date (date of diagnosis for cases). The baseline questionnaire also collected weight in pounds by decade of life from 20 years through 70 years of age. The follow-up questionnaire contained similar questions asking self-reported height, body weight in pounds one year after diagnosis, current weight, maximum weight since time of diagnosis and minimum weight since time of diagnosis. Measures of BMI and weight change at each corresponding time point were calculated from these variables; details will be discussed under variable definition.

**Physical Activity.** Recreational physical activity was measured using a modification of the instrument developed by Bernstein and colleagues used in previous studies of physical activity and breast cancer risk (8). The baseline questionnaire included a screening question asking “Have you ever participated in any physical activities or exercises on a regular basis—that is for at least 1 hour per week for 3 months or more in any year?” (2) with those subjects answering affirmatively considered having ever participated in physical activity. These subjects were then asked to list each activity separately and then answer more detailed questions for each one:

1. “In what activity did you (first/next) participate on a regular basis?”
2. “Looking at the calendar, at what age did you start (ACTIVITY) regularly?”
3. “At what age did you stop (ACTIVITY)?”
4. “For how many years did you (ACTIVITY) regularly? For how many months each year did you do this?”
(5) “On average, about how many hours per week did you actually (ACTIVITY)?”

Eng (4) describes in detail the procedure used for calculating physical activity variables from this data, which will be summarized here. For the subjects with detailed physical activity data the number of months per year of each activity was converted to number of hours per week. Where an activity was reported without a corresponding duration, 12 months per year was imputed for non-seasonal activities while the average number of months per year was imputed for seasonal activities. For each activity a metabolic equivalent of energy expenditure (MET) score was assigned using a database derived from previously published values (9). The MET is a standard unit of energy expenditure for physical activity research (10) and is defined as the energy expended while sitting quietly, approximately one kilocalorie per kilogram of bodyweight per hour. When an activity reported by the study subjects did not have a corresponding MET-score published then a similar activity with a published MET value was chosen. For each activity the MET value was then multiplied by duration of activity in number of hours per week, which was added across all activities for each subject for each year of life since menarche to arrive at a measure of the total number of MET-hours per week for each subject.

In the follow-up questionnaire subjects were asked if they had participated in any physical activities or exercises on a regular basis for at least 1 hour a week for 3 months or more in any year since the date of diagnosis. If an affirmative answer was recorded then detailed questions regarding type, duration, frequency and age at commencement were asked, similar to the baseline questionnaire, for up to 12 separate activities with the time period queried relevant to the time since the reference date (diagnosis) for the case-control study.
The average number of MET-hours per week in the 5 years since diagnosis was calculated in the same manner described above for the baseline measure.

**Behavioral, Clinical and Histopathologic Covariates.** The baseline and follow-up questionnaires were administered by trained study staff and ascertained a number of lifestyle measures (1, 2). The baseline questionnaires collected detailed data on reproductive and menstrual history, smoking, alcohol intake, hormone therapy and oral contraceptive use. Diet was measured using a self-completed modified semi-quantitative Block food frequency questionnaire (FFQ) that contained approximately 100 line items corresponding to different foods or food groups and included measures of portion size and frequency of consumption. This data was used to calculate a broad set of dietary measures including frequency of consumption of standard portions of various foods as well as a comprehensive set of micronutrient intakes, macronutrient levels and total energy intake. Total energy are potential confounders for these analyses.

At baseline, clinical and histopathological variables were established from medical chart abstraction and included disease stage, treatment modality, ER/PR status and other relevant tumor characteristics. The follow-up interview also included questions regarding treatment modality in an attempt to ascertain this variable for those women who were pre-treatment or had yet to complete a course of treatment at the time of enrollment on the study (1). During the follow-up study medical records were abstracted for 598 women and the details regarding treatment regimen were compared to that self-reported in the follow-up interview. Kappa coefficients comparing self-report and medical records were high for all three treatment modalities examined: radiation therapy $\kappa=0.97$, chemotherapy $\kappa=0.96$ and
hormone therapy κ=0.92 (7). Data on tumor size was ascertained through the New York State Cancer Registry.

**Results from Previous Analyses**

Several published manuscripts, based on data from the parent case-control study, include results that are relevant to the proposed project. A recent report by Eng et. al. (11) examined the effect of changes in body mass over the life course on risk of incident breast cancer in the LIBCSP. The authors report that among postmenopausal women, those who gained more than 15 kg since age 20 years had a significant increase in risk of breast cancer (OR: 1.6, 95% CI: 1.11-2.26) compared to women who maintained their weight at age 20. This study also showed that women who reduced weight over their lifetime had a decrease in the estimated risk of breast cancer (OR: 0.55, 95% CI: 0.32-0.96) suggesting that modification of body size is associated with decreased risk.

In a separate report of the effect of lifetime physical activity in this cohort Eng (4) showed that postmenopausal women who ever regularly exercised had a notable decrease in risk of breast cancer compared to women who never regularly exercised (OR: 0.81, 95% CI: 0.65-1.01). This study also showed that postmenopausal women who were active for at least 6.4 hours per week had a similar decrease in risk of breast cancer (OR: 0.73, 95% CI: 0.54-0.99). Both studies by Eng and colleagues suggest that changes in body mass as well as physical activity levels are associated with breast cancer risk, which make it plausible that changes in these factors would also be related to survival post diagnosis.

Using the data from the LIBCSP follow-up study, Cleveland et al. report that women who are obese one year prior to diagnosis (BMI >= 30.0 kg/m²) have an increased risk of breast cancer related death compared to ideal weight women for both premenopausal (HR:
2.85, 95% CI: 1.30-6.23) and postmenopausal cases (HR: 1.88, 95% CI: 1.04-3.34) with similar associations observed for all-cause mortality. Also, among women who were premenopausal when diagnosed with breast cancer, those who gained more than 16 kg between age 20 and 1 year prior to diagnosis had a markedly greater risk of death from all causes (HR: 2.45, 95% CI: 0.96-6.27) and breast cancer (HR: 2.09, 95% CI: 0.80-5.48) compared to women whose weight remained stable. A similar association was observed among postmenopausal women diagnosed with breast cancer who gained more than 12.7 kg after age 50 up to the year before diagnosis compared to those whose weight remained stable (all-cause mortality HR: 2.69, 95% CI: 1.63-4.43; breast cancer mortality HR: 2.95, 95% CI: 1.36-6.43). Body size was also found to modify the relationship between smoking and survival in this cohort as Sagiv and colleagues found (12). The authors report that the association between current active smoking (compared to never smokers) and increased risk of death from all causes was stronger among women who were obese at diagnosis (HR: 2.10, 95% CI: 1.03-4.27) than overweight (HR: 1.10, 95% CI: 0.52-2.35) or normal weight women (HR: 1.00, 95% CI: 0.54-1.86) with similar associations observed for breast cancer specific mortality. Examining dietary exposures that may be relevant to the relationship between energy balance and mortality, Fink et. al. (3) report that intake one year prior to diagnosis of 46 or more half-cup servings per week of fruits and vegetables is associated with an increase in survival among women diagnosed with invasive breast cancer compared to those who ate fewer than 18 half-cup servings a week (HR: 0.68, 95% CI: 0.42-1.09). In a separate report, Fink and colleagues also found an approximately 40-50% reduction in risk of all-cause and breast cancer specific death comparing extreme quintiles of intakes of flavones, isoflavones and anthocyanidins (13).
Data Analysis

Variable Definitions for Outcome and Exposures

**All-cause and Breast Cancer Specific Mortality.** A variable corresponding to length of time from baseline interview to either date of death or date of last follow-up (December 31, 2005) was created. Dichotomous indicator variables were created to indicate if the time variable indicates death from any cause and death from breast-cancer related cause or if the subject was alive (or died from a non-breast cancer related illness) at the associated time point.

**Changes in Body Size.** Changes in body size from baseline (one year before diagnosis) to the various follow-up assessments were categorized into groups corresponding to weight loss (>5% loss from pre-diagnosis weight), maintenance (within 5% of pre-diagnosis weight) and gain (2 groups, 5-10% gain from pre-diagnosis weight and >10% gain from prediagnosis weight). These cutpoints were determined in accordance with other recent reports (14) as well as to coincide with public health recommendations (15). For examination of effect modification, weight gain was combined into a single group (>5% gain from pre-diagnosis weight) in order to avoid sparse categories.

**Physical Activity Levels.** Physical activity levels, expressed in MET-hours/week were categorized as 0 MET-hours/week (inactive), 0.1-9.0 MET-hours/week and >9 MET-hours/week. This categorization was again chosen to correspond to previous reports (16, 17), facilitating comparison of our results with them as well as providing meaningful categories, as the highest category corresponds roughly to 2-3 hours a week of brisk walking. With the issue of a significant amount of missing data, it is important to remember that *a priori* cutpoints for both body size and physical activity are necessary as the distribution of the
missing data is unobserved, and therefore empirically-derived cutpoints are therefore inappropriate.

**Missing Data: Background**

Although outcomes are available on all 1,508 case subjects, only 1,033 women (68.5%) participated in the follow-up interview, yielding a large proportion of missing data for post-diagnosis body size and physical activity. Data on missing exposures and covariates is common in population based research and basic methods used to address this issue have been recently reviewed in the epidemiology literature (18). Missing data may arise from several different processes defined by the relationships between the probability that the data is missing, other covariates and the unobserved value of the missing variable.

In their classic text, Little and Rubin (19) propose a taxonomy that describes the mechanisms that may lead to missing data. Data is considered missing completely at random (MCAR) if the probability that data is missing is independent of both observed and unobserved data. Under this mechanism the observed data essentially constitutes a random sample of values from all subjects. If data is MCAR then a complete-case analysis, where the analysis is performed on all subjects with completely observed data, will yield valid estimates. If the probability that data is missing depends only upon observed variables then the mechanism is termed missing at random (MAR). Under this scenario a complete-case analysis will yield biased estimates, however analytic methods that account for the probability of missingness conditional on observed variables, such as multiple imputation and weighted estimating equations, are appropriate. The most problematic situation arises when the probability that data is missing depends upon unobserved values of the missing variable, for example if the probability that income is reported is dependent upon the unobserved value of income. When the probability that a variable is missing depends upon
its unobserved value, and possibly observed variables, the mechanism is referred to as not missing at random (NMAR). In MCAR and MAR the missing data mechanism is considered ignorable (its consideration is not required in the modeling process), while under NMAR the missing data mechanism is non-ignorable (one must account for the distribution that determined if a data element was missing). Valid estimation under non-ignorable missingness requires an approach that simultaneously incorporates a model for the probability that data is missing, a model for the distribution of the values of the missing variable and a model relating the variable to the outcome of interest.

The concern in this project is that follow-up data on weight change and physical activity may be non-ignorably missing. As described previously, many women were deceased at the time of the follow-up interview while a substantial number were ill or otherwise unable to complete the interview. If post-diagnosis weight changes or levels of physical activity of these women contributed to their death or inability to complete the interview then this would create a mechanism of non-ignorable missingness for these variables and a complete case analysis, imputation or a weighted estimation approach would yield biased estimates of the association between these variables and survival.

**Proportional Hazards Models Under Complete Data**

This analysis utilizes proportional hazards regression to relate the exposures of interest to survival. Proportional hazards regression is a statistical approach that allows for regression modeling of the rate of an event occurrence as a function of one or more variables (20-22). Here I sketch out the version of the proportional hazards model used here under the assumption that all data is completely observed. This brief description will be extended into the selection model in the next section, with the more formal derivation discussed in chapter 3. Assume we have data on a sample of \( n \) independent subjects and for subject \( i \) we observe
the variable $y_i$ representing the observed follow-up time for subject $i$ and indicator $\delta_i$ which equals 1 if $y_i$ corresponds to an event (e.g. death), and 0 if it represents a censored observation. To define the piecewise exponential model, we divide the time axis into $J$ discrete intervals $(s_{j-1}, s_j]$ for $j=1, \ldots, J$ with $s_0 = 0$ and $s_J$ greater than the maximum of the \{y_i\}. Each subject provides a series of longitudinal measurements for $p$ completely observed covariates indexed by observation $j$ by $x_{ij}$ for $j = 1, \ldots, J$, some elements of which may be constant (such as a fixed covariate measured only once, at baseline). The piecewise exponential hazards model is then defined with the hazard function:

$$\lambda(y_i | x_{ij}, \beta_j, \lambda_j) = \lambda_j \exp \{ x_{ij} \beta_j \} \quad \text{for} \quad y \in (s_{j-1}, s_j]$$  \hspace{1cm} (1)

where $\beta_j$ is the $p \times 1$ vector of coefficients on the vector of covariates $x_{ij}$ and $\lambda_j$ represents the baseline hazard within interval $j$. The piecewise exponential proportional hazards model is similar to the Cox proportional hazards model except the baseline hazards are explicitly defined to be constant within each interval $(s_{j-1}, s_j]$. The advantages of the piecewise exponential model are that it allows for the shape of the baseline hazard to change across the time interval and facilitates estimation of the hazard ratios from a familiar distribution without regard to tied failure times that are sometimes problematic for the standard Cox proportional hazards model (21, 22). Note that here, for convenience, it is assumed that the covariate measurement times correspond to the intervals of the piecewise exponential model, but this is not a requirement; this assumption will be relaxed in chapter 3.

**Selection Models for Survival Data with Missing Covariates**

Herring et. al. (23) propose a general technique that extends the fixed-time Cox proportional hazards model when some covariates may be missing, and allows for the missing data mechanism to be potentially non-ignorable. It is this method that I extended to allow
covariates that vary over time. This modeling paradigm, called a selection model, requires
the specification of the joint distribution of three processes: (1) the probability that data is
missing, (2) the distribution of the variable with missing values and (3) the relationship
between said variable and the outcome, time to death.

In addition to the covariates $x_{ij}$ defined previously, consider $z_{ij}$ a vector of length $q$ of
additional covariates, some of which may be missing at some point in time. Corresponding
to each variable in $z_{ij}$ is an indicator of missingness for that variable contained in the vector
$r_{ij}=(r_{ij1}, \ldots, r_{ijq})$ where $r_{ijl}=1$ if $z_{ijl}$ is missing and 0 otherwise, for $l=1, \ldots, q$. The selection
model allows specification of the joint distribution of $(r_i, y_i, z_i | x_i)$ with the goal of obtaining
unbiased estimates of the regression parameters. In general, the complete data joint
distribution of $(r_i, y_i, z_i | x_i)$ may be expressed as a series of conditional densities:

$$p(r, y, z | x, \beta, \alpha, \phi) = p_r(r | y, z, x, \phi) \times p_y(y | z, x, \beta) \times p_z(z | x, \alpha) \quad (2)$$

The parameters $\phi$ and $\alpha$ index the distribution of the missing data mechanism and the
missing covariates and are nuisance parameters which are not of inferential interest.

**Distribution of the missing data process.** Accounting for the nonignorability of the
missing data mechanism requires specification of a model for the missing data mechanism $r_i$,
which is assumed to be dependent upon the unobserved value the corresponding variable
would have taken if it were observed. As shown in Ibrahim, Lipsitz and Chen (24) and
Stubbendick and Ibrahim (25) modeling the missing data mechanism $r_i$ as a series of one
dimensional conditional distributions is effective at reducing the number of nuisance
parameters while maintaining correlation between the longitudinal observations and allowing
for non-monotone patterns of missingness (25, 26). For the joint distribution of $r_i$, we
specify a distribution for each $r_{ijl}$ sequentially conditioning over the other missingness
indicators at measurement \( j \), previous missingness indicators for all variables at all measurements prior to \( j \), the corresponding vector of completely observed and possibly missing covariates, \( x_{ij} \) and \( z_{ij} \), respectively, event time \( y_i \) and vector of parameters \( \phi_{ij} \):

\[
p_x(r_i | y_i, x_i, z_i) = p(r_{i(q)} | r_{i1}, \ldots, r_{i(q-1)}, r_{i(j-1)}, \ldots, r_{i1} | x_{i1}, z_{i1}, y_i, \phi_{ij}) \times \ldots
\]

\[
\times p(r_{i1(j-2)} | r_{i1(j-1)}, \ldots, r_{i1(q-1)}, y_i, \phi_{(j-1)q}) \times \ldots
\]

\[
\times p(r_{i11} \mid r_{i1}, r_{i2}, \ldots, r_{i11} | x_{i1}, z_{i1}, y_i, \phi_{1q}) \times \ldots \times p(r_{i11} \mid x_{i1}, z_{i1}, y_i, \phi_{11})
\] (3)

Sequentially conditioning on previous measurements approximates a correlation structure similar to what would be obtained using random effects models without the need to specify the random effect (25, 27). A series of logistic regressions may be used to model these conditional distributions as each \( r_{ij} \) is dichotomous. The contribution to the complete-data likelihood for subject \( i \) corresponding to the missing data mechanism is thus given by equation (3).

**Distribution of the observed failure times.** The distribution of observed failure times conditional on observed and unobserved covariates is given by the proportional hazards model defined in equation (1), but explicitly including the potentially missing variables \( z_{ij} \). The hazard function is then:

\[
\lambda(y_i | x_{ij}, z_{ij}, \beta, \beta_x, \lambda) = \lambda_x \exp(x_{ij}' \beta + z_{ij}' \beta_x) \quad \text{for} \quad y \in (s_{j-1}, s_j]
\]

where now \( \beta_x \) is the \( q \times 1 \) vector of coefficients on the vector of covariates \( z_{ij} \). The density for the observed failure time \( y_i \) is then:

\[
p_y(y_i | x_{ij}, z_{ij}, \beta, \beta_x, \lambda) = \left( \lambda_x \exp(x_{ij}' \beta + z_{ij}' \beta_x) \right)^{\delta_i} \exp(-\Lambda_i(y_i)) \exp(x_{ij}' \beta + z_{ij}' \beta_x)
\]

for \( y_i \in (s_{j-1}, s_j] \) with cumulative hazard function:

66
\[
\Lambda_j(y_i) = \left( (y_i - s_{j-1}) \lambda_j \exp(x_{ij}' \beta_1 + z_{ij}' \beta_2) + \sum_{g=1}^{j-1} (s_g - s_{g-1}) \lambda_g \exp(x_{ig}' \beta_1 + z_{ig}' \beta_2) \right).
\]

We further let \( \lambda = (\lambda_1, \ldots, \lambda_J) \)' denote the \( J \times 1 \) vector of baseline hazards \( \lambda_j \) and let \( \Delta_{ij} \) be an indicator of if subject \( i \) died or was censored in interval \( j \) (i.e. \( y_i \in (s_{j-1}, s_j] \)). The \( i^{th} \) contribution to the complete data likelihood for the piecewise exponential model is then:

\[
p_j(y_i \mid x_i, z_i, \beta_1, \beta_2, \lambda) = \prod_{j=1}^{J} \left( \lambda_j \exp(x_{ij}' \beta_1 + z_{ij}' \beta_2) \right)^{\Delta_{ij} \delta_i} \times 
\exp\left\{ -\Delta_{ij} \Lambda_j(y_i) \exp(x_{ij}' \beta_1 + z_{ij}' \beta_2) \right\}.
\]

If we define \( t_{ij} = \min(y_i, s_{j+1}) - s_j \) if \( y_i \geq s_j \) and 0 if \( y_i < s_j \) to be the length of the overlap from the beginning of interval \( j \) to the end of the interval or until failure time \( y_i \), then it can be shown that the likelihood function given by equation (4) is equivalent to one where \( \Delta_{ij} \delta_i \) follows a Poisson distribution with mean \( t_{ij} \lambda(y_i \mid x_{ij}, z_{ij}, \beta_1, \beta_2, \lambda_j) \).

**Distribution of missing data.** For the joint distribution of the missing covariates \( z_i \) we again follow the strategy suggested by Lipsitz and Ibrahim (28), Ibrahim et al. (24) and Stubbendick and Ibrahim (25) by specifying a sequence of one dimensional conditional distributions. We specify a model for each \( z_{il} \) sequentially conditioning over the other \( z \) variables at measurement \( j \), all \( z \) variables at previous times, the corresponding vector of completely observed covariates, \( x_{ij} \), event time \( y_i \), and \( \alpha=(\alpha_{l1}, \ldots, \alpha_{lj}) \)' where each \( \alpha_{lj} \) is a vector of parameters indexing the distribution for each covariate \( l \) for measurement \( j \). The joint distribution of the \( z \) variables for subject \( i \) is then:
Expression of the distribution of covariates this way allows considerable flexibility in choice of distribution for each \( z_{ijt} \), accommodating continuous and categorical variables, as well as offering a convenient way to account for intra-subject correlation without specification of a random effect. Once again, one should strive for a parsimonious specification of this joint distribution to avoid specification issues. Equation (5) then represents the \( i^{th} \) contribution to the marginal likelihood for \( z \).

Substituting equations (3), (4) and (5) into equation (2), yields the complete data likelihood:

\[
\ell(\beta_1, \beta_2, \lambda, \alpha, \phi) = \prod_{i=1}^{n} p(r_i, y_i, z_i | x_i, \beta_1, \beta_2, \lambda, \alpha, \phi) = \prod_{i=1}^{n} p_x(r_i | y_i, z_i, x_i, \phi) p_y(y_i | z_i, x_i, \beta_1, \beta_2, \lambda) p_z(z_i | x_i, \alpha) \]

with densities \( p_x(\cdot) \), \( p_y(\cdot) \) and \( p_z(\cdot) \) defined above. Previous work with similar models has made use of the expectation-maximization (EM) algorithm to obtain the parameter estimates (23). However here I propose a Fully Bayesian (FB) approach using vague priors on the parameters \( \beta_1, \beta_2, \lambda, \alpha \), and \( \phi \), which will produce estimates equivalent to the frequentist analysis using EM. The FB framework is also less computationally demanding than the EM framework for this model.
The joint posterior distribution of the parameters is proportional to the product of the conditional distribution of the observed data given the parameters and the joint prior distribution of the model parameters \( p(\beta_1, \beta_2, \lambda, \alpha, \phi) \):

\[
p(\beta_1, \beta_2, \lambda, \alpha, \phi \mid y, r, x, z) \propto \prod_{i=1}^{n} \int_{\mathbf{z}_i} \left( p(y_i, r_i, x_i, z_i \mid \beta_1, \beta_2, \lambda, \alpha, \phi) \right) d\mathbf{z}_i \times p(\beta_1, \beta_2, \lambda, \alpha, \phi) \quad (6)
\]

If non-informative priors are specified for \((\beta_1, \beta_2, \lambda, \alpha, \phi)\) then the posterior means and standard deviations of the parameters will be similar to maximum likelihood. In the analyses to follow, I used the Gibbs Sampler (29) to sample from the posterior distribution given by equation (6). Although somewhat computationally intensive (but less intensive than EM), the FB approach here provides a very straightforward way to estimate parameters from a complex model, especially variance and covariance parameters.

In the Bayesian paradigm inferences on model parameters (e.g. \(\beta_1, \beta_2, \lambda, \alpha, \phi\)) are obtained through estimation of their distribution conditional on the observed data, also referred to as the posterior distribution of the parameters. For notational convenience let \(\beta=(\beta_1, \beta_2)\) denote the vector of coefficients on both the fully observed and potentially missing covariates from the proportional hazards model. The posterior distribution is proportional to the product of the conditional distribution of the data \((x\text{ and } z)\) given the parameters and prior distributions on the model parameters:

\[
p(\beta, \lambda, \alpha, \phi \mid y, r, x, z) \propto p(y, r, x, z \mid \beta, \lambda, \alpha, \phi) \times p(\beta) \times p(\lambda) \times p(\alpha) \times p(\phi) . \quad (7)
\]

The conditional distribution of the data, \(p(y, r, x, z \mid \beta, \lambda, \alpha, \phi)\) is proportional to the likelihood function from the frequentist approach (equation 6) and the terms \(p(\beta), p(\lambda), p(\alpha)\) and \(p(\phi)\) represent the prior distributions on the model parameters, representing a priori expectations on them. The Bayesian approach then seeks to obtain means and variances for
the parameters of interest \((\beta, \lambda, \alpha, \phi)\) from their marginal posterior distributions. These Bayesian estimates are equivalent to frequentist point estimates for these parameters if non-informative distributions are specified for the priors (30). A non-informative distribution is one which exhibits little influence on the posterior, allowing the likelihood to dominate.

To solve for the parameters of this model I used the WinBUGS software program (31, 32) that makes use of the Gibbs sampler, an algorithm for generating samples from marginal posterior distributions without requiring knowledge of the form of the density (29). The Gibbs sampler allows the generation of a large sample of data on the parameters \((\beta, \lambda, \alpha, \phi)\) from which sample means and variances may be calculated in order to make statistical inferences. To obtain an empirical estimate of the joint posterior distribution \(p(\beta, \lambda, \alpha, \phi | y, r, x, z)\), the Gibbs sampler generates a series of draws of \((\beta, \lambda, \alpha, \phi)\) from the corresponding full conditional distributions [e.g. \(p(\beta | \lambda, \alpha, \phi, y, r, x, z)\)]. This sequence of draws on each variable ultimately converges in distribution to the marginal posterior distributions of interest (e.g. \(p(\beta | y, r, x, z)\)) from which estimates of the mean and variance may be calculated (29).

Specifically, the steps of the algorithm are as follows:

0. Specify a set of initial values of the parameters, \((\beta^{(0)}, \alpha^{(0)}, \phi^{(0)}, \lambda^{(0)})\)
1. Set the step counter \(k=1\).
2. Draw \(\beta^{(k)}\) from the conditional distribution \([\beta | \alpha^{(k-1)}, \phi^{(k-1)}, \lambda^{(k-1)}, y, r, x, z]\)
3. Draw \(\lambda^{(k)}\) from the conditional distribution \([\lambda | \beta^{(k)}, \alpha^{(k-1)}, \phi^{(k-1)}, y, r, x, z]\)
4. Draw \(\alpha^{(k)}\) from the conditional distribution \([\alpha | \beta^{(k)}, \lambda^{(k)}, \phi^{(k-1)}, y, r, x, z]\)
5. Draw \(\phi^{(k)}\) from the conditional distribution \([\phi | \beta^{(k)}, \lambda^{(k)}, \alpha^{(k)}, y, r, x, z]\)
6. Set \(k=k+1\). Repeat step 2, a total of \(N\) times.
7. The sequence of \(N\) “observations” of \((\beta^{(k)}, \lambda^{(k)}, \alpha^{(k)}, \phi^{(k)})\) will then essentially consist of a sample from the marginal posterior distributions for each variable.

From equation 7, one can see that the full conditional distributions in this algorithm are proportional to the likelihood multiplied by the prior distribution, for example:
\[ p(\beta \mid \lambda, \alpha, \phi, y, r, x, z) \approx p(y, r, x, z \mid \beta, \lambda, \alpha, \phi) \times p(\beta). \]

For the distributions of the exposures \((x, z)\), probability of missingness \((r)\) and event time \((y)\)
I use the forms described above. For the parameters from the proportional hazards regression
\((\beta, \lambda)\), logistic regression \((\phi)\) and missing data \((\alpha)\) models I specify prior distributions as
normal with mean zero and large variance:

\[ \beta_i \sim N(0, \sigma_{\beta}) \quad \forall \ i = 1, \ldots, p \]
\[ \phi_i \sim N(0, \sigma_{\phi}) \quad \forall \ i = 1, \ldots, s \]
\[ \alpha_i \sim N(0, \sigma_{\alpha}) \quad \forall \ i = 1, \ldots, q \]

with \(\sigma\) on the order of 1,000,000 while the each of the baseline hazards from the proportional
hazards regression, \(\lambda\), are assumed to follow independent Gamma distributions with
common parameters. I ran the Gibbs sampler for several thousand iterations to achieve
convergence (usually 25,000-30,000 iterations, depending on the complexity of the model),
and then an additional 25,000-30,000 iterations after convergence to generate a sample for
inference. Estimates of regression parameters are obtained by calculating the mean value for
each sample of parameters (e.g. the mean of the sample for each of the \(\beta\) parameters), with
the 2.5\(^{th}\) and 97.5\(^{th}\) percentiles defining the 95% credible intervals, the Bayesian analog to
frequentist confidence intervals (30). For substantive interpretation I exponentiate (anti-log
transform) the parameters and limits on the credible intervals. Although the credible
intervals are theoretically distinct from frequentist confidence intervals (as they represent the
interval which has a 95% probability of containing the true parameter value), the use of non-
informative priors yields interval estimates that are similar to standard frequentist confidence
intervals.
Confounding

Model selection procedures have not been developed for missing data problems, and most importantly the iterative inclusion and exclusion of covariates into the models will fundamentally affect the assumptions regarding the missing data process. This fact makes it impossible to disentangle the confounding effect of an included variable with the effect of the associated change in the missing data model. Consequently, potential confounders were examined using a directed acyclic graph and confirmed through a change-in-estimate approach on the complete dataset of 1,033 subjects. Although this method is less than ideal it should still successfully identify confounders of the associations of interest. Potential confounders for the body size and physical activity analyses were those variables that are thought to be associated with both post-diagnosis weight change and mortality or post-diagnosis physical activity levels and mortality, respectively. Separately for all-cause and breast cancer specific mortality I used a change-in-estimate approach on the complete dataset to identify potential confounding variables for the final analysis. Using a Cox proportional hazards regression model each variable was independently examined for its impact on the relationship between exposure and time to death. Those variables that change the natural logarithm of the ratio of the adjusted to unadjusted hazard ratio (lnCoHR = ln[adjusted HR/unadjusted HR]) by 10% or more were included in a multivariate model.

For the analysis of post-diagnosis weight change potential confounders included those variables potentially associated with both post-diagnosis weight change and survival (33-38): post-diagnosis physical activity (metabolic equivalent task hours per week, MET-hrs/wk; continuous), total energy at baseline interview (kcal/day; continuous), menopausal status (premenopausal, postmenopausal), stage (in situ, invasive), chemotherapy regimen (yes/no), radiation therapy (yes/no), hormone therapy (yes/no), nodal involvement (yes/no), tumor size
greater than 2 cm (yes/no), estrogen receptor positive tumor (ER status; yes/no) and progesterone receptor positive tumor (PR status; yes/no). The final confounder list included age, chemotherapy regimen, tumor size, ER status and PR status, which was also consistent with identification of confounders using the analysis of a directed acyclic graph. In addition to weight change, other covariates with significant amounts of missing data included chemotherapy, tumor size, ER status and PR status with 32.2%, 31.6%, 34.0% and 34.3% missing out of the 1,508 cases, respectively. Models for the variables chemotherapy, ER status, PR status and tumor size were included in the selection model along with the model for weight change. Treatment and tumor characteristics were unlikely to be non-ignorably missing and therefore did not require specification of models for their missing data mechanisms. Variables with minor amounts of missing data included menopausal status (1.99% missing), pre-diagnosis BMI (1.13% missing), adult weight change (from age 20 to 1 year before diagnosis; 1.66% missing), education (0.40% missing) and income (0.27% missing); so these few subjects were excluded from the analysis as these amounts were unlikely to influence the results. The analysis of weight change ultimately included 1,436 women, 292 who died during follow-up with 156 of these deaths attributed to breast cancer.

For post-diagnosis physical activity, potential confounders included variables that have been shown to be related to both survival and post-diagnosis physical activity levels, and included age, pre-diagnosis body mass index (BMI, weight in kg/squared height in meters) in the year prior to diagnosis, chemotherapy treatment (yes/no), radiation therapy (yes/no) and hormone therapy (yes/no) (39, 40). I also considered menopausal status at diagnosis (premenopausal/postmenopausal), total energy intake in the year prior to the baseline interview, stage (in situ vs. invasive), ER status (yes/no), PR status (yes/no), tumor
size (< 2 cm vs. >= 2 cm) and nodal involvement (yes/no). The final models included age, chemotherapy treatment, radiation treatment, hormone therapy and tumor size, which were also consistent with assessment of confounding using a directed acyclic graph. In addition to physical activity, other covariates with significant amounts of missing data included chemotherapy, radiation treatment, hormone therapy and tumor size with 32.2%, 33.0%, 32.1%, 31.6% missing, respectively. For each of these variables, models were incorporated into the analysis to impute their values, similar to the model required for physical activity levels. Other variables with missing data included baseline physical activity (0.93%) menopausal status (1.99% missing), pre-diagnosis BMI (1.13% missing), adult weight change (from age 20 to 1 year before diagnosis; 1.66% missing), education (0.40% missing) and income (0.27% missing). These small amounts were unlikely to influence the results and so the additional cost of estimating models for each of these variables exceeded any benefit. I therefore excluded subjects with any of these variables missing. Our final analysis then included 1,423 women (94.4% of the original 1,508 cases). Among these women, 291 died during the follow-up with 156 of these deaths due to breast cancer.

**Effect Measure Modification**

Given the low number of events and the influence of missing data, the power for examining effect modification in this study was limited and therefore I examined interactions in an exploratory and descriptive manner using stratified analysis. For the analysis of post-diagnosis weight change, I examined effect modification by baseline body mass index (BMI; <25 kg/m² vs. >= 25 kg/m²), adult weight gain from age 20 years to 1 year before reference date (>3 kg loss, maintain within 3 kg and >3 kg gain) as well as the interaction with time (before and after 2 years post-diagnosis), which additionally captured possible departure from the proportionality assumption. For the physical activity analysis I stratified results by
pre-diagnosis physical activity levels (0 MET-hours/week, 0.1-9.0 MET-hours/week, >9.0 MET-hours/week), baseline (BMI <25 kg/m$^2$ vs. >= 25 kg/m$^2$) and various measures of time (before and after 1 year post-diagnosis; before and after 2 years post-diagnosis). Formal tests for interactions were not conducted as these were considered exploratory and $p$-values are not calculated in the Bayesian paradigm.

**Sensitivity Analysis**

One benefit of the selection model approach is the ability to use it to easily perform sensitivity analysis by altering the form of the models for probability of missingness. As in Herring et al. (23) I compared the results from models with assumptions (1) that the probability of missing is independent of observed and unobserved data (MCAR) using a complete-case analysis in Stata v. 10.0, as well as (2) that the probability of missing is dependent on observed covariates, but not unobserved data (MAR), formally omitting specification of the probability of missingness (equation (3)) and (3) that the probability of missing is dependent upon observed and missing data (non-ignorable missing), requiring explicit specification of equation (3).

The results of this sensitivity analysis are presented in chapter 3. I only conducted a complete sensitivity analysis on the analysis of post-diagnosis weight change, and I estimated models that assumed a non-ignorably missing data mechanism for the detailed weight change and physical activity chapters (chapters 4 and 5). It was encouraging that results from the weight change analysis appeared largely invariant to changes in the MAR and NMAR assumptions, as we believed weight change to be more likely to be non-ignorably missing given that it is such a sensitive issue.
**Study Power**

Although the final analysis was performed in a Fully Bayesian framework, and I did not conduct formal hypothesis testing, since the proposed analysis will yield estimates similar to the frequentist approach, a power analysis may provide a general idea of the ability of this study to detect certain effect sizes with credible intervals that exclude the null. Given the significance of missing data for this study and the fact that power calculations do not exist for missing data models, I show power calculations for two different scenarios: for the minimum sample size (e.g. if a complete case analysis were to be conducted among those completing the follow-up questionnaire, n=1033) and for the maximum sample size (e.g. assuming complete follow-up data for the entire cohort, n=1436 for weight change analysis and n=1423 for physical activity analysis, after omitting subjects with missing data on confounders as described above) which will provide the most useful information for this analysis. A significance level of 5% will be assumed for hypothesis tests and calculations are based on the logrank test using the sample size routine in Stata 10.0 (Stata Corporation, College Station, TX). For each of the aims, I assumed distributions of weight change and physical activity levels recently published in similar follow-up studies, applied to the expected sample sizes in this cohort.

**Aim 1. Determine if post-diagnosis changes in bodyweight are associated with breast cancer survival:** Power calculations for this aim are based on a comparison of survival rates between those in the weight maintenance group and those in the group corresponding to the highest gain. Applying the distribution of weight change observed in the analysis by Caan et al. (14) (>5% loss: 17%, maintenance within 5%: 47%, 5-10% gain: 16%, >10% gain: 19%) to the sample sizes presented here, for the complete data sample of 1,033 subjects I calculate that there will be 489 and 200 subjects in the maintenance and highest gain group,
respectively. For the full sample of 1,436 subjects I calculated 679 and 278 subjects in the maintenance and highest gain group, respectively.

Table 2 shows power calculations comparing survival times using a 2-sided log-rank test to compare survival in these groups; as I expect a deleterious effect of increasing weight the comparisons are presented assuming hazard ratios greater than 1.0. Power to compare maintainers to those with the greatest weight gain was excellent for modest associations (>95% power to detect a HR of at least 1.5).

Aim 2. Determine if post-diagnosis levels of physical activity are associated with breast cancer survival: Power calculations for this aim will be based on a comparison of survival rates between those in the group of lowest post-diagnosis physical activity and those in the group corresponding to the highest level of post-diagnosis physical activity and as before, I present results comparing to a larger group consisting of the two greatest levels of activity combined. I used the distribution of post-diagnosis physical activity levels reported by Irwin et al. (17) (0 MET-hours/week: 17%, 0.1-9.0 MET-hours/week: 43%, >9.0 MET-hours/week: 40%) applied to the sample sizes here. These calculations yield 171 subjects in the lowest level, 416 subjects in the highest level for the complete data sample of 1,033. For the full sample of 1,423 subjects I calculated 236 and 573 subjects in the lowest and highest activity group, respectively. Power calculations for this aim will be presented assuming a protective effect of increasing activity, or for hazard ratios less than 1.0. As shown in table 3, power to compare the lowest to those with the highest levels of post-diagnosis physical activity was excellent for modest associations (>92% power to detect a HR of 0.75 or less).

Strengths and Limitations

The primary strength of this study is that the question of how post-diagnosis levels in body size and physical activity levels affect survival after breast cancer diagnosis is largely
unresolved. Results from this analysis address important public health questions, which will impact on lifestyle recommendations for women who are diagnosed with the disease. The physical activity questionnaire used in this study is comprehensive in that it allows study subjects to specify activities in an open-ended format and allows for accurate assessment of activity over a wide span of time (8); consequently, it is considered to be the best assessment tool for population-based studies of physical activity and breast cancer etiology (41).

Another significant advantage of this study is the novel approach employed to account for the missing follow-up data in this cohort. This technique is unique and is applicable to outcomes other than survival. An additional benefit of this approach is that it allows a straightforward method for conducting sensitivity analyses to evaluate the effect of different missing data assumptions (MCAR, MAR or non-ignorable) on the parameter estimates. Use of the Gibbs Sampler through software such as BUGS (31), although computationally intensive, is straightforward and should be accessible to the methodologically astute epidemiologist.

The cases for this study form a population-based cohort, and although not generalizable to the entire United States population will yield valuable data on the relationship between lifestyle factors and breast cancer survival. A notable weakness of this study is that the assessments used for the primary exposures (height, weight, physical activity) are self-reported and as such may tend to be reported with error. Anthropometric measures in particular tend to be mis-reported with heavier people tending to underreport weight, while lighter people tend to overreport (42). Lastly, the presence of a substantial portion of missing data is a weakness of this dataset proposed for this work. Techniques to account for missing data are useful, however not a substitute for having the data that was not observed. Although an analytical approach is not as ideal as having the data it is an
appropriate strategy, yielding statistically unbiased estimators and making efficient use of existing resources.
Summary

To examine the association between post-diagnosis changes in body size, levels of physical activity and survival after diagnosis with breast cancer I used data from the Long Island Breast Cancer Study Follow-up, a population-based cohort of 1,508 women diagnosed with breast cancer during 1996 and 1997 from Long Island, New York. Data was collected shortly after diagnosis and again approximately five years later and includes anthropometric and physical activity data as well as sociodemographic, medical and lifestyle variables. Date and cause of death were ascertained using the National Death Index. Out of the 1,508 subjects completing the baseline interview only 1,033 completed the follow-up assessments resulting in a large proportion of missing exposure data. To address this issue I outlined a selection model from the biostatistics literature that provides a flexible framework in which to model the missing data process. I used time-to-event models to relate post diagnosis changes in body size and post diagnosis levels of physical activity to time of death and specify a model to describe the distribution of the exposures in the total study population. A model to explain the probability that an observation is missing is also specified; by modifying the form of this model I performed sensitivity analysis to explore the impact of different assumptions on the mechanisms that lead to the missing data.
Table 2.1. Selected characteristics of breast cancer cases at diagnosis (n=1,508), Long Island Breast Cancer Study Project, 1996-1997.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Cases</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age at diagnosis (yrs)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;40</td>
<td>89</td>
<td>5.9</td>
</tr>
<tr>
<td>40-&lt;50</td>
<td>318</td>
<td>21.1</td>
</tr>
<tr>
<td>50-&lt;60</td>
<td>410</td>
<td>27.2</td>
</tr>
<tr>
<td>60-&lt;70</td>
<td>353</td>
<td>23.4</td>
</tr>
<tr>
<td>&gt;=70</td>
<td>338</td>
<td>22.4</td>
</tr>
<tr>
<td><strong>Menopausal Status</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Premenopausal</td>
<td>472</td>
<td>31.9</td>
</tr>
<tr>
<td>Postmenopausal</td>
<td>1,006</td>
<td>68.1</td>
</tr>
<tr>
<td>Missing</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>1,411</td>
<td>93.8</td>
</tr>
<tr>
<td>Other</td>
<td>94</td>
<td>6.2</td>
</tr>
<tr>
<td>Missing</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td><strong>Hispanic Ethnicity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>57</td>
<td>3.8</td>
</tr>
<tr>
<td>No</td>
<td>1,448</td>
<td>96.2</td>
</tr>
<tr>
<td>Missing</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than High School</td>
<td>183</td>
<td>12.2</td>
</tr>
<tr>
<td>High School</td>
<td>538</td>
<td>35.8</td>
</tr>
<tr>
<td>At Least Some College</td>
<td>781</td>
<td>51.8</td>
</tr>
<tr>
<td>Missing</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td><strong>Marital Status</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ever Married</td>
<td>1,443</td>
<td>95.8</td>
</tr>
<tr>
<td>Never Married</td>
<td>64</td>
<td>4.2</td>
</tr>
<tr>
<td>Missing</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><strong>Parity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nulliparous</td>
<td>198</td>
<td>13.1</td>
</tr>
<tr>
<td>Parous</td>
<td>1310</td>
<td>86.9</td>
</tr>
<tr>
<td><strong>Alcohol use</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>588</td>
<td>39.0</td>
</tr>
<tr>
<td>Ever</td>
<td>920</td>
<td>61.0</td>
</tr>
<tr>
<td><strong>Smoking</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>675</td>
<td>44.8</td>
</tr>
<tr>
<td>Former</td>
<td>543</td>
<td>36.0</td>
</tr>
<tr>
<td>Current</td>
<td>290</td>
<td>19.2</td>
</tr>
<tr>
<td>Missing</td>
<td>47</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.1 (continued).  Selected characteristics of breast cancer cases at diagnosis (n=1,508), Long Island Breast Cancer Study Project, 1996-1997.

<table>
<thead>
<tr>
<th>Family History Breast Cancer</th>
<th>Yes</th>
<th>No</th>
<th>First Degree</th>
<th>295</th>
<th>19.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hormone Replacement Therapy Use</td>
<td>Never</td>
<td>1,096</td>
<td>72.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ever</td>
<td>408</td>
<td>27.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMI at reference date</td>
<td>&lt;25</td>
<td>683</td>
<td>45.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>25-29.9</td>
<td>476</td>
<td>31.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;=30</td>
<td>332</td>
<td>22.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage</td>
<td>In situ</td>
<td>235</td>
<td>15.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Invasive</td>
<td>1,273</td>
<td>84.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nodal Involvement</td>
<td>No 463</td>
<td>77.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes 134</td>
<td>22.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>911</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tumor size</td>
<td>&lt; 2 cm</td>
<td>749</td>
<td>76.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;= 2 cm</td>
<td>234</td>
<td>23.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>453</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ER status</td>
<td>Negative 265</td>
<td>26.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>730</td>
<td>73.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>513</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PR status</td>
<td>Negative 355</td>
<td>35.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>636</td>
<td>64.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>517</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Radiation Treatment</td>
<td>Yes</td>
<td>625</td>
<td>60.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No 401</td>
<td>39.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>482</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemotherapy Treatment</td>
<td>Yes 423</td>
<td>41.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No 599</td>
<td>58.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>486</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2.1 (continued).  Selected characteristics of breast cancer cases at diagnosis (n=1,508), Long Island Breast Cancer Study Project, 1996-1997.

<table>
<thead>
<tr>
<th>Hormone Therapy Treatment</th>
<th>Yes</th>
<th>No</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>616</td>
<td>393</td>
<td>499</td>
</tr>
<tr>
<td></td>
<td>61.1</td>
<td>38.9</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.2. Power to detect varying hazard ratios for weight change, calculations based on the log-rank test statistic, assuming 5% significance level, complete data (n=1033 total; maintenance n=489; highest gain n=200) and full sample (n=1436 total; maintenance n=679; highest gain n=278).

<table>
<thead>
<tr>
<th>Min. Detectable Hazard Ratio</th>
<th>Power Highest level vs. Lowest level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Complete Data</td>
</tr>
<tr>
<td>1.33</td>
<td>95%</td>
</tr>
<tr>
<td>1.5</td>
<td>99%</td>
</tr>
<tr>
<td>1.75</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 2.3. Power to detect varying hazard ratios for post-diagnosis levels of physical activity categorized to three levels, calculations based on the log-rank test statistic, complete data (n=1033 total; inactive: 0 MET-hours/week n=171; highest activity: >9.0 MET-hours/week n=416) and full sample (n=1423 total; inactive: 0 MET-hours/week n=236; highest activity: >9.0 MET-hours/week n=573).

<table>
<thead>
<tr>
<th>Min. Detectable Hazard Ratio</th>
<th>Power Highest level vs. Lowest level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Complete Data</td>
</tr>
<tr>
<td>0.75</td>
<td>92%</td>
</tr>
<tr>
<td>0.67</td>
<td>99%</td>
</tr>
<tr>
<td>0.50</td>
<td>100%</td>
</tr>
<tr>
<td>0.33</td>
<td>100%</td>
</tr>
<tr>
<td>0.25</td>
<td>100%</td>
</tr>
</tbody>
</table>
References


CHAPTER 3
A PROPORTIONAL HAZARDS REGRESSION MODEL WITH NON-IGNORABLY MISSING TIME-VARYING COVARIATES

Abstract

Missing covariate data is common in observational studies of time to an event, especially when covariates are repeatedly measured over time. Failure to account for the missing data can lead to bias or loss of efficiency, especially when the data are non-ignorably missing. Previous work has focused on the case of fixed covariates rather than those that are repeatedly measured over the follow-up period, so here we present a selection model that allows for proportional hazards regression with time-varying covariates when some covariates may be non-ignorably missing. We develop a fully Bayesian model and obtain posterior estimates of the parameters via the Gibbs sampler in WinBUGS. We illustrate our model with an analysis of post-diagnosis weight change and survival after breast cancer diagnosis in the Long Island Breast Cancer Study Project (LIBCSP) follow-up study. Our results indicate that post-diagnosis weight gain is associated with lower all-cause and breast cancer specific survival among women diagnosed with new primary breast cancer. Our sensitivity analysis showed only slight differences between models with different assumptions on the missing data mechanism yet the complete case analysis yielded markedly different, and sometimes implausible results.

Introduction

Studies of survivorship are often plagued by missing covariate data, especially when assessments are made longitudinally and deal with lifestyle or behavioral characteristics that
may be sensitive in nature. A formal treatment of missing data requires consideration of the process that leads to the incomplete observations, such as the taxonomy suggested by Little and Rubin (1). Data is considered missing completely at random (MCAR) if the probability that data is missing is independent of both observed and unobserved data. Under this scenario the observed data essentially constitutes a random sample of values from all subjects and thus a complete-case analysis, which uses data only on those subjects with no missing observations, will yield unbiased parameter estimates. If the probability that data is missing depends only upon fully observed variables then the data is referred to as missing at random (MAR). The missing data mechanism is often called ignorable in this case. The most problematic situation arises when the probability that data is missing depends upon unobserved values of the missing variable, which is what we suspect for our data. When the probability that a variable is missing depends upon its unobserved value then the data is referred to as not missing at random (NMAR) and the missing data mechanism is referred to as non-ignorable. Valid estimation under non-ignorable missingness requires simultaneously accounting for the probability that data is missing, the distribution of the values of the missing variable and the relationship between the potentially incomplete variable to the outcome of interest.

The majority of the literature on missing covariates in proportional hazards regression has focused on frequentist methods for baseline MAR covariates (2-10). The Bayesian approach to survival analysis with covariate data that is MAR is described in detail by Ibrahim, Chen and Sinha (11). Frequentist methods for non-ignorably missing covariates in survival analysis have been presented by Leong, Lipsitz and Ibrahim (12) and Herring, Ibrahim and Lipsitz (13), however both of these methods apply only to baseline (fixed)
covariates, and the method by Leong and colleagues requires them to also be dichotomous. The selection model outlined by Herring et al. (13) specifies the joint distribution of the survival times, missing covariates and missingness indicator through a series of one-dimensional conditional distributions and uses a Monte Carlo Expectation Maximization (MCEM) algorithm for parameter estimation. Here, we propose a model that extends this approach by allowing the covariate values to vary over time, and we present a computationally easier alternative to the MCEM algorithm for parameter estimation.

The motivation for developing this model was lies in our interest in identifying factors that may be associated with survival among women with breast cancer. Our objective is to evaluate how post-diagnosis changes in weight over time affect survival through an analysis of data from the follow-up to the LIBCSP (14). An issue for this analysis is that a significant portion of the study subjects are missing data for one or more follow-up assessments of bodyweight making bias or loss of efficiency of serious concern if we limit our investigation to only those subjects with complete data. Specifically, given the stigma associated with being overweight, we suspect that those subjects with missing data on body size may tend to be heavier than those who responded, making this subset of subjects a less than representative sample of the study population. In this case standard proportional hazards models would be inappropriate. With repeated measurements of body size at and after diagnosis, our primary covariate of interest is time-varying, requiring us to develop the model we present here. To our knowledge there has been no previous work addressing inference for selection models with non-ignorably missing time-varying covariates.

The following section of the paper will outline our notation and describe the selection model in general. We then describe specific models for each of the conditional distributions:
the missingness indicator given time-to-event and covariates, the time-to-event model given
 covariates and the distribution of the missing covariates. We then describe our estimation
 approach and then illustrate our model with an example of an analysis of changes in
 bodyweight over time and survival after breast cancer diagnosis, using data from the follow-
 up to the Long Island Breast Cancer Study (14). We conclude with a discussion of the results
 and the methodology.

The Selection Model

Here we outline a selection model for proportional hazards regression with time varying
covariates, which is defined by the joint distribution of the event times, missing covariates
and the mechanism that describes the probability of missingness. This joint distribution is
specified through a series of conditional distributions: (1) the probability the covariate data
is missing conditional on event time and (possibly unobserved) covariates, (2) the
distribution of event time conditional on covariates and (3) the marginal distribution of the
missing covariates dependent only on fully-observed variables. We begin by outlining the
notation for the model.

Assume we have data on a sample of \( n \) independent subjects and for subject \( i \) denote
the event time by \( T_i \) and censoring time by \( C_i \). For each of the \( n \) subjects we observe the
variable \( y_i = \min(T_i, C_i) \) and indicator of failure \( \delta_i \) which takes on the value 1 if \( y_i \)
corresponds to an occurrence of an event (i.e. \( T_i \leq C_i \)), and 0 if it represents a censored
observation (i.e. \( T_i > C_i \)). We further assume independence between \( T_i \) and \( C_i \). Each
subject provides a series of longitudinal measurements for \( p + q \) variables where for the \( k^{th} \)
measurement we denote the vector of \( p \) completely observed variables by \( x_{ik} = (x_{ik1}, \ldots, x_{ikp}) \)
and \( q \) variables with potentially missing values by \( z_{ik} = (z_{ik1}, \ldots, z_{ikq}) \) measured at times \( v_{ik} \)
for \( k = 1, \ldots, K_i \), where \( K_i \geq 1 \). Elements of the \( \mathbf{z}_{ik} \) vector are missing for only some subjects at some of the measurement points so associated with each variable in \( \mathbf{z}_{ik} \) is an indicator of missingness for that variable contained in the vector \( \mathbf{r}_{ik} = (r_{ik1}, \ldots, r_{ikq}) \) where \( r_{ikl} = 1 \) if \( z_{ikl} \) is missing and \( r_{ikl} = 0 \) otherwise, for \( l = 1, \ldots, q \) and \( k = 1, \ldots, K_i \). The notation \( \mathbf{x}_i \), \( \mathbf{z}_i \), and \( \mathbf{r}_i \) will refer to matrices of size \( K_i \times p \), \( K_i \times q \), and \( K_i \times q \) respectively, representing the set of all \( K_i \) measurements for each vector of variables for each subject \( i \).

In general, estimation of a proportional hazards regression of \( y \) on \( [\mathbf{x} \mathbf{z}] \) using complete covariate data (only where \( r_{ikl} = 1 \)) will yield biased estimates of the regression parameters as it does not account for the distribution of the missing variables and more importantly the possibility that the reason they are missing may be related to their unobserved values. The selection model allows us to specify the joint distribution of \( (\mathbf{r}_i, y_i, \mathbf{z}_i | \mathbf{x}_i) \) allowing us to account for these relationships with the goal of obtaining unbiased estimates of the regression parameters. In general, the complete data joint distribution of \( (\mathbf{r}_i, y_i, \mathbf{z}_i | \mathbf{x}_i) \) may be expressed as a series of conditional densities:

\[
p(\mathbf{r}_i, y_i, \mathbf{z}_i | \mathbf{x}_i, \phi, \alpha) = p_r(\mathbf{r}_i | y_i, \mathbf{z}_i, \mathbf{x}_i, \phi) \times p_y(y_i | \mathbf{z}_i, \mathbf{x}_i, \beta) \times p_z(\mathbf{z}_i | \mathbf{x}_i, \alpha)
\]

The parameters \( \phi \) and \( \alpha \) index the distribution of the missing data mechanism and the missing covariates and are nuisance parameters which are not of inferential interest. The remainder of this section describes the specification of these conditional densities and the form of the complete data likelihood in detail.

**Models for the missing data mechanism**

The assumption of nonignorability of the missing data process requires specification of the distribution of the probability of missingness, which is assumed to be dependent upon the
unobserved value the corresponding variable would have taken if it were observed. We follow Ibrahim, Lipsitz and Chen (15) and Stubbendick and Ibrahim (16) by modeling the missing data mechanism $r_i$ as a series of one dimensional conditional distributions, which is effective at reducing the number of nuisance parameters while maintaining correlation between the longitudinal observations and allowing for non-monotone patterns of missingness (16, 17). For the joint distribution of $r_i$, we specify a distribution for each $r_{ikl}$ sequentially conditioning over the other missingness indicators at measurement $k$, previous missingness indicators for all variables at all measurements prior to $k$, the corresponding vector of completely observed and possibly missing covariates, $x_{ik}$ and $z_{ik}$, respectively, event time $y_i$ and vector of parameters $\phi_{kl}$:

$$
p_{\epsilon}(r_i \mid y_i, x_i, z_i) = p(r_{iKq} \mid r_{iK,1}, \ldots, r_{iK(q-1)}, r_{iK(q-1)}, \ldots, r_{iK(q)}, x_i, z_i, y_i, \phi_{Kq}) \times \ldots \times p(r_{i1} \mid r_{i1,1} \times \ldots, r_{i1(q-1)}, x_i, z_i, y_i, \phi_{1}) \times \\
p_{\epsilon}(r_{iKq} \mid r_{i1,1} \times \ldots, r_{i1(q-1)}, x_i, z_i, y_i, \phi_{1}) \times \\
p_{\epsilon}(r_{iKq} \mid r_{i1,1} \times \ldots, r_{i1(q-1)}, x_i, z_i, y_i, \phi_{1}) \times \ldots \times p(r_{i1} \mid x_i, z_i, y_i, \phi_{1})
$$

(2)

Sequentially conditioning on previous measurements approximates a correlation structure similar to what would be obtained using random effects models without the need to specify the random effect (16, 18). A series of logistic regressions may be used to model these conditional distributions as each $r_{ikl}$ is dichotomous. The contribution to the complete-data likelihood for subject $i$ corresponding to the missing data mechanism is thus given by equation (2).

Although the specification above appears quite complicated, in practice the number of measurements $K_i$ are likely to be small and it may be realistic to assume that only some
subset of the variables in $z$ are non-ignorably missing, and therefore the number of variables requiring specification of missingness models is fewer than $q$. Although it may be tempting to include a large number of variables and cross-products into the missing data models the analyst should strive for the most parsimonious specification possible as these models can easily become unidentifiable (13, 17, 19). Herring et al. (13) and Ibrahim, Zhu and Tang (20) suggest a strategy for model selection for the missing data mechanism of these models to help avoid issues of identifiability.

**Model for the time-to-event**

We consider here a Cox piecewise exponential hazard model to describe the relationship between event time and the covariates. To define the piecewise exponential model, we divide the time axis into $J$ discrete intervals $(s_{j-1}, s_j]$ for $j = 1, \ldots, J$ with $s_0 = 0$ and $s_J$ greater than the maximum of the $\{y_i\}$. The measurement times for the covariate vector are assumed to fall at the boundaries of the intervals although it is possible for a measurement to span multiple intervals (e.g. if measurements on $[x \ z]$ are taken every 2 years but the intervals $(s_{j-1}, s_j]$ correspond to 1 year each). Thus, since the number of covariate measurements $K_i \leq J$ then we define a notation so the indexes on each of the covariates match the index for the intervals of the piecewise exponential model. Then for subject $i$ within interval $j$, we define $x^*_i = (x_{ij1}^*, x_{ij2}^*, \ldots, x_{ijp}^*)'$ where $x_{ijl}^* = x_{ikl}$ and $z^*_i = (z_{ij1}^*, z_{ij2}^*, \ldots, z_{ijp}^*)'$ where $z_{ijl}^* = z_{ikl}$ for $k$ and $l$ such that $v_{ik} \leq s_{j-1} < s_{j} < v_{i,k+1}$. For example, assume that covariate measurements are made at year 1 and year 3, yet the intervals for the piecewise exponential model correspond to one year each. Then the covariate values at the second year ($x_{i2}^*$) will be carried forward from the first year, while those from the third year will reflect the
measurements taken at that time. Thus \( x_{ij}^* \) denotes the \( p \times 1 \) vector of fully observed covariate values and \( z_{ij}^* \) denotes the \( q \times 1 \) vector of possibly missing covariate values corresponding to the \( j^{th} \) interval for \( j = 1, \ldots, J \). We then define the piecewise exponential hazards model with the hazard function:

\[
\lambda(y_i | x_{ij}^*, z_{ij}^*, \beta_1, \beta_2, \lambda_j) = \lambda_j \exp\left( x_{ij}^* \beta_1 + z_{ij}^* \beta_2 \right) \text{ for } y_i \in (s_{j-1}, s_j]
\]

where \( \beta_1 \) is the \( p \times 1 \) vector of coefficients on the vector of covariates \( x_{ij}^* \) and \( \beta_2 \) is the \( q \times 1 \) vector of coefficients on the vector of covariates \( z_{ij}^* \). The density for the observed failure time \( y_i \) is then:

\[
p_y(y_i | x_{ij}^*, z_{ij}^*, \beta_1, \beta_2, \lambda_j) = \left( \lambda_j \exp( x_{ij}^* \beta_1 + z_{ij}^* \beta_2 ) \right)^{d_i} \exp(-\Lambda_j(y_i)) \exp( x_{ij}^* \beta_1 + z_{ij}^* \beta_2 )
\]

for \( y_i \in (s_{j-1}, s_j] \) with cumulative hazard function:

\[
\Lambda_j(y_i) = \left( (y_i - s_{j-1}) \lambda_j \exp( x_{ij}^* \beta_1 + z_{ij}^* \beta_2 ) + \sum_{g=1}^{j-1} (s_g - s_{g-1}) \lambda_g \exp( x_{ig}^* \beta_1 + z_{ig}^* \beta_2 ) \right).
\]

We further let \( \lambda = (\lambda_1, \ldots, \lambda_J)' \) denote the \( J \times 1 \) vector of baseline hazards \( \lambda_j \) and let \( \Delta_{g} \) be an indicator of if subject \( i \) died or was censored in interval \( j \) (i.e. \( y_i \in (s_{j-1}, s_j] \)). The \( i^{th} \) contribution to the complete data likelihood for the piecewise exponential model is then:

\[
p_y(y_i | x_i, \beta_1, \beta_2, \lambda) = \prod_{j=1}^{J} \left( \lambda_j \exp( x_{ij}^* \beta_1 + z_{ij}^* \beta_2 ) \right)^{d_i} \times \exp\left\{ -\Delta_{g} \Lambda_j(y_i) \exp( x_{ij}^* \beta_1 + z_{ij}^* \beta_2 ) \right\}
\]

where \( x_{ij}^* = x_{ik} \) and \( z_{ij}^* = z_{ik} \) with \( k \) and \( j \) such that \( v_{ik} \leq s_{j-1} < v_{i,k+1} \). If we define

\[ t_{ij} = \min(y_i, s_{j+1}) - s_j \text{ if } y_i \geq s_j \text{ and } 0 \text{ if } y_i < s_j \]

to be the length of the overlap from the beginning of interval \( j \) to the end of the interval or until failure time \( y_i \), then it can be shown
that the likelihood function given by equation (3) is equivalent to one where \( \Delta_j \delta_i \) follows a Poisson distribution with mean \( t_j \lambda (y_i \mid z_{ij}^*, x_{ij}^*, \beta_1, \beta_2, \lambda_j) \).

**Models for Missing Covariates**

For the joint distribution of the missing covariates \( z_i \) we again follow the strategy suggested by Lipsitz and Ibrahim (21), Ibrahim et al. (15) and Stubbendick and Ibrahim (16) by specifying a sequence of one dimensional conditional distributions. We specify a model for each \( z_{ikl} \) sequentially conditioning over the other \( z \) variables at measurement \( k \), all \( z \) variables at previous times, the corresponding vector of completely observed covariates, \( x_{ik} \), event time \( y_i \), and \( \alpha = (\alpha_{i1}, \ldots, \alpha_{K, q})' \) where each \( \alpha_{kl} \) is a vector of parameters indexing the distribution for each covariate \( l \) for measurement \( k \). The joint distribution of the \( z \) variables for subject \( i \) is then:

\[
p_z(z_i \mid y_i, x_i) = \prod p(z_{iK,q} \mid z_{iK,1}, \ldots, z_{iK,(q-1)}, z_{i(K-1),i}, \ldots, z_{i1,i}, x_{i1,y_i,\alpha_{K,q}}) \times \cdots \times p(z_{i1,1} \mid z_{i1,1}, x_{i1,1,y_i,\alpha_{1,1}}) \times p(r_{11} \mid x_{i1}, z_{i1}, y_i, \alpha_{11})
\]

Expression of the distribution of covariates this way allows considerable flexibility in choice of distribution for each \( z_{ikl} \), accommodating continuous and categorical variables, as well as offering a convenient way to account for intra-subject correlation without specification of a random effect. Once again, one should strive for a parsimonious specification of this joint distribution to avoid specification issues. Equation (4) then represents the \( i^{th} \) contribution to the marginal likelihood for \( z \).
Estimation

Substitution of equations (2), (3) and (4) into (1) yields the complete data likelihood:

\[
\ell(\beta_1, \beta_2, \lambda, \alpha, \phi) = \prod_{i=1}^{n} p(r_i, y_i, z_i | x_i, \beta_1, \beta_2, \lambda, \alpha)
\]

\[
= \prod_{i=1}^{n} p_r(r_i | y_i, z_i, x_i) p_y(y_i | z_i, x_i, \beta_1, \beta_2, \lambda) p_z(z_i | x_i, \alpha)
\]

with densities \( p_r(\cdot) \), \( p_y(\cdot) \) and \( p_z(\cdot) \) defined above. Previous work with similar models has made use of the EM algorithm to obtain the parameter estimates (13). However here we illustrate a Fully Bayesian (FB) approach using vague priors on the parameters \( \beta_1, \beta_2, \lambda, \alpha \) and \( \phi \) which will produce estimates equivalent to the frequentist analysis using EM and also yield variance estimates that are much easier to obtain than with the EM framework. The FB framework is also less computationally demanding than the EM framework for this model.

The joint posterior distribution of the parameters is proportional to the product of the conditional distribution of the observed data given the parameters and the joint prior distribution of the model parameters \( p(\beta_1, \beta_2, \lambda, \alpha, \phi) \):

\[
p(\beta_1, \beta_2, \lambda, \alpha, \phi | y, r, x, z) \propto \prod_{i=1}^{n} \int p(y_i, r_i, z_i | x_i, \beta_1, \beta_2, \alpha, \gamma, \phi, \lambda) dz_i \times p(\beta_1, \beta_2, \lambda, \alpha, \phi).
\]

(5)

If non-informative priors are specified for \( (\beta_1, \beta_2, \lambda, \alpha, \phi) \) then the posterior means and standard deviations of the parameters will be similar to maximum likelihood. We use the Gibbs Sampler (22) to sample from the posterior distribution given by equation (5).

Although somewhat computationally intensive (but less intensive than EM), the FB approach here provides a very straightforward way to estimate parameters from a complex model, especially variance and covariance parameters.
Example

We apply this model to an analysis of data from the follow-up study to the LIBCSP, to evaluate whether time-varying post-diagnosis changes in body size are related to survival among women with newly diagnosed breast cancer.

Description of the Long Island Breast Cancer Follow-up Study

The details of the LIBCSP are discussed elsewhere (14) but briefly, the parent study is a population based case-control study of breast cancer among women in Nassau and Suffolk counties on Long Island, New York conducted between August 1996 and July 1997. Cases consisted of 1,508 women with newly diagnosed in situ or invasive breast cancer; of these, 1,414 women agreed to be contacted at a later date for follow-up interviews. For those who agreed to participate in the follow-up, the case subjects or their proxy were contacted by mail approximately 5 years after initial diagnosis of breast cancer and informed consent was obtained via telephone follow-up calls. Of the 1,414 women who initially agreed to participate, 316 subsequently refused or were unable to be contacted. Of the remaining 1,098 subjects who agreed to the follow-up interview, only 1,033 case subjects or proxies (68.5% of the original 1,508 women) actually completed the interviewer-administered questionnaire (23). The follow-up interview ascertained information similar to that gathered in the baseline questionnaire but relevant to the time period since diagnosis including treatment, reproductive history, smoking and alcohol use, and as well as body size and physical activity. Date and cause of death were ascertained for all 1,508 women using the National Death Index (24) with median follow-up time of 8.8 years (range: 0.2-9.4 years).

Relevant to this analysis, the follow-up questionnaire ascertained body size (weight in pounds and height in inches) at diagnosis, one year post-diagnosis and at time of response to questionnaire for those subjects still living, or one year prior to death for interviews.
completed by proxy for subjects who were deceased at the follow-up but living longer than one year (specific timing for final follow-up measurement varied between 2 and 7 years post diagnosis). Overall refusal to participate in the follow-up interview and non-response to specific questions among people still alive at each timepoint resulted in percentages of missing data on body size of 47.6%, 49.4% and 33.9% at baseline, 1-year post diagnosis and final follow-up. Our concern is that heavier women may not have responded to the questionnaire in general, or to the body size questions specifically, due to self-conscious feelings or other reasons related to the amount of their weight, creating a non-ignorable mechanism for missing body size data. With the body size variables we calculate percent change in body weight between the year prior to diagnosis and $k^{th}$ measurement 

\[
(100\times(\text{weight at measurement } k - \text{weight one year before diagnosis})/\text{weight one year before diagnosis})
\]

for $k=1$ (at baseline), 2 (at one year) and 3 (at time of interview or one year prior to death).

Other fixed covariates included in this analysis (measured only once, at diagnosis) are indicators of chemotherapy regimen (yes/no), tumor size greater than 2 centimeters (yes/no), estrogen receptor positive tumor (ER status, yes/no) and progesterone receptor positive tumor (PR status, yes/no). Each of these covariates also exhibited a significant amount of missing data with 32.2%, 31.6%, 34.0% and 34.3% missing, respectively. The overlap between these variables with missing values was small, and since a complete case analysis requires all variables to be observed the resulting completely observed dataset, which excluded those with missing post-diagnosis change in bodyweight or a missing value for any covariate contained 499 subjects. Note, however, that the percentage missing for any one of these variables was moderate. Other important covariates included menopausal status, education,
adult weight change and body mass index (BMI) one year prior to diagnosis, which were each missing for less than 2% of subjects. We also include data on age at diagnosis, which is fully observed for all women. For the small amount of missing data on menopausal status, education, adult weight change and prediagnosis BMI we will exclude these subjects from the analysis, however for the remaining 1,455 subjects we will specify a selection model to account for the significant amount of missing data on follow-up body size, treatment and tumor characteristics. Out of the 1,455 women included in this analysis, 292 died during the follow-up period with 156 of those deaths attributed to breast cancer.

**Selection Model**

We model the time since diagnosis for subject \( i \) (denoted \( duri \)) as a piecewise exponential model with \( J=10 \) one-year intervals. Percent change in bodyweight for subject \( i \) in interval \( j \) corresponding to measurement \( k \), denoted as \( pcwt_{ij}^* = pcwt_{ik} \) for \( s_{j-1} \in (v_{ik}, v_{i(k+1)}] \), was categorized into four categories using indicator functions \( I_{(a,b)}(x) \) where \( I_{(a,b)}(x)=1 \) if \( x \in (a,b) \) and 0 otherwise. The four categories represent those who lost more than 5% of their pre-diagnosis body weight (\( pcwt_{ij}^* < -5 \)), those who maintained within 5% of their pre-diagnosis bodyweight (\( pcwt_{ij}^* \geq -5 \) and \( pcwt_{ij}^* \leq 5 \)), those who gained between 5% and 10% of their pre-diagnosis weight (\( pcwt_{ij}^* > 5 \) and \( pcwt_{ij}^* < 10 \)) and those who gained 10% or more of their prediagnosis bodyweight \( pcwt_{ij}^* \geq 10 \), omitting the category corresponding to those maintaining weight as the referent group. We also include fixed covariates continuous age at diagnosis (\( dxage_i \)), indicators for chemotherapy treatment (\( chemo_i \)), ER status (\( erstati \)), PR status (\( prstati \)) and tumor size > 2 centimeters (\( tumori \)), yielding the hazard function for our time to event model:
\[
\lambda(dur_i | x_i, z_i, \beta_1, \beta_2, \beta_3, \lambda_j) = \lambda_j \exp(\beta_{11} dxage_i + \beta_{12} I_{\infty < s} (pcwt_i^*) + \\
\beta_{21} I_{s < 5} (pcwt_i^*) + \beta_{23} I_{10, \infty} (pcwt_i^*) + \beta_{24} \text{chemoi} + \\
\beta_{25} \text{erstat}_i + \beta_{26} \text{prstat}_i + \beta_{27} \text{tumor}_i)
\]  

for \(dur_i \in (s_{j-1}, s_j]\). For our analysis we assume that only percent change in weight (pcwt_\(i_k\)) is potentially non-ignorably missing, while chemotherapy treatment (chemoi) and tumor characteristics tumor size > 2 centimeters (tumor_\(i\)), ER status (erstat_\(i\)) and PR status (prstat_\(i\)) are ignorably missing as we believe that their missingness is unlikely to be related to either unknown or known variables. Therefore only one missing data mechanism need be specified: \(r_{ik} = 1\) if subject \(i\) was missing body size responses at measurement \(k\) for \(k = 1, \ldots, K_i\) where \(K_i = 1, 2\) or \(3\) and \(r_{ik} = 0\) if the value was present. Then, from equation (2) for \(K_i = 3\) we have:

\[
\begin{align*}
\rho_i(r_i | y_i, x_i, z_i, \phi) &= p(r_{i3} | r_{i2}, r_{i1}, dxage_i, pcwt_{i3}, t_{i3}, \phi_3) \times \\
p(r_{i2} | r_{i1}, dxage_i, pcwt_{i2}, t_{i2}, \phi_2) \times \\
p(r_{i1} | dxage_i, pcwt_{i1}, t_{i1}, y_i, \phi_1)
\end{align*}
\]

where \(dxage_i\) is age at diagnosis and \(t_{ik}\) is time in years since diagnosis to measurement \(k\), both fully observed. The modification of equation (7) for \(K_i = 1\) or \(2\) is straightforward. We specify each of the conditional distributions on the right hand side of equation (6) with a logistic regression model.

Using equation (4) we express the joint distribution of the missing time-varying covariates percent change in weight (pcwt_\(i_k\)) for \(k = 1, \ldots, 3\), and fixed (baseline) covariates chemotherapy treatment (chemoi), tumor size > 2 centimeters (tumor_\(i\)), ER status (erstat_\(i\)), PR status (prstat_\(i\)), as functions of continuously measured year of follow-up (for the 3rd measurement, \(v_{i3}\)) menopausal status at diagnosis (menpstat_\(i\)), BMI in the year prior to
diagnosis \((bmiref_i)\), income reported at diagnosis \((income_i)\), years of education completed \((education_i)\) and age at diagnosis \((dxage_i)\) as:

\[
p_z^*(z_i | y_i, x_i, \phi) = p(pcwt_{i3} | pcwt_{i2}, \ldots, pcwt_{ik}, \ldots, pcwt_{i1}, \ldots, \text{chemo}_i, \ldots, dxage_i, \ldots, \text{menpstat}_i, \ldots, bmiref_i, \ldots, v_{i3}, \ldots, \alpha_7) \times \]

\[
p(pcwt_{i2} | pcwt_{i1}, \ldots, \text{chemo}_i, \ldots, dxage_i, \ldots, \text{menpstat}_i, \ldots, bmiref_i, \ldots, v_{i6}) \times \]

\[
p(pcwt_{i1} | \ldots, \text{chemo}_i, \ldots, dxage_i, \ldots, \text{menpstat}_i, \ldots, bmiref_i, \ldots, \alpha_5) \times \]

\[
p(\text{chemo}_i | dxage_i, income_i, education_i, \ldots) \times p(\text{erstat}_i | dxage_i, \ldots) \times \]

\[
p(prstat_i | dxage_i, \ldots) \times p(tumor_i | dxage_i, income_i, education_i, \ldots). \tag{8}\]

We model the conditional distributions of percent change in weight as linear regression models while the dichotomous treatment and tumor characteristic variables are modeled as using logistic regression models. Note that in the case of the linear regression models the parameter vector \(\alpha_i\) contains not only slope, but variance terms as well. The choice of covariates for the models for percent change in bodyweight were determined based on consensus of previous studies on postdiagnosis weight change among breast cancer patients \((25, 26)\). The models for treatment variables were selected in the interest of parsimony and to represent those variables we believe to be associated with access to care. An alternative option for the models for percent change in weight would be to generate a single four-level ordinal categorical variable for \(pcwt_{ik}\) with the corresponding indicator variables in the piecewise exponential model, and ordinal logistic regression models for the categorical variable in the joint distribution given by equation \((8)\). However, given the inherently continuous nature of the underlying variable the method presented here yields an equivalent and more intuitive specification.

We selected noninformative priors for the unknown parameters in the model. For the slope parameters for the regression models \((\beta_1, \beta_2, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \phi_1, \phi_2,\)
we specified independent normal distributions with zero mean and precision of \(10^{-5}\) and for the baseline hazards \((\lambda)\) and variance parameters for the linear regression models we specified vague gamma distributions. Estimation was performed using the Gibbs sampler in WinBUGS 1.4 (27) run for 25,000 iterations with an additional 25,000 burn-in. To evaluate the robustness of our model to assumptions on the missing data mechanism we also estimated this model assuming that the longitudinal body size variable was MAR, therefore omitting the specification for the missing data mechanism, equation (7). For comparison, we also estimated a complete-case analysis using maximum likelihood for an equivalent piecewise exponential model using Stata v. 10.0 (College Station, TX). Each set of models was estimated for both all-cause and breast-cancer specific mortality.

In table 3.1 we report the parameters from the piecewise exponential model under the various missing data assumptions for all-cause mortality. Post-diagnosis weight loss appears to be associated with poor survival across all missing data assumptions as this is likely indicative of women with more advanced disease or otherwise less than robust health. Post diagnosis weight gain is also positively associated with all-cause mortality for all model assumptions. However the associations in the complete case analysis in relation to moderate weight gain (5-10% of prediagnosis weight) are markedly attenuated compared to those that account for the missing data. The magnitudes of the associations with post diagnosis weight change are similar between the MAR and NMAR models, although the estimate of effect for more moderate gain (5-10%) appears attenuated in the NMAR model compared to the MAR model, with the confidence interval for both estimates including the value corresponding to a null effect. Most notable is that the credible interval for larger weight gain (>10%) excludes the null effect in both missing data models yet includes it in the complete case analysis. In
both missing data models, compared to women who maintain their prediagnosis weight, moderate weight gain is associated with a modest increase in risk of death (MAR posterior log-hazard ratio (lnHR): 0.31, 95% credible interval: -0.42, 0.99; NMAR posterior lnHR: 0.18, 95% credible interval: -0.55, 0.84) while larger gain is associated with a much greater risk of death (MAR posterior lnHR: 1.15, 95% credible interval: 0.50, 1.76; NMAR posterior lnHR: 0.72, 95% credible interval: 0.06, 1.36). Estimates of effect for the covariates appear nearly identical between the missing data models, and the same direction and similar magnitude compared to the complete case analysis.

Table 3.2 shows the results for breast cancer related deaths. Compared to the models for all-cause mortality, we observe similar patterns across the different missing data assumptions. The effect of moderate weight gain is attenuated, while the effect of larger weight gain is intensified in the models that account for missing data.

**Discussion**

We have presented a model for the analysis of time-to-event data with time-varying covariates when data on some covariates may be missing and have proposed an easy to implement solution strategy. We employed this model in an analysis of the association of longitudinal changes in bodyweight and survival after diagnosis with breast cancer in a large, population-based case-control study where we were concerned that data on post-diagnosis changes in body size may be non-ignorably missing. Our findings from the analysis suggest that weight gain after diagnosis is associated with greater mortality, both from any cause and specifically for death due to breast cancer with greater weight gain associated with a larger effect. Through sensitivity analysis we found that results from the complete case analysis are significantly attenuated compared to results from our selection model. In the case of breast cancer specific mortality, the results are contrary to what we expect, while the models that
account for the missing data produce estimates that appear more reasonable. Overall, models assuming an ignorable and non-ignorable missing data mechanism yielded equivalent conclusions although slight differences in parameter estimates were observed, especially for women with modest weight gain.

Models for missing data can be useful analytic tools yet there are no substitutes for complete data. The assumptions on the missing data mechanism are untestable and in some cases selection models similar to the one we illustrate have shown to be quite sensitive to misspecification (13). When employing selection models careful consideration must be given to the form of the model for the missing data mechanism—the desire for a thorough and accurate specification of the model must be balanced with parsimony as convergence can become problematic for models with many parameters. Although computationally intensive, the Bayesian approach to parameter estimation that we employed is easy to implement and accessible to analysts with a wide variety of computational ability. The ideal situation would be to completely observe data on all subjects, however this is unlikely to ever happen in reality, especially in longitudinal population-based studies. For cases where there is concern about the potential for covariate data to be missing at random, techniques such as the one we propose here offer a practical means of analyzing such incomplete datasets.
Table 3.1. Coefficient estimates (posterior log-hazard ratios) and 95% credible intervals from piecewise exponential proportional hazards model for all-cause mortality in the Long Island Breast Cancer Study Project under different missing data assumptions.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Complete Case* (n=499)</th>
<th>MAR† (n=1461)</th>
<th>NMAR§ (n=1461)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change in bodyweight</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;5% loss</td>
<td>1.39 (0.84, 1.94)</td>
<td>1.55 (1.14, 1.98)</td>
<td>1.30 (0.91, 1.72)</td>
</tr>
<tr>
<td>5-10% gain</td>
<td>0.04 (-0.89, 0.96)</td>
<td>0.31 (-0.42, 0.99)</td>
<td>0.18 (-0.55, 0.84)</td>
</tr>
<tr>
<td>&gt;10% gain</td>
<td>0.77 (-0.01, 1.55)</td>
<td>1.15 (0.50, 1.76)</td>
<td>0.72 (0.06, 1.36)</td>
</tr>
<tr>
<td>Chemotherapy treatment</td>
<td>0.51 (-0.02, 1.04)</td>
<td>0.68 (0.32, 1.03)</td>
<td>0.73 (0.36, 1.10)</td>
</tr>
<tr>
<td>ER positive tumor</td>
<td>-0.67 (-1.25, -0.08)</td>
<td>-0.50 (-0.84, -0.15)</td>
<td>-0.50 (-0.85, -0.15)</td>
</tr>
<tr>
<td>PR positive tumor</td>
<td>-0.27 (-0.85, 0.29)</td>
<td>-0.28 (-0.62, 0.05)</td>
<td>-0.28 (-0.60, 0.04)</td>
</tr>
<tr>
<td>Tumor size &gt; 2cm</td>
<td>0.71 (0.25, 1.18)</td>
<td>0.69 (0.40, 0.99)</td>
<td>0.68 (0.38, 0.97)</td>
</tr>
<tr>
<td>Age at diagnosis</td>
<td>0.05 (0.03, 0.08)</td>
<td>0.05 (0.03, 0.06)</td>
<td>0.04 (0.03, 0.06)</td>
</tr>
</tbody>
</table>

* Complete case analysis reports log-hazard ratios and 95% confidence intervals from maximum likelihood. Excludes subjects with missing data on one or more covariates.
† Specifies model for distribution of missing covariates.
§ Specifies model for distribution of missing covariates and missing data indicator for change in bodyweight.

Table 3.2. Coefficient estimates (log-hazard ratios) and 95% credible intervals from piecewise exponential proportional hazards model for breast cancer mortality in the Long Island Breast Cancer Study Project under different missing data assumptions.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Complete Case* (n=499)</th>
<th>MAR† (n=1461)</th>
<th>NMAR§ (n=1461)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change in bodyweight</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;5% loss</td>
<td>2.10 (1.24, 2.95)</td>
<td>1.83 (1.24, 2.47)</td>
<td>1.70 (1.10, 2.32)</td>
</tr>
<tr>
<td>5-10% gain</td>
<td>0.39 (-0.98, 1.76)</td>
<td>0.18 (-0.99, 1.21)</td>
<td>0.12 (-1.02, 1.14)</td>
</tr>
<tr>
<td>&gt;10% gain</td>
<td>0.79 (-0.40, 1.98)</td>
<td>1.07 (0.20, 1.90)</td>
<td>1.48 (0.87, 2.11)</td>
</tr>
<tr>
<td>Chemotherapy treatment</td>
<td>0.70 (-0.05, 1.44)</td>
<td>1.06 (0.54, 1.59)</td>
<td>1.12 (0.59, 1.65)</td>
</tr>
<tr>
<td>ER positive tumor</td>
<td>-0.23 (-1.08, 0.63)</td>
<td>-0.37 (-0.85, 0.11)</td>
<td>-0.37 (-0.84, 0.10)</td>
</tr>
<tr>
<td>PR positive tumor</td>
<td>-0.47 (-1.29, 0.35)</td>
<td>-0.40 (-0.85, 0.05)</td>
<td>-0.39 (-0.84, 0.06)</td>
</tr>
<tr>
<td>Tumor size &gt; 2cm</td>
<td>1.21 (0.59, 1.84)</td>
<td>1.12 (0.73, 1.49)</td>
<td>1.09 (0.70, 1.47)</td>
</tr>
<tr>
<td>Age at diagnosis</td>
<td>0.02 (-0.01, 0.06)</td>
<td>0.02 (0.00, 0.03)</td>
<td>0.02 (0.01, 0.04)</td>
</tr>
</tbody>
</table>

* Complete case analysis reports log-hazard ratios and 95% confidence intervals from maximum likelihood. Excludes subjects with missing data on one or more covariates.
† Specifies model for distribution of missing covariates.
§ Specifies model for distribution of missing covariates and missing data indicator for change in bodyweight.
References


CHAPTER 4
POST-DIAGNOSIS CHANGE IN BODYWEIGHT AND SURVIVAL AFTER BREAST CANCER DIAGNOSIS

Abstract

Weight gain after diagnosis is common among women with breast cancer, yet results have been inconsistent among the few studies examining its effects on survival. This study examines these effects among 1,436 women diagnosed with a first primary breast cancer in 1996-1997, on Long Island, NY. Shortly after diagnosis and again after approximately 5 years, subjects were interviewed to assess factors related to breast cancer, including anthropometric measures. Weight was assessed at each decade of life from age 20 years, one year before, at and one year after diagnosis and at the time of follow-up. Mortality through the end of 2005 was assessed using the National Death Index. Proportional hazards regression with time-varying covariates was used while employing a selection model to account for missing data. Compared to women who maintained their pre-diagnosis weight (+/- 5%), those who gained more than 10% after diagnosis had worse survival [hazard ratio (HR) (95% credible interval): 2.08 (1.05-3.94)]. The effect was more pronounced during the first 2 years after diagnosis [>5% gain, before 2 years, all-cause HR: 2.94 (0.49-17.43); after 2 years HR: 1.46 (0.83-2.49)]. For breast-cancer mortality, the association with post-diagnosis weight gain was stronger among women overweight before diagnosis [HR: 2.15 (0.71-6.17)] compared to ideal weight before diagnosis [HR: 1.18 (0.35-4.06)], and for women who had gained > 3 kilograms over adulthood before diagnosis [HR: 3.82 (2.04-
compared to no gain weight before diagnosis [HR: 1.13 (0.24-5.18)]. These results suggest the importance of weight maintenance for women after breast cancer diagnosis.

**Introduction**

For the more than 2 million female breast cancer survivors in the United States (1) the identification of modifiable factors that may positively influence the survivorship experience is imperative. Of particular interest is the effect of changes in body size after diagnosis on survival, as many women gain weight after being diagnosed with breast cancer (2, 3). Weight gain among women with breast cancer is often observed in response to treatment with chemotherapy, as well as among women with younger age at diagnosis, lower pre-diagnosis BMI, those who are premenopausal, and present with later disease stage at diagnosis (2-7). Higher levels of adipose tissue are associated with greater circulating levels of hormones such as estrogen, insulin and related growth factors, which increase proliferation of mammary cells and are associated with breast carcinogenesis (8-10).

The prognostic significance of obesity at or before diagnosis (3) is well established, however, the question of how changes in bodyweight after diagnosis affect survival has been addressed by relatively few studies with inconsistent results (3, 11-13). Since pre-diagnosis body size is often associated with post-diagnosis changes (2-7), it is unclear if the observed increase in risk associated with pre-diagnosis body size is merely a reflection of any adverse effect of post-diagnosis changes in body composition, and thus the observed reduction in survival for pre-diagnosis body size is only an artifact of its correlation with post-diagnosis changes. Alternatively, it is possible that both pre-and post-diagnostic weight changes adversely influence survival independently of each other, or even synergistically. Furthermore, none of the previous studies have employed longitudinal measures of post-diagnosis weight change, thus have been unable to examine if the effect may vary over time.
The objective for this analysis was to examine the effect of post-diagnosis body size changes over time on survival after breast cancer diagnosis using data from a large, population-based cohort of women with a recent diagnosis of breast cancer on Long Island, New York, and determine if the adverse effects of pre-diagnosis body size are attenuated after accounting for post-diagnosis weight change. Additionally, possible modification of the effect of post-diagnosis weight change was examined by stratification on baseline body size, pre-diagnosis adult weight gain and timing of post-diagnosis weight change.

**Methods**

For this analysis we used data from the follow-up to the Long Island Breast Cancer Study Project (LIBCSP). The parent LIBCSP is a population-based case-control study of breast cancer conducted on Long Island, New York (14) with a follow-up study of case women to determine factors associated with survival after breast cancer diagnosis. The study was approved by the Institutional Review Board of participating institutions.

**Study Population.** Cases were English speaking adult women with a first primary in situ or invasive breast cancer diagnosed between August 1, 1996 to July 31, 1997, from Nassau and Suffolk counties in New York. Potentially eligible women were identified through pathology departments of participating hospitals and their physicians were contacted to confirm the diagnosis and obtain permission to contact the patients for participation in the study. A total of 1,508 eligible cases (82.1%) agreed to participate in the parent study and completed the baseline questionnaires. Signed informed consent was obtained for all subjects prior to data collection.

During the parent case-control study 94 women refused to be contacted at a later date for participation in the follow-up study. For those who agreed to participate in the follow-up, the case subjects or their proxy were contacted by mail approximately 5 years after initial
diagnosis of breast cancer and informed consent was obtained via telephone follow-up calls. Of the 1,414 women who initially agreed to participate in the follow-up study, 60 subsequently refused at the initial mail contact, 65 refused when contacted by telephone, 18 refused due to illness, 22 were unable to complete the interview, 55 were lost to follow-up and 96 were deceased with no identifiable proxy to complete the interview. Of the remaining 1,098 subjects who agreed to the follow-up interview, 65 were only able to provide information on their first course of treatment for their original breast cancer diagnosis. Ultimately 1,033 case subjects (68.5% of the original 1,508 women) actually completed the follow-up interview (15). The follow-up interview ascertained information similar to that gathered in the baseline questionnaire but relevant to the time period since diagnosis.

**Outcome Assessment.** The outcome for this study, date and cause of death, was determined by the National Death Index (NDI) (16), a centralized database of death records maintained by the National Center for Health Statistics which is considered the standard source of mortality data for epidemiologic research (17). For women who were determined to be deceased from the NDI records we constructed two indicators: (1) breast cancer related death (breast cancer listed as primary or secondary cause of death, International Classification of Disease code 174.9 or C-50.9) and (2) death from any cause. These indicators, combined with time of death, form the outcomes for this analysis. Of the 1,508 cases from the parent study, there were a total of 308 deaths as of December 31, 2005, and the largest single cause of death was attributed specifically to breast cancer (n = 164).

**Body Size Assessment.** At baseline the initial questionnaire (18) collected self-reported data on height in inches and body weight in pounds at age 20 years and at 1 year prior to date of diagnosis. The baseline questionnaire also collected weight in pounds by
decade of life from 20 years through 70 years of age. The follow-up questionnaire assessed self-reported height, body weight in pounds at diagnosis, one year after diagnosis and at time of follow-up interview or one year before death for interviews on deceased subjects completed by proxy. This yielded three assessments of body size at and after diagnosis.

Percent change in body weight was calculated between the year prior to diagnosis and at diagnosis, 1-year post diagnosis and at time of follow-up interview as 100*(weight at follow-up measurement - weight one year before diagnosis)/weight one year before diagnosis. Weight change was categorized as >5% loss, maintain within 5%, 5-10% gain and greater than 10% weight gain; these categories were used to compare our results to other reports (11) and correspond to categories of weight loss used in weight management recommendations aimed at cancer patients (19). To avoid small counts within strata when assessing effect modification, weight change was categorized as >5% loss, maintain within 5% and >5% gain. Other body size variables used in our analysis included BMI 1 year before diagnosis (<25 kilograms/meters\(^2\) (kg/m\(^2\)), 25-30 kg/m\(^2\) and >=30 kg/m\(^2\)) and adult weight change, which reflected change in body weight from age 20 up to 1 year before diagnosis (>3 kg loss, maintenance within 3 kg and >3 kg gain). These categorizations of pre-diagnosis body size reflect those used in previous analyses of this data (20).

**Covariates.** Risk factor information was assessed by structured questionnaire, which was administered by trained interviewers during in-home visits [at baseline, (http://epi.grants.cancer.gov/LIBCSP/projects/Questionnaire.html)] or telephone (at follow-up). Information collected included known and suspected risk and prognostic factors for breast cancer, including menopausal status, education, income, anthropometric measures, and physical activity.
For case women who signed a medical record release form (97.7%), tumor stage and estrogen and progesterone receptor (ER/PR) status were ascertained from the medical records as part of the parent case-control study. For the follow-up, signed medical record release forms were obtained again, and medical records were abstracted for 598 women to obtain the details regarding complete course of treatment for the first primary breast cancer diagnosis. The abstracted treatment data were then compared to the self-reported information obtained as part of the telephone follow-up interview. Kappa coefficients comparing self-report and medical records were high for all three treatment modalities examined: radiation therapy $\kappa=0.97$, chemotherapy $\kappa=0.96$ and hormone therapy $\kappa=0.92$ (20), and thus the self-reported data are included in these analyses. Data on tumor size was obtained from the New York State Cancer Registry.

At baseline, approximately 98% of the respondents completed a self-administered modified Block food frequency questionnaire (FFQ) that took an average of approximately 30 minutes to complete (14). The FFQ assessed frequency and relative portion size for 101 food items in the year before the baseline interview.

**Statistical Analysis.** Non-response to specific questions among people still alive at each timepoint, coupled with the overall non-response to the follow-up interview, resulted in percentages of missing data on weight of 47.6%, 49.4% and 33.9% at time of diagnosis, 1-year post diagnosis and final follow-up, respectively among all 1,508 women. Given such substantial percentages of missing data, and the sensitivity of assessing body size, there was concern that body size data may be not missing at random (NMAR), a condition which arises when the probability that data is missing is dependent upon the unobserved values (21, 22). In this study, the potential issue was that nonresponse to the follow-up questionnaire may be
more likely among heavier women, who may be more psychologically sensitive to questions regarding bodyweight or who may be in poor health caused by being overweight. To address this issue a selection model for the analysis of survival data with non-ignorably missing time-varying covariates was used (23). A proportional hazards regression was specified to estimate the effect of post-diagnosis weight change over time on survival (all-cause and breast cancer specific). The main variables in this model included post-diagnosis change in body size (>5% loss, maintain within 5% (referent category), 5-10% gain, >10% gain), BMI one year before diagnosis (<25 kg/m² (referent category), >= 30 kg/m²) and adult weight change from age 20 (> 3 kg loss, maintain within 3 kg (referent category), > 3 kg gain), and were adjusted for age (continuous).

For models of effect modification post-diagnosis weight change was categorized as >5% loss, maintain within 5% and >5% gain and included its product with time (< 2 years, >= 2 years), BMI one year before diagnosis (categorized as <25 kg/m² and >= 25 kg/m²) and adult weight gain from age 20 years to 1 year before diagnosis (any loss or maintain within 3 kg, gain >= 3 kg). The time interaction of two years was chosen since recent studies of post-diagnosis weight change assessed women at approximately this time (11, 12) and to distinguish the possible effects of weight gain shortly after diagnosis. Collapsing the categories for the measures of pre- and post-diagnosis body size was to prevent unestimable effects due to small counts within categories.

Potential confounders in this analysis included those variables potentially associated with both post-diagnosis weight change and survival: post-diagnosis physical activity (metabolic equivalent task hours per week, MET-hrs/wk; continuous), total energy at baseline interview (kcal/day; continuous), menopausal status (premenopausal,
postmenopausal), stage (in situ, invasive), chemotherapy regimen (yes, no), radiation therapy (yes, no), hormone therapy (yes, no), nodal involvement (yes, no), tumor size greater than 2 cm (yes, no), estrogen receptor positive tumor (ER status; yes, no) and progesterone receptor positive tumor (PR status; yes, no). The final models included those confounders that altered the estimates of the hazard ratios for change in bodysize by >10% in a complete case analysis using Stata (version 10.0, College Station, TX) as selection models are impractical for the usual algorithm used for confounder identification. The final confounder list included age, chemotherapy regimen, tumor size, ER status and PR status, which was also consistent with identification of confounders using the analysis of a directed acyclic graph.

In addition to the proportional hazards model, the selection model requires specification of a model for the missing covariates and the dichotomous indicator of the missing data. The change in body size variable at each follow-up measurement was modeled as a linear regression dependent upon previous change in body size variables, as well as age, chemotherapy indicator, menopausal status and BMI before diagnosis, variables established in the literature to be consistently associated with post-diagnosis weight change (2-7). In addition to weight change, other covariates with significant amounts of missing data included chemotherapy, tumor size, ER status and PR status with 32.2%, 31.6%, 34.0% and 34.3% missing out of the 1,508 cases, respectively. Chemotherapy and tumor size were modeled as logistic regressions against age, income (< $20,000; $20,000-49,999; $50,000-89,999 and >= $90,000) and education (high school or less, some college, college graduate and post-college), while hormone receptor status indicators were modeled as logistic regressions as a function of age. Treatment and tumor characteristics were unlikely to be non-ignorably missing and therefore did not require specification of models for their missing data.
mechanisms. The probability that data was missing for change in body size at any given observation is modeled as a logistic regression with age, weight change at the current and all previous observations, and missingness indicator at all previous observations. The parameters of these ancillary models are not of inferential interest and are only needed to provide unbiased estimates of the survival model. Variables with minor amounts of missing data included menopausal status (1.99% missing), pre-diagnosis BMI (1.13% missing), adult weight change (from age 20 to 1 year before diagnosis; 1.66% missing), education (0.40% missing) and income (0.27% missing); so, we excluded these few subjects from the analysis as these amounts were unlikely to influence our results.

This analysis ultimately included 1,436 women, 292 who died during our follow-up with 156 of these deaths attributed to breast cancer. Median survival time was 8.80 years after diagnosis with a range from 0.23 to 9.41 years. Most women were postmenopausal at time of diagnosis. Women in this cohort were between 25.1 to 98.1 years of age with average of 58.8 years. Less than half of the women received chemotherapy treatment, and most tumors were either ER-positive or PR-positive. Less than 20% of the tumors were 2 cm or larger. Sociodemographic and anthropometric characteristics of the study sample are shown in table 1.

A fully Bayesian approach to parameter estimation was employed by specifying vague prior distributions on the model parameters and the Gibbs sampler in WinBUGS 1.4 (24) was used to sample from their posterior distribution. The sampler was run for 60,000 iterations with the first 20,000 discarded as a burn-in sample, retaining every 5th iteration to reduce serial correlation. Posterior hazard ratios (HR) were calculated by exponentiating the mean of the samples for the beta coefficients (log-hazard ratios) from the proportional
hazards model and posterior credible intervals were calculated by exponentiating the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles of these samples. The use of vague priors yields posterior point estimates and credible intervals that are similar to those from standard frequentist analysis.

**Results**

As shown in table 1, among women with complete data on at least one follow-up measure of body size, some 55% were able to maintain their pre-diagnosis body size. However, among the nearly 22% of women who reported gaining weight after diagnosis, mortality was increased, as shown in table 2. The effect estimates for mortality associated with moderate post-diagnosis weight gain [all-cause posterior HR: 1.19 (0.58-2.31) and breast cancer-specific posterior HR: 1.13 (0.36-3.11)] were increased even after adjustment for pre-diagnostic BMI and adult weight gain, and similar to those from the unadjusted models that were reported previously (23). With these additional adjustments, mortality risk was doubled for large post-diagnosis weight gain [all-cause posterior HR: 2.08 (1.05-3.94), and breast cancer-specific posterior HR: 2.14 (0.86-5.06)], although the estimate for breast-cancer specific mortality was attenuated compared to the unadjusted models (see table 2). Among women who lost weight after diagnosis, the adverse effect on survival remained, with increased effect estimates for all-cause [HR: 3.56 (2.39-5.48)] and breast-cancer specific mortality [HR: 4.29 (2.42-7.82)], even after adjusting for weight loss prior to diagnosis.

In these models that consider post-diagnosis weight change, being overweight (BMI 25.0-29.9 kg/m<sup>2</sup>) at time of diagnosis was not strongly associated with an increased risk of either all-cause mortality [HR: 0.89 (0.65-1.20)] or breast cancer-specific deaths [HR: 1.05 (0.69-1.60)]. Although the adverse effects associated with obesity at diagnosis (BMI >= 30 kg/m<sup>2</sup>) are still evident [all-cause mortality HR: 1.25 (0.91-1.71); breast cancer-specific mortality HR: 1.35 (0.87-2.08)], the magnitude of the effect is attenuated from the
pronounced effects previously reported for this study population (20). Similarly, once the
effects of post-diagnosis weight change are considered, the adverse effects associated with
pre-diagnostic adult weight loss conveyed a moderate effect on all-cause [HR: 1.63 (0.90-
2.98)], but not breast cancer-specific mortality [HR: 1.04 (0.36-2.76)], and adult weight gain
prior to diagnosis showed little effect on either outcome.

Effect modifications due to timing of post-diagnosis weight gain, pre-diagnosis BMI
and pre-diagnosis adult weight change are presented in table 3. Across all of these models
the greatest increases in risk are seen among women who lost weight after diagnosis, which
again may reflect poorer health in general. The deleterious effects of post-diagnosis weight
gain appear stronger when weight is gained shortly after diagnosis [within first 2 years of
diagnosis, all-cause posterior HR: 3.02 (0.51-17.44) and breast cancer-specific HR: 2.31
(0.40-12.78)], although the wide credible intervals indicate imprecise estimates. Weight gain
after two years shows similar effects for both all-cause and breast cancer-specific mortality,
with moderate increases in risk of death compared to women who maintained their weight.
The effect of post-diagnosis weight gain is slightly greater among overweight and obese
women (BMI >= 25 kg/m²) than among women of ideal weight (BMI < 25 kg/m²), with the
difference being greatest for breast cancer-related deaths [BMI < 25 kg/m², posterior HR: 1.18
(0.35-4.06) and BMI >= 25 kg/m², posterior HR: 2.15 (0.71-6.17)]. Among women
who maintained or lost weight before diagnosis, the effect of post-diagnosis weight gain was
slightly lower compared to women who gained weight as an adult, however once again this
difference was greatest when only deaths due to breast cancer were considered [pre-diagnosis
weight maintainers, posterior HR: 1.13 (0.24-5.18) and pre-diagnosis weight gainers,
posterior HR: 3.82 (2.04-7.37)].
Discussion

Previously reported findings (23) on weight change after diagnosis in this cohort of 1,436 women diagnosed with in situ or invasive breast cancer focused on the methodologic approach for considering time varying non-ignorably missing covariate data in a survival analysis. That previous analysis (23) did not include adjustments for pre-diagnosis levels of BMI or adult weight change; nor did it consider effects associated with the timing of the weight gain relative to diagnosis, namely those that occurred soon after diagnosis, presumably during chemotherapy, versus weight gain that occurred later, after the first course of treatment is completed. The unadjusted results of the previous analysis suggested a moderate 19% increase in all-cause mortality among women gaining between 5-10% of their pre-diagnosis bodyweight, with a much larger 105% increase among women who gained more then 10%; similar increases in risk were noted for breast cancer-specific mortality. The greatest risk – over three-fold—was found among women who lost weight after diagnosis, yet this is likely to reflect adverse effects associated with advanced disease or otherwise poor health.

In the analysis presented here, even after adjusting for pre-diagnosis body size variables in this analysis, moderate increases in risk of all-cause and breast cancer-specific mortality were still observed among women who gained between 5-10% of their pre-diagnosis weight at any time after diagnosis. Large increases in mortality risk were found for both outcomes among women who gained more than 10% of their pre-diagnosis weight after diagnosis. When pre-diagnosis anthropometric measures of adiposity were included in the models the effects for larger weight gain were somewhat attenuated from the analysis that did not include pre-diagnosis body size, primarily for breast-cancer specific mortality, although the effects were essentially unchanged for moderate weight gain. The effect of weight gain
on mortality appeared stronger within the first 2 years after diagnosis compared to after 2 years, although study power was limited to draw firm conclusions regarding this difference. The effect of post-diagnosis weight gain on breast cancer specific mortality was also stronger among women who had gained weight as an adult before diagnosis.

Fat mass is thought to affect breast cancer etiology since visceral adipose tissue is metabolically active (25) and affects a number of pathways that are involved in carcinogenesis. Visceral adipose tissue is associated with increases in sex hormones such as estradiol and decreases in related binding proteins, especially sex hormone binding globulin (SHBG) (26) as well as insulin and insulin like growth factors (10, 27) which can promote a hormonal environment that encourages proliferation of both normal and cancerous mammary cells (8-10). Weight gain after adolescence is primarily associated with accumulation of this metabolically active adipose tissue and therefore changes in body size may be an etiologically relevant exposure (28). This issue is of particular concern as weight gain after diagnosis in breast cancer patients is well documented, with the amount ranging from approximately 1 kg to more than 10 kg within the first two years (3).

The relationship between breast cancer survival and post-diagnosis changes in body size is far less studied than the relationship with pre-diagnostic body size. Only a handful of studies have directly examined the relationship between weight change post-diagnosis and survival with inconsistent results (11-13, 29-33). Findings have generally been null when modest increases in weight have been considered (31-33), although significant decreases in survival have been noted when larger changes in weight or BMI were studied, often in subgroup analyses (12, 29, 30). Caan and colleagues failed to find an association between post-diagnosis weight change and survival, however the authors did note an association with
pre-diagnosis body size (11). Nichols et al. found results more similar to those reported here, namely that weight gain of at least 5 kg was associated with an increase in all-cause, breast cancer specific and cardiovascular mortality. Generally, studies with null findings have been limited by small sample size (31-33) and short follow-up (31, 33), possibly explaining the lack of associations. Additionally, timing of follow-up measures is likely to contribute to the differences in findings. Our study was able to utilize anthropometric measures from diagnosis to death or censoring, allowing us to estimate the effect of this time-varying exposure as well as estimate differences in effect across relevant time periods. We observed that the deleterious effect of post-diagnosis weight gain was more pronounced closer to diagnosis, which was not assessed in other recent studies that used body size measures at least 2 years post diagnosis (11, 12).

Our findings regarding weight loss are consistent with recent reports by Caan et al. (11) and Nichols et al. (13) who observed that weight loss was associated with significantly greater mortality after breast cancer diagnosis, both reporting hazard ratios over 2 in the greatest category of weight loss for death from any cause. Data from the Nurses’ Health Study also showed a greater risk of death among those who lost weight, although the associations were not statistically significant (12). Recent recommendations regarding weight loss for breast cancer patients (34) have been made based on the observation that greater BMI at diagnosis is associated with poor survival, however post-diagnosis weight loss is not supported, and appears to be contradicted, by recent observational reports, including this one. As stated previously, it is unclear if this observed association between weight loss and survival is due to a distinct effect of the weight loss, or the fact that those
who are near death are likely to be losing weight. This issue could be clarified by studies that include data on intent of weight loss among breast cancer survivors.

Obesity at diagnosis and weight gain pre-diagnosis are established indicators of poorer prognosis as discussed in several recent reviews (3, 34-36), although most of the studies conducted to date have not accounted for post-diagnosis changes in body size. Similarly, a recent report that utilized the Long Island breast cancer cohort observed hazard ratios ranging from 1.63 to 2.85 for women who were obese before diagnosis (20). However, in the analysis reported here, once the effects of post-diagnosis changes in weight on mortality are considered, the associations with pre-diagnosis anthropometric measures are attenuated. Our new findings suggest that the effect of pre-diagnosis body size may be driven, at least partly, by its association with post diagnosis weight change. Additionally, our observation of a more pronounced adverse effect for post-diagnosis weight gain among women who also gained weight as an adult before diagnosis underscores the importance of avoiding increases in adiposity at any point in a woman’s life trajectory.

Strengths of this study include its population-based study design and relatively large sample size. Also, an innovative analytical approach to the treatment of missing exposure and confounder data was employed. Complete case analysis is common since it is automatically carried out by most software packages, however it usually reduces statistical efficiency, and can yield biased effect estimates in all but the most rigid conditions. Ad hoc adjustment for missing data is still common in epidemiology, such as some variant of the “missing indicator” method or improper imputation, which can perform even worse than complete case analysis (22). Formal treatment of missing data is crucial to accurate inference, and the selection model approach employed here can account for potentially non-
ignorably missing covariates, which is when bias and statistical efficiency are of greatest concern (21). Even though the methodology used here is theoretically sound, it is important to keep in mind that even the most rigorous statistical model is no substitute for having the data that was unobserved. Missing data models rely on untestable assumptions and can be quite sensitive to changes in specification (37).

A significant strength of this analysis is that data from multiple assessments of change in body size over the follow-up experience were utilized, starting at and near diagnosis, which allowed for determination of differential effects based on timing of weight change which, as of yet, has gone unaddressed. A limitation of this analysis is the use of self-reported body size measures, which leaves open the potential for measurement error. However, self-reported anthropometric measures have been shown to be highly correlated when compared to measurements taken in a clinic setting (38). Data from the NHANES III study showed that in addition to self reported and measured weight being highly correlated, older women, who make up most of this cohort, tend to report their weight accurately (39). Also, self- and interviewer-obtained measurements have shown nearly identical associations in a recent analysis of pre-diagnosis BMI and survival after breast cancer diagnosis in a similar cohort (40). The use of proxy interviews could also be a source of bias, however these accounted for a small portion of our study sample (<8%), and a recent detailed report comparing the use of proxy and case assessments illustrated that proxy assessments of anthropometric measures yielded nearly identical associations compared to those completed by the case subject (41). The results from our stratified models should be interpreted as exploratory given that statistical power was reduced for subgroup analysis. Additionally, the categorization of pre-diagnosis BMI and pre-diagnosis weight gain yielded a different
percentage of cases in each subgroup, notably 80% of the cohort appeared in the upper category of pre-diagnosis weight gain, which likely contributed to the differences noted, as BMI and weight gain are highly correlated.

In summary, these findings suggest that weight maintenance after breast cancer diagnosis should be encouraged, especially among women who have gained weight as an adult before diagnosis. The time period immediately after diagnosis may be especially relevant for weight maintenance, when treatment-related weight gain is common. Although the associations with pre-diagnosis BMI were attenuated in our study, maintenance of a healthy weight should nevertheless be encouraged at all times as it may promote weight maintenance after diagnosis, and conveys other health benefits.

<table>
<thead>
<tr>
<th>Variable</th>
<th>N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deaths through December 31, 2005</td>
<td></td>
</tr>
<tr>
<td>All cause</td>
<td>292 (20.3)</td>
</tr>
<tr>
<td>Breast-cancer specific</td>
<td>156 (10.8)</td>
</tr>
<tr>
<td>Weight change from 1-year before diagnosis to date of diagnosis</td>
<td></td>
</tr>
<tr>
<td>&gt;5% loss</td>
<td>197 (26.0)</td>
</tr>
<tr>
<td>Maintain ±5%</td>
<td>458 (60.4)</td>
</tr>
<tr>
<td>5-10% gain</td>
<td>72 (9.5)</td>
</tr>
<tr>
<td>&gt;10% gain</td>
<td>31 (4.1)</td>
</tr>
<tr>
<td>Missing</td>
<td>678</td>
</tr>
<tr>
<td>Weight change from 1-year before diagnosis to 1-year post diagnosis</td>
<td></td>
</tr>
<tr>
<td>&gt;5% loss</td>
<td>168 (23.1)</td>
</tr>
<tr>
<td>Maintain ±5%</td>
<td>400 (55.0)</td>
</tr>
<tr>
<td>5-10% gain</td>
<td>95 (13.1)</td>
</tr>
<tr>
<td>&gt;10% gain</td>
<td>64 (8.8)</td>
</tr>
<tr>
<td>Missing</td>
<td>709</td>
</tr>
<tr>
<td>Weight change from 1-year before diagnosis to time of follow-up interview</td>
<td></td>
</tr>
<tr>
<td>&gt;5% loss</td>
<td>200 (21.2)</td>
</tr>
<tr>
<td>Maintain ±5%</td>
<td>414 (43.9)</td>
</tr>
<tr>
<td>5-10% gain</td>
<td>161 (17.1)</td>
</tr>
<tr>
<td>&gt;10% gain</td>
<td>168 (17.8)</td>
</tr>
<tr>
<td>Missing</td>
<td>493</td>
</tr>
<tr>
<td>ER status</td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>701 (73.6)</td>
</tr>
<tr>
<td>Negative</td>
<td>252 (26.4)</td>
</tr>
<tr>
<td>Missing</td>
<td>483</td>
</tr>
<tr>
<td>PR status</td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>609 (64.2)</td>
</tr>
<tr>
<td>Negative</td>
<td>340 (35.8)</td>
</tr>
<tr>
<td>Missing</td>
<td>487</td>
</tr>
<tr>
<td>Chemotherapy</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>404 (41.4)</td>
</tr>
<tr>
<td>No</td>
<td>573 (58.7)</td>
</tr>
<tr>
<td>Missing</td>
<td>459</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumor size</td>
<td></td>
</tr>
<tr>
<td>&lt; 2 cm</td>
<td>749 (76.2)</td>
</tr>
<tr>
<td>&gt;= 2 cm</td>
<td>234 (23.8)</td>
</tr>
<tr>
<td>Missing</td>
<td>453</td>
</tr>
<tr>
<td>BMI 1 year before diagnosis</td>
<td></td>
</tr>
<tr>
<td>&lt;18.5 kg/m²</td>
<td>25 (1.7)</td>
</tr>
<tr>
<td>18.5-24.9 kg/m²</td>
<td>639 (44.5)</td>
</tr>
<tr>
<td>25.0-29.9 kg/m²</td>
<td>455 (31.7)</td>
</tr>
<tr>
<td>&gt;=30.0 kg/m²</td>
<td>317 (22.1)</td>
</tr>
<tr>
<td>Weight change from age 20 to 1 year before diagnosis</td>
<td></td>
</tr>
<tr>
<td>&lt; 3kg gain or any loss</td>
<td>240 (16.7)</td>
</tr>
<tr>
<td>&gt; 3kg gain</td>
<td>1,196 (83.3)</td>
</tr>
<tr>
<td>Age at diagnosis (mean: 58.8 yrs, sd: 12.6 yrs)</td>
<td></td>
</tr>
<tr>
<td>20-29.9</td>
<td>10 (0.7)</td>
</tr>
<tr>
<td>30-39.9</td>
<td>77 (5.4)</td>
</tr>
<tr>
<td>40-49.9</td>
<td>306 (21.3)</td>
</tr>
<tr>
<td>50-59.9</td>
<td>381 (26.5)</td>
</tr>
<tr>
<td>60-69.9</td>
<td>340 (23.7)</td>
</tr>
<tr>
<td>70-79.9</td>
<td>274 (19.1)</td>
</tr>
<tr>
<td>80-89.9</td>
<td>45 (3.1)</td>
</tr>
<tr>
<td>90+</td>
<td>3 (0.2)</td>
</tr>
<tr>
<td>Menopausal Status</td>
<td></td>
</tr>
<tr>
<td>Premenopausal</td>
<td>462 (32.2)</td>
</tr>
<tr>
<td>Postmenopausal</td>
<td>974 (67.8)</td>
</tr>
<tr>
<td>Education</td>
<td></td>
</tr>
<tr>
<td>High school or less</td>
<td>688 (47.9)</td>
</tr>
<tr>
<td>Some college</td>
<td>342 (23.8)</td>
</tr>
<tr>
<td>College graduate</td>
<td>186 (13.0)</td>
</tr>
<tr>
<td>Post college education</td>
<td>220 (15.3)</td>
</tr>
<tr>
<td>Income</td>
<td></td>
</tr>
<tr>
<td>&lt; $20,000</td>
<td>174 (12.1)</td>
</tr>
<tr>
<td>$20,000-49,999</td>
<td>560 (39.0)</td>
</tr>
<tr>
<td>$50,000-89,999</td>
<td>427 (29.7)</td>
</tr>
<tr>
<td>&gt;=$90,000</td>
<td>275 (19.2)</td>
</tr>
</tbody>
</table>
Table 4.2. Hazard ratios (and 95% credible intervals) for the association between post-diagnosis changes in body weight and all-cause and breast cancer-specific mortality among women newly diagnosed with a first primary breast cancer in 1996-1997 in Long Island, NY and followed through 2005.

<table>
<thead>
<tr>
<th>Hazard ratio (95% credible interval)</th>
<th>All-cause Mortality (292 deaths/1,436 subjects)</th>
<th>Breast Cancer-specific Mortality (156 deaths/1,436 subjects)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1*</td>
<td>Model 2†</td>
</tr>
<tr>
<td>Post-diagnosis weight change</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;5% loss</td>
<td>3.68 (2.47-5.57)</td>
<td>3.56 (2.39-5.48)</td>
</tr>
<tr>
<td>±5% maintain (ref)</td>
<td>1.</td>
<td>1.</td>
</tr>
<tr>
<td>5-10% gain</td>
<td>1.19 (0.58-2.31)</td>
<td>1.23 (0.59-2.39)</td>
</tr>
<tr>
<td>&gt;10% gain</td>
<td>2.05 (1.06-3.90)</td>
<td>2.08 (1.05-3.94)</td>
</tr>
<tr>
<td>BMI 1 year before diagnosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;25 kg/m² (ref)</td>
<td></td>
<td>1.</td>
</tr>
<tr>
<td>25-30 kg/m²</td>
<td>0.89 (0.65-1.20)</td>
<td></td>
</tr>
<tr>
<td>≥30 kg/m²</td>
<td>1.25 (0.91-1.71)</td>
<td></td>
</tr>
<tr>
<td>Weight change from age 20 to 1 year before diagnosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;3kg loss</td>
<td>1.63 (0.90-2.98)</td>
<td></td>
</tr>
<tr>
<td>±3kg maintain (ref)</td>
<td>1.</td>
<td></td>
</tr>
<tr>
<td>&gt;3kg gain</td>
<td>0.93 (0.63-1.41)</td>
<td></td>
</tr>
</tbody>
</table>

* Model 1 adjusted for age at diagnosis, chemotherapy treatment, ER status, PR status and tumor size. These results presented previously (23).
† Model 2 includes covariates from Model 1 as well as BMI 1 year prior to breast cancer diagnosis and weight change from age 20 up to 1 year before breast cancer diagnosis.
Table 4.3. Hazard ratios (and 95% credible intervals) for the association between post-diagnosis changes in body weight and all-cause and breast cancer-specific mortality, stratified by time, pre-diagnosis BMI and pre-diagnosis adult weight change among women newly diagnosed with a first primary breast cancer in 1996-1997 on Long Island, NY, and followed through 2005.

<table>
<thead>
<tr>
<th>Post-diagnosis weight change</th>
<th>Time Since Diagnosis</th>
<th>Pre-diagnosis BMI</th>
<th>Pre-diagnosis adult weight gain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>All-cause Mortality (292 deaths/1,436 subjects)</td>
<td>Breast Cancer-specific Mortality (156 deaths/1,436 subjects)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Before 2 years</td>
<td>After 2 years</td>
</tr>
<tr>
<td>&gt;5% loss</td>
<td></td>
<td>7.87 (2.47-37.41)</td>
<td>3.39 (2.25-5.21)</td>
</tr>
<tr>
<td>±5% maintain (ref)</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>&gt;5% gain</td>
<td></td>
<td>2.94 (0.49-17.43)</td>
<td>1.46 (0.83-2.49)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Post-diagnosis weight change</th>
<th>&lt; 25</th>
<th>&gt;= 25</th>
<th>&lt; 3 kg</th>
<th>&gt;= 3kg</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;5% loss</td>
<td></td>
<td>4.50 (2.52-8.22)</td>
<td>3.27 (2.02-5.73)</td>
<td>4.42 (1.94-10.59)</td>
</tr>
<tr>
<td>±5% maintain (ref)</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>&gt;5% gain</td>
<td></td>
<td>1.47 (0.70-3.13)</td>
<td>1.69 (0.80-3.61)</td>
<td>1.18 (0.35-4.06)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Post-diagnosis weight change</th>
<th>&lt; 3 kg</th>
<th>&gt;= 3kg</th>
<th>&lt; 3 kg</th>
<th>&gt;= 3kg</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;5% loss</td>
<td></td>
<td>4.10 (1.30-12.57)</td>
<td>4.78 (3.06-7.65)</td>
<td>4.04 (0.77-19.47)</td>
</tr>
<tr>
<td>±5% maintain (ref)</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>&gt;5% gain</td>
<td></td>
<td>1.96 (0.73-4.84)</td>
<td>2.58 (1.63-4.18)</td>
<td>1.13 (0.24-5.18)</td>
</tr>
</tbody>
</table>

* Models adjusted for pre-diagnosis BMI, pre-diagnosis adult weight gain, age at diagnosis, chemotherapy treatment, ER status, PR status and tumor size.
References


CHAPTER 5
POST-DIAGNOSIS PHYSICAL ACTIVITY LEVELS AND SURVIVAL AFTER BREAST CANCER DIAGNOSIS AMONG WOMEN ON LONG ISLAND, NY.

Abstract

Physical activity is associated with physiological responses thought to beneficially affect survival after breast cancer diagnosis, yet results of epidemiologic studies undertaken to evaluate this hypothesis have been inconsistent. Effects of post-diagnosis activity on survival, including stratification by time since diagnosis and by pre-diagnosis body size, were examined in a cohort of 1,423 women diagnosed with in situ or invasive breast cancer in 1996-1997 on Long Island, New York. Subjects were interviewed soon after diagnosis and again after approximately 5 years to assess breast cancer-related factors, including recreational physical activity before and after diagnosis. Date and cause of death through 2005 were determined from the National Death Index. Adjusted estimates were obtained using a proportional hazards regression with time-varying covariates and a selection model to account for missing data. Survival was improved among women who were highly active after diagnosis (>9.0 MET-hours/week) compared to inactive women (0 MET-hours/week) for all-cause [hazard ratio (HR) (95% credible interval): 0.23 (0.15-0.39)] and breast cancer-specific mortality [HR: 0.22 (0.13-0.36)], even with adjustment for pre-diagnosis activity. The beneficial effect of physical activity on breast cancer-specific mortality was evident in all subgroups examined, including: (1) within the first 2 years since diagnosis [HR: 0.17 (0.04-0.58)] as well as 2+ years since diagnosis [HR: 0.17 (0.08-0.33)]; and (2) among women who were overweight in the year prior to diagnosis [HR: 0.13 (0.04-0.32)] as well as those who
were not overweight [HR: 0.23 (0.01-0.47)]. These findings show that post-diagnosis physical activity is associated with improved survival among women with breast cancer.

Introduction

Higher levels of physical activity have been linked to a decreased risk of developing breast cancer (1), which is believed to reflect the ability of physical activity to decrease circulating estrogen, improve insulin sensitivity and improve immune function as well as favorably influence adiposity (2). As these mechanisms are also believed to influence prognosis after a diagnosis of breast cancer, physical activity has gained interest as a potential modifiable lifestyle factor that may improve survival among women with breast cancer. Several recent studies have found that increased physical activity after diagnosis of breast cancer is associated with better survival (3-6), yet the magnitude of these effects is inconsistent and none have addressed if timing of physical activity relative to diagnosis alters these effects. These questions have important public health and clinical implications for the 150,000 women who are diagnosed with breast cancer every year as well as the 2+ million breast cancer survivors in the United States (7).

To assess the effect of physical activity after diagnosis on survival among women with breast cancer we conducted an analysis using data from a large, population-based cohort of women on Long Island, New York who were recently diagnosed with breast cancer. We also considered possible modification of these effects by timing, pre-diagnosis physical activity levels and baseline body size.

Methods

This study employs data from the follow-up to the Long Island Breast Cancer Study Project (LIBCSP), a population-based case-control study of breast cancer risk (8). The follow-up study followed newly diagnosed breast cancer cases from the parent study with the objective
of assessing factors associated with survival after diagnosis. Both the parent and follow-up studies were approved by the Institutional Review Board of participating institutions.

**Study Population.** Participants in the follow-up study were English speaking adult women with a first primary in situ or invasive breast cancer diagnosed in 1996-1997 from Nassau and Suffolk counties in New York and enrolled in the parent LIBCSP case-control study. Incident breast cancer cases were identified through pathology departments of participating hospitals and their physicians were contacted to confirm the diagnosis and obtain permission to contact the patients. Of the 1,837 eligible cases, a total of 1,508 (82.1%) agreed to participate and provided signed informed consent. Of these, 1,414 women agreed to be contacted at a later date for participation in the follow-up. These women were contacted by mail approximately 5 years after diagnosis and informed consent was obtained by telephone with 1,098 women ultimately consenting to the follow-up interview. Of the 316 who subsequently refused to participate, 60 (19%) refused at mail contact, 65 (21%) refused at telephone contact, 18 (6%) would not participate due to illness, 22 (7%) were unable to complete the interview, 55 (17%) were lost to follow-up and 96 (30%) were deceased with no identifiable proxy. Of those agreeing to participate, 1,033 subjects (68.5% of the original 1,508 women) completed the follow-up interview (9), which gathered information relevant since diagnosis.

**Outcome Assessment.** Date and cause of death through December 31, 2005 were established using the National Death Index (NDI) (10), which is considered a standard source of mortality data for epidemiologic research (11). With data from the NDI, for the 1,508 cases from the parent study, we constructed a variable indicating death from any cause (n=308) as well as a variable indicating death due to breast cancer (n=164), which was
determined if breast cancer was listed as the primary or secondary cause of death, using International Classification of Disease code 174.9 or C-50.9. Cases without a death record in the NDI database were determined to be alive as of December 31, 2005. These variables indicating cause of death or censoring, along with the variable indicating length of time since diagnosis, were the outcomes for our analysis.

**Physical Activity Assessment.** Recreational physical activity was assessed through personal interviews, using a modification of the instrument developed by Bernstein and colleagues developed in a previous study of physical activity and breast cancer (12). The baseline questionnaire included a screening question asking “Have you ever participated in any physical activities or exercises on a regular basis—that is for at least 1 hour per week for 3 months or more in any year?” (13); those subjects answering affirmatively were considered having ever participated in physical activity. These subjects were then asked to list each activity separately, and then answer more detailed questions for each one:

1) “In what activity did you (first/next) participate on a regular basis?”
2) “Looking at the calendar, at what age did you start (ACTIVITY) regularly?”
3) “At what age did you stop (ACTIVITY)?”
4) “For how many years did you (ACTIVITY) regularly? For how many months each year did you do this?”
5) “On average, about how many hours per week did you actually (ACTIVITY)?”

For the subjects with detailed physical activity data the number of months per year of each activity was converted to number of hours per week. Where an activity was reported without a corresponding duration, 12 months per year was imputed for non-seasonal activities while the average number of months per year was imputed for seasonal activities. For each activity a metabolic equivalent of energy expenditure (MET) score was assigned using a database derived from previously published values (14). The MET is a standard unit
of energy expenditure for physical activity research (15) and is defined as the energy expended while sitting quietly, approximately one kilocalorie per kilogram of bodyweight per hour. When an activity reported by the study subjects did not have a corresponding MET-score published then a similar activity with a published MET value was chosen. For each activity the MET value was then multiplied by duration of activity in number of hours per week, which was added across all activities for each subject and averaged to arrive at a measure of the average total number of MET-hours per week for each subject.

In the follow-up questionnaire, subjects were asked if they had participated in any physical activities or exercises on a regular basis for at least 1 hour a week for 3 months or more in any year since the date of diagnosis. If an affirmative answer was recorded then detailed questions regarding type, duration, frequency and age at commencement were asked, as described in the baseline questionnaire, for any number of activities, with the time period queried relevant to the time since the diagnosis. From this we calculated an average number of MET-hours per week in the time since diagnosis averaged over the entire follow-up period, and for women with complete data on when activities started and stopped, we also calculated the average number of MET-hours per week for each year from diagnosis up to the time of the follow-up interview, as described above. This yielded a maximum of 7 follow-up measures of physical activity, beginning in the year of diagnosis.

**Covariates.** Self-reported data on potential covariates were gathered through interviewer-administered questionnaires at baseline (in person) and at follow-up (by telephone) and included assessment of menopausal status, education, income, treatment modalities and other factors known and suspected to influence the development and prognosis of breast cancer including anthropometric measures (height in meters and weight
in kilograms in the year before diagnosis) and lifetime physical activity assessments. Approximately 98\% of the respondents completed a self-administered modified Block food frequency questionnaire (FFQ) during the baseline interview, which assessed frequency and portion size for 101 food items in the year prior (8). Data on tumor stage and estrogen and progesterone receptor (ER/PR) status were gathered from medical records of the 1,402 women who signed a medical record release form at the time of the baseline questionnaire. Treatment for the first primary diagnosis of breast cancer and other tumor characteristics were gathered from medical records for 598 of the women who signed a medical record release at the time of the follow-up questionnaire. The treatment data gathered from the medical record matched closely the self-reported data from the follow-up questionnaire (kappa coefficients: radiation therapy $\kappa=0.97$, chemotherapy $\kappa=0.96$ and hormone therapy $\kappa=0.92$ (16)), and thus the self-reported data are used in these analyses. Data on tumor size was obtained from the New York State Cancer Registry.

**Statistical Analysis.** Due to nonresponse to the follow-up questionnaire 33.5\% (n=506) of the study sample were missing information on post-diagnosis physical activity levels. In addition, there was missing information on start and stop dates for 10.6\% (n=160) of the sample, which prevented matching these activities to specific timepoints. However, since these women provided information on absolute duration, it was possible to calculate an average physical activity level over the entire follow-up period. For women with complete data on when activities began and ended, physical activity levels were also created for each year after diagnosis.

To account for missing data in this study we utilized a selection model for the analysis of survival data with time-varying covariates (17). The selection model accounts for
the possibility that data may be not missing at random (NMAR), a condition where the probability that data is missing is dependent upon the unobserved data. The model requires specification of (1) the outcome model, here described with a proportional hazards regression, (2) linear and logistic regression models for the missing covariates and (3) a model describing the probability that data is missing.

The main exposure for this analysis was post-diagnosis physical activity, which was categorized into 0 MET-hrs/wk (referent category), 0.01-9.00 MET-hrs/wk and >9 MET-hrs/wk. For each year of follow-up, the corresponding physical activity variable pertained to activity undertaken in the year prior, which would reflect the etiologically relevant exposure period of recent activity. Within the selection model, the missing physical activity data at each time period was modeled as a linear regression as a function of age at diagnosis, chemotherapy and radiation treatment, menopausal status at diagnosis, body mass index (BMI, weight in kg/squared height in meters) 1-year before diagnosis and lagged physical activity measurements back to diagnosis. Potential confounders for this analysis included variables that have been shown to be related to both survival and post-diagnosis physical activity levels, and included age, pre-diagnosis body mass index in the year prior to diagnosis, chemotherapy treatment, radiation therapy treatment and hormone therapy (18, 19). We also considered menopausal status at diagnosis, total energy intake in the year prior to the baseline interview, stage (in situ vs. invasive), estrogen receptor (ER) status, progesterone receptor (PR) status, tumor size (< 2 cm vs. >= 2 cm) and nodal involvement (yes vs. no). In the final models we included those confounders that changed the effect estimates of the physical activity-survival association by >10% using a complete case analysis. The final models included age, chemotherapy treatment, radiation treatment,
hormone therapy and tumor size, which were also consistent with assessment of confounding using a directed acyclic graph. MET-hrs/wk of physical activity at each timepoint was modeled as a linear regression with age, chemotherapy treatment, radiation therapy, chemotherapy-radiation therapy interaction, menopausal status, pre-diagnosis BMI and all previous measures of MET-hrs/wk. In addition to physical activity, other covariates with significant amounts of missing data included chemotherapy, radiation treatment, hormone therapy and tumor size with 32.2%, 33.0%, 32.1%, 31.6% missing, respectively. For each of these variables, a logistic regression model was specified with age, income (< $20,000; $20,000-49,999; $50,000-89,999 and >= $90,000) and education (high school or less, some college, college graduate and post-college). Within the selection model, the probability that physical activity data was missing was modeled as a logistic regression with age and physical activity assessments.

Other variables with missing data included baseline physical activity (0.93%) menopausal status (1.99% missing), pre-diagnosis BMI (1.13% missing), adult weight change (from age 20 to 1 year before diagnosis; 1.66% missing), education (0.40% missing) and income (0.27% missing). These small amounts were unlikely to influence our results and so the additional cost of estimating models for each of these variables exceeded any benefit, therefore we excluded subjects with any of these variables missing.

Our final analysis then included 1,423 women (94.4% of the original 1,508 cases). Among these women, 291 died during the follow-up with 156 of these deaths due to breast cancer. Median survival time was 8.80 years with times ranging from 0.23 to 9.41 years. Table 1 shows descriptive statistics for sociodemographic and select prognostic factors for the subjects in our analysis. Most women in the study were postmenopausal at diagnosis and
ages ranged from 25 to 91 years. Less than half of the women reported receiving chemotherapy treatment after their diagnosis, while the majority received radiation therapy or hormone therapy.

The model was estimated within a Fully Bayesian framework, specifying vague prior distributions on the model parameters and using the Gibbs sampler in WinBUGS 1.4 (20) to sample from their posterior distribution. We ran the sampler for 50,000 iterations, discarding the first 25,000 as a burn-in sample and retained every 5\textsuperscript{th} iteration to reduce serial correlation. Posterior hazard ratios (HR) and corresponding 95\% credible intervals were calculated by exponentiating the mean of the samples for the log-hazard ratios from the proportional hazards model (beta coefficients) and the 2.5\textsuperscript{th} and 97.5\textsuperscript{th} percentiles of these samples, respectively. Specifying vague priors on the regression parameters yields posterior point estimates and credible intervals that are similar to those from frequentist analysis.

Results

As shown in Table 5.1, most women with breast cancer (76.1\%) reported some form of recreational physical activity after diagnosis, with the majority of those with complete data reporting more than 9 MET-hours per week. This level of activity is equivalent to approximately 108 minutes per week of brisk walking (4 miles per hour, a moderate-intensity activity) or 68 minutes per week of jogging at 5 miles per hour (a vigorous-intensity activity) (21).

Table 5.2 displays the association between mortality and the average physical activity over the entire post-diagnosis follow-up period. Moderate levels (0.1-9.0 MET-hours/week) were associated with improved survival compared to women who were inactive (0 MET-hours/week) for both all-cause [HR (95\% credible interval): 0.41 (0.27-0.62)] and breast cancer-specific mortality [HR: 0.40 (0.25-0.68)], adjusted for pre-diagnostic activity levels
and other relevant factors listed in the methods section. The beneficial effects were even greater for the highest level of activity [>9.0 MET-hours/week vs. 0 MET-hours/week; all-cause HR: 0.21 (0.15-0.30), breast cancer-specific mortality HR: 0.12 (0.07-0.20)].

Table 5.3 shows the estimates for the effects of average post-diagnosis levels of activity on survival stratified by levels of pre-diagnosis physical activity. Strong beneficial effects for the highest physical activity levels undertaken after diagnosis were observed across groups, regardless of the average activity level undertaken before diagnosis. There is a slight suggestion, however, of a stronger inverse association for moderate activity among women who were inactive before diagnosis [among women who expended 0 MET-hours/week before diagnosis, comparing 0.1-9.0 MET-hours/week vs. 0 MET-hours/week, all-cause mortality HR: 0.31 (0.15-0.59)] compared to those who were moderately active or highly active before diagnosis [among women who expended 0.1-9.0 MET-hours week before diagnosis, HR: 0.49 (0.23-1.01); among women who expended >9.0 MET-hours week before diagnosis HR: 0.56 0.25-1.36)]. Similar patterns were observed for breast-cancer specific mortality.

When we considered variable, yearly physical activity levels over time (table 5.4), results for the most-active women were similar to the estimates obtained when we considered only average physical activity levels. For example, for >9.0 MET-hours/week vs. 0 MET-hours/week, the hazards ratio was 0.23 (0.15-0.39) for all-cause mortality, and 0.22 (0.13-0.36) for breast cancer specific mortality. A stronger inverse association, however, was noted for those who were moderately active when variable levels were taken into consideration, especially for breast-cancer specific mortality [0.1-9.0 MET-hours/week vs. 0
MET-hours/week, all-cause mortality HR: 0.35 (0.15-0.70), and breast cancer specific mortality HR: 0.22 (0.13-0.36)].

In table 5.5 we present results stratified by timing of effect to determine if activity before or after diagnosis, or both, is most relevant to enhancing survival. This approach also facilitates comparison of our results to those of other studies, who assessed activity around 2 years or more post-diagnosis. For all-cause mortality, the beneficial effect of variable, yearly physical activity was somewhat greater within the 2 years immediately after diagnosis [>9.0 MET-hours/week vs. 0 MET-hours/week HR: 0.15 (0.03-0.47)] than 2+ years post-diagnosis [>9.0 MET-hours/week vs. 0 MET-hours/week HR: 0.25 (0.15-0.39)], although there are substantial reductions in mortality for both time periods. In contrast, these differences by time since diagnosis were not observed when only deaths due to breast cancer were considered [>9.0 MET-hours/week vs. 0 MET-hours/week, with 2 years HR: 0.17 (0.04-0.58), and after 2 years HR: 0.17 (0.08-0.33)]. Additionally, after 2 years, the effect of moderate activity was nearly equal to the effect of high activity for both all-cause and breast cancer-specific mortality.

To further isolate the effects of post-diagnosis activity from pre-diagnosis activity, we also stratified our analyses by the time period just before and after 1-year post-diagnosis, with the former reflecting recent pre-diagnosis physical activity, and the latter corresponding to post-diagnosis activity. As shown in table 5.5, higher physical activity levels appear to have similar effects on all-cause mortality regardless of timing [before 1-year post-diagnosis HR comparing high activity to inactivity: 0.21 (0.03-0.93); after 1-year post-diagnosis HR comparing high activity to inactivity: 0.25 (0.15-0.40)]. However, for breast cancer related deaths, activity after diagnosis [HR comparing high activity to inactivity: 0.16 (0.09-0.29)]
appeared to be slightly more beneficial than activity before [HR comparing high activity to inactivity: 0.43 (0.06-2.32)], although both time periods showed substantial reductions in mortality in relation to varying physical activity levels.

As shown in table 5.5, results were also stratified by body size at the time of diagnosis, a well-established indicator of prognosis among women with breast cancer. Post-diagnosis activity improved survival for both overweight and non-overweight women, although the effect appeared to be slightly stronger among women who were not overweight in the year prior to diagnosis [BMI < 25 kg/m$^2$ in the year before diagnosis, all-cause mortality HR comparing high activity to inactivity: 0.09 (0.03-0.22)] than those who were overweight [BMI $\geq$ 25 kg/m$^2$ in the year prior to diagnosis, all-cause mortality HR comparing high activity to inactivity: 0.41 (0.24-0.70)].

**Discussion**

We observed that greater levels of physical activity undertaken after diagnosis were associated with substantially lower risk of death from any cause as well as death due to breast cancer in a large, population-based cohort of women who were diagnosed with a first primary breast cancer in 1996-1997. Enhanced survival associated with post-diagnosis physical activity was evident over several different scenarios: (1) when activity levels were considered as average levels over the entire follow-up versus when yearly variations were considered; and (2) when adjustments were made for pre-diagnosis levels of activity and body size, and other parameters relevant to prognosis. The beneficial effect of physical activity appeared slightly stronger in the time period following a diagnosis than activity undertaken before diagnosis, and also among women who were not overweight in the year before diagnosis for both all-cause and breast cancer specific mortality. Nevertheless, the
risk of death was substantially reduced among women who were physically activity after diagnosis in all subgroups that we examined.

Physical activity is associated with several metabolic consequences that may favor survival from both breast cancer and cardiovascular disease (22), which were the first and second most common cause of death, respectively, in our cohort of women with breast cancer. Especially relevant to cancer, higher levels of physical activity are associated with reduction in hormones believed to function as tumor promoters, such as estrogen and insulin. The primary mechanism for these effects is through a reduction in adipose tissue (23, 24), especially metabolically active visceral fat (25), which is a significant source of endogenous estrogen, especially among postmenopausal women (26). Independent of its effect on adiposity, greater physical activity is also associated with an increase in the amount of sex hormone binding globulin (SHBG) and improvement in insulin sensitivity (27). While hormonal pathways offer the most convincing explanations for a protective effect of physical activity on carcinogenesis, moderate physical activity may improve the immune response over both sedentary and exceptionally active individuals (28), possibly by promoting killer-cell, macrophage and cytokine activity (29, 30) as well as upregulating antioxidant enzyme activity (31), which may protect against DNA damage.

Until recently, studies of the effect of physical activity on survival after breast cancer diagnosis focused on pre-diagnostic activity, and results were inconsistent. Two small studies of pre-diagnosis physical activity and survival failed to find associations with breast cancer prognosis (32, 33), although protective effects were noted by Abrahamson et al. (34) and Irwin et al. (5). Only recently have a handful of investigations been conducted on the effects of post-diagnosis physical activity (3-6) with most reporting an approximately 40-
50% reduction in risk of death among highly active women (3, 4, 6). However, the most recent report, by Irwin and colleagues (5), observed a very strong protective effect of being physically active 2 years after diagnosis; for women who expended 9 or more MET-hours/week compared to those who were inactive mortality was reduced by two-thirds (all-cause mortality HR: 0.33, 95% CI: 0.15-0.73), which is similar to the magnitude of effect we observed when considering the effect of this same level of activity after 2 years (all-cause mortality HR: 0.27).

While our findings are in general agreement with a protective effect of increasing physical activity on mortality among breast cancer survivors, our associations were generally stronger in magnitude than those reported by previous studies. Although the categorization of physical activity levels differed across studies, a more likely potential explanation may be the significant differences in study design. Our study followed women forward from date of diagnosis, while the previous analyses included women who were well into their survivorship experience, usually around 2 to 3 years post-diagnosis but as much as 4 years in one study (6), and over 10 years in another (3, 35). Excluding women who do not survive past the first several years could induce length-biased sampling (36) and create a sample of subjects who are more likely to survive regardless of their activity levels. These cohorts are thus not generally representative of all women who are diagnosed with breast cancer. Additionally, the instrument we used for physical activity assessment allowed us to obtain longitudinal measures of physical activity over the entire follow-up period, while most previous studies utilized data from single time periods (e.g. in the year prior to interview). Timing of physical activity assessments is important, as previous research has suggested that among breast cancer survivors, physical activity levels tend to decline during the first year, but show an
upward trend after one year, although only about half of the women return to their pre-diagnosis activity levels by 3 years (19). Failing to fully capture the return to higher levels of physical activity among those who survive longer could partly explain the differences in reported associations.

Strengths of our study include its large size and population-based study design that included women with breast cancer from the time of diagnosis. Additionally, we were able to ascertain physical activity longitudinally from date of diagnosis through time of the follow-up interview, which allowed for evaluation of the effect of temporal changes on the effect of physical activity on survival.

We also employed a novel and rigorous modeling approach to deal with missing data, which is far superior to the *ad hoc* methods often employed in epidemiologic analyses (37). Excluding subjects with any missing data, referred to as complete case analysis, has the potential to induce bias in parameter estimates and results in reduction in statistical efficiency. Methods that use a category for missing data or improper methods of imputation can often perform worse than complete case analysis and are similarly not recommended (37). The Bayesian imputation technique applied here is theoretically sound, amenable to sensitivity analysis and applicable to a wide range of outcome models, including linear and logistic regression (38).

Potential limitations of the study reported here are worthy of mention. These include the use of self-reported physical activity (39), however the comprehensive instrument we used was developed specifically for the study of physical activity and breast cancer (12) and has been successfully used in other studies (40, 41). The use of proxy interviews is also a potential source of bias for physical activity data (42), however the number of proxy
interviews in our dataset was small (<8%). Additionally, a recent report found that data gathered by proxy interview provided nearly identical associations compared to index reports of physical activity (43), suggesting that proxy interviews of physical activity data are reliable. It is also possible that the strong inverse associations noted could be due to healthier women being more physically active while those who were sicker were more inclined to be less active. However, this source of bias is unlikely to entirely explain the effects entirely as we adjusted for tumor characteristics and treatment, and additionally the results were robust when considering physical activity later in the survivorship experience, when such effects would likely be attenuated as those with more severe disease would be likely to have died within the first several years. Also, we had limited power to detect effect modification, therefore our stratified analyses should be interpreted as exploratory, and should be confirmed by other studies. Although a formal methodology was employed to address missing data, the ideal situation would be to have a data set with all variables fully observed that is representative of all breast cancer survivors. Models for missing data may be sensitive to model specification (44), which could have potentially influenced these results.

In summary, our results indicate that physical activity undertaken after diagnosis with breast cancer increases survival among women with breast cancer. We found some suggestion that the beneficial effect may vary slightly by the time since diagnosis, as well as by body size just before diagnosis, however a protective effect was consistently seen in all stratified analyses. Since most women, especially those with a breast cancer diagnosis, do not achieve the recommended amount of physical activity each day (19), these findings suggest that this is an important area for public health interventions.

<table>
<thead>
<tr>
<th></th>
<th>N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deaths as of December 31, 2005</td>
<td></td>
</tr>
<tr>
<td>All cause</td>
<td>291 (20.5)</td>
</tr>
<tr>
<td>Breast-cancer specific</td>
<td>156 (11.0)</td>
</tr>
<tr>
<td>Physical activity over entire follow-up period after diagnosis</td>
<td></td>
</tr>
<tr>
<td>0 MET-hrs/wk</td>
<td>226 (23.9)</td>
</tr>
<tr>
<td>0.1-9.0 MET-hrs/wk</td>
<td>182 (19.2)</td>
</tr>
<tr>
<td>&gt;9 MET-hrs/wk</td>
<td>538 (56.9)</td>
</tr>
<tr>
<td>Missing</td>
<td>477</td>
</tr>
<tr>
<td>Physical activity before diagnosis</td>
<td></td>
</tr>
<tr>
<td>0 MET-hrs/wk</td>
<td>582 (40.9)</td>
</tr>
<tr>
<td>0.1-9.0 MET-hrs/wk</td>
<td>355 (25.0)</td>
</tr>
<tr>
<td>&gt;9 MET-hrs/wk</td>
<td>486 (34.1)</td>
</tr>
<tr>
<td>Physical activity 1 year after diagnosis</td>
<td></td>
</tr>
<tr>
<td>0 MET-hrs/wk</td>
<td>213 (27.1)</td>
</tr>
<tr>
<td>0.1-9.0 MET-hrs/wk</td>
<td>94 (11.9)</td>
</tr>
<tr>
<td>&gt;9 MET-hrs/wk</td>
<td>480 (61.0)</td>
</tr>
<tr>
<td>Missing</td>
<td>619</td>
</tr>
<tr>
<td>Physical activity 2 years after diagnosis</td>
<td></td>
</tr>
<tr>
<td>0 MET-hrs/wk</td>
<td>193 (25.1)</td>
</tr>
<tr>
<td>0.1-9.0 MET-hrs/wk</td>
<td>96 (12.5)</td>
</tr>
<tr>
<td>&gt;9 MET-hrs/wk</td>
<td>480 (62.4)</td>
</tr>
<tr>
<td>Missing</td>
<td>602</td>
</tr>
<tr>
<td>Physical activity 3 years after diagnosis</td>
<td></td>
</tr>
<tr>
<td>0 MET-hrs/wk</td>
<td>183 (23.8)</td>
</tr>
<tr>
<td>0.1-9.0 MET-hrs/wk</td>
<td>96 (12.5)</td>
</tr>
<tr>
<td>&gt;9 MET-hrs/wk</td>
<td>490 (63.7)</td>
</tr>
<tr>
<td>Missing</td>
<td>560</td>
</tr>
<tr>
<td>Physical activity 4 years after diagnosis</td>
<td></td>
</tr>
<tr>
<td>0 MET-hrs/wk</td>
<td>177 (23.2)</td>
</tr>
<tr>
<td>0.1-9.0 MET-hrs/wk</td>
<td>89 (11.7)</td>
</tr>
<tr>
<td>&gt;9 MET-hrs/wk</td>
<td>496 (65.1)</td>
</tr>
<tr>
<td>Missing</td>
<td>542</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Physical activity 5 years after diagnosis</th>
<th>0 MET-hrs/wk</th>
<th>183 (24.7)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1-9.0 MET-hrs/wk</td>
<td>83 (11.2)</td>
</tr>
<tr>
<td></td>
<td>&gt;9 MET-hrs/wk</td>
<td>474 (64.0)</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>524</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Physical activity 6 years after diagnosis</th>
<th>0 MET-hrs/wk</th>
<th>349 (62.3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1-9.0 MET-hrs/wk</td>
<td>30 (5.4)</td>
</tr>
<tr>
<td></td>
<td>&gt;9 MET-hrs/wk</td>
<td>181 (32.3)</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>668</td>
</tr>
</tbody>
</table>

| ER status | Negative | 251 (26.6) |
|           | Positive | 694 (73.4) |
|           | Missing | 478 |

| PR status | Negative | 338 (35.9) |
|           | Positive | 603 (64.1) |
|           | Missing | 482 |

| Chemotherapy | No | 568 (58.8) |
|              | Yes | 398 (41.2) |
|              | Missing | 457 |

| Radiation therapy | No | 379 (39.1) |
|                   | Yes | 591 (60.9) |
|                   | Missing | 453 |

| Hormone therapy | No | 367 (38.5) |
|                 | Yes | 586 (61.5) |
|                 | Missing | 470 |

| Tumor size | < 2 cm | 442 (80.5) |
|           | >= 2 cm | 107 (19.5) |
|           | Missing | 874 |

<table>
<thead>
<tr>
<th>BMI 1 year before diagnosis</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;18.5 kg/m²</td>
<td>25 (1.8)</td>
<td></td>
</tr>
<tr>
<td>18.5-24.9 kg/m²</td>
<td>633 (44.5)</td>
<td></td>
</tr>
<tr>
<td>25.0-29.9 kg/m²</td>
<td>451 (31.7)</td>
<td></td>
</tr>
<tr>
<td>&gt;=30.0 kg/m²</td>
<td>314 (22.1)</td>
<td></td>
</tr>
</tbody>
</table>

Weight change from age 20 to 1 year before diagnosis

<table>
<thead>
<tr>
<th>Weight change</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 3kg gain or any loss</td>
<td>238 (16.7)</td>
<td></td>
</tr>
<tr>
<td>&gt; 3kg gain</td>
<td>1,185 (83.3)</td>
<td></td>
</tr>
</tbody>
</table>

Age at diagnosis (mean: 58.8 yrs, sd: 12.7 yrs)

<table>
<thead>
<tr>
<th>Age at diagnosis</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>20-29.9</td>
<td>10 (0.7)</td>
<td></td>
</tr>
<tr>
<td>30-39.9</td>
<td>77 (5.4)</td>
<td></td>
</tr>
<tr>
<td>40-49.9</td>
<td>302 (21.2)</td>
<td></td>
</tr>
<tr>
<td>50-59.9</td>
<td>377 (26.5)</td>
<td></td>
</tr>
<tr>
<td>60-69.9</td>
<td>337 (23.7)</td>
<td></td>
</tr>
<tr>
<td>70-79.9</td>
<td>272 (19.1)</td>
<td></td>
</tr>
<tr>
<td>80-89.9</td>
<td>45 (3.2)</td>
<td></td>
</tr>
<tr>
<td>90+</td>
<td>3 (0.2)</td>
<td></td>
</tr>
</tbody>
</table>

Menopausal Status at diagnosis

<table>
<thead>
<tr>
<th>Menopausal Status</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Premenopausal</td>
<td>456 (32.0)</td>
<td></td>
</tr>
<tr>
<td>Postmenopausal</td>
<td>967 (68.0)</td>
<td></td>
</tr>
</tbody>
</table>

Education

<table>
<thead>
<tr>
<th>Education</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>High school or less</td>
<td>683 (48.0)</td>
<td></td>
</tr>
<tr>
<td>Some college</td>
<td>339 (23.8)</td>
<td></td>
</tr>
<tr>
<td>College graduate</td>
<td>185 (13.0)</td>
<td></td>
</tr>
<tr>
<td>Post college education</td>
<td>216 (15.2)</td>
<td></td>
</tr>
</tbody>
</table>

Income at diagnosis

<table>
<thead>
<tr>
<th>Income at diagnosis</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; $20,000</td>
<td>174 (12.2)</td>
<td></td>
</tr>
<tr>
<td>$20,000-49,999</td>
<td>555 (39.0)</td>
<td></td>
</tr>
<tr>
<td>$50,000-89,999</td>
<td>420 (29.5)</td>
<td></td>
</tr>
<tr>
<td>&gt;=$90,000</td>
<td>274 (19.3)</td>
<td></td>
</tr>
</tbody>
</table>

* Average MET-hours per week for recreational physical activity over relevant interval. Note that the total number of subjects available for each assessment declines over time as subjects leave the cohort due to death or censoring.
Table 5.2. Hazard ratios (and 95% credible intervals) for the association between all-cause and breast cancer-specific mortality, and average post-diagnosis physical activity levels (MET-hrs/wk) over the entire follow-up period, among a cohort of women with a first primary breast cancer on Long Island, NY, and followed through December 31, 2005.

<table>
<thead>
<tr>
<th>Post-diagnosis Physical Activity</th>
<th>Hazard ratio* (95% credible interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All-cause Mortality</td>
</tr>
<tr>
<td>Average Met-hrs/wk</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1.</td>
</tr>
<tr>
<td>0.1-9.0</td>
<td>0.41 (0.27-0.62)</td>
</tr>
<tr>
<td>&gt;9.0</td>
<td>0.21 (0.15-0.30)</td>
</tr>
</tbody>
</table>

* Models adjusted for age, chemotherapy treatment, radiation therapy, hormone therapy, tumor size and pre-diagnosis physical activity levels (MET-hrs/wk).

Table 5.3. Hazard ratios (and 95% credible intervals) for the association between all-cause and breast cancer specific mortality and average post-diagnosis physical activity levels (PA) over entire follow-up period, stratified by pre-diagnosis physical activity (PA) levels, among women diagnosed with a first primary breast cancer in 1996-1997 on Long Island, NY, and followed through December 1, 2005.

<table>
<thead>
<tr>
<th>Post-diagnosis Physical Activity</th>
<th>All-cause Mortality (292 deaths/1,436 subjects)</th>
<th>Breast Cancer-Specific Mortality (156 deaths/1,436 subjects)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre-diagnosis PA 0 Met-hrs/wk</td>
<td>Pre-diagnosis PA 0.1-9.0 Met-hrs/wk</td>
</tr>
<tr>
<td>Average Met-hrs/wk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1.</td>
<td>1.</td>
</tr>
<tr>
<td>0.1-9.0</td>
<td>0.31 (0.15-0.59)</td>
<td>0.49 (0.23-1.01)</td>
</tr>
<tr>
<td>&gt;9.0</td>
<td>0.26 (0.15-0.44)</td>
<td>0.21 (0.09-0.43)</td>
</tr>
</tbody>
</table>

* Models adjusted for age, chemotherapy treatment, radiation therapy, hormone therapy and tumor size.
Table 5.4. Hazard ratios (and 95% credible intervals) for the association between all-cause and breast-cancer specific mortality and post-diagnosis physical activity levels (met-hrs/wk) assessed yearly over entire follow-up, among women diagnosed with a first primary breast cancer in 1996-1997 on Long Island NY, and followed through December 31, 2005.

<table>
<thead>
<tr>
<th>Post-diagnosis Physical Activity</th>
<th>Hazard ratio* (95% credible interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yearly Met-hrs/wk</td>
<td>All-cause Mortality</td>
</tr>
<tr>
<td>0</td>
<td>1.</td>
</tr>
<tr>
<td>0.1-9.0</td>
<td>0.35 (0.15-0.70)</td>
</tr>
<tr>
<td>&gt;9.0</td>
<td>0.23 (0.15-0.39)</td>
</tr>
</tbody>
</table>

* Models adjusted for age, chemotherapy treatment, radiation therapy, hormone therapy and tumor size.
Table 5.5. Hazard ratios (and 95% credible intervals) for the association between all-cause and breast cancer-specific mortality and yearly post-diagnosis physical activity (PA) levels, stratified by time since diagnosis and pre-diagnosis BMI, among women diagnosed with a first primary breast cancer in 1996-1997 on Long Island, NY, and followed through December 31, 2005.

<table>
<thead>
<tr>
<th>Time since diagnosis:</th>
<th>Hazard ratio* (95% credible interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Post-diagnosis PA</strong></td>
<td><strong>All-cause Mortality</strong></td>
</tr>
<tr>
<td>(292 deaths/1,436 subjects)</td>
<td>(156 deaths/1,436 subjects)</td>
</tr>
<tr>
<td>Yearly Met-hrs/wk</td>
<td>0-2 years</td>
</tr>
<tr>
<td>0</td>
<td>1.</td>
</tr>
<tr>
<td>0.1-9.0</td>
<td>0.39 (0.11-1.11)</td>
</tr>
<tr>
<td>&gt;9.0</td>
<td>0.15 (0.03-0.47)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time since diagnosis:</th>
<th>Time since Diagnosis:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yearly Met-hrs/wk</td>
<td>0-1 year</td>
</tr>
<tr>
<td>0</td>
<td>1.</td>
</tr>
<tr>
<td>0.1-9.0</td>
<td>0.61 (0.15-1.99)</td>
</tr>
<tr>
<td>&gt;9.0</td>
<td>0.21 (0.03-0.93)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BMI one year before diagnosis</th>
<th>BMI one year before diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yearly Met-hrs/wk</td>
<td>&lt; 25</td>
</tr>
<tr>
<td>0</td>
<td>1.</td>
</tr>
<tr>
<td>0.1-9.0</td>
<td>0.17 (0.03-0.60)</td>
</tr>
<tr>
<td>&gt;9.0</td>
<td>0.09 (0.03-0.22)</td>
</tr>
</tbody>
</table>

* Models adjusted for age, chemotherapy treatment, radiation therapy, hormone therapy and tumor size.
References


CHAPTER 6
DISCUSSION

Summary

A statistical model to address potentially non-ignorably missing time-varying covariates in survival analysis was developed. This technique, and a slight modification, was applied to two analyses of a large, population-based follow-up study of women with a recent diagnosis of breast cancer to assess if post-diagnosis weight change and physical activity levels were associated with survival.

The results showed that women who had gained weight after diagnosis had worse survival when deaths from all causes as well as deaths specifically due to breast cancer were considered. When considering post-diagnosis weight change, the effects for pre-diagnosis body size that have been consistently associated with poorer prognosis in the recent literature, body mass index (BMI) before diagnosis and adult weight gain before diagnosis, were noticeably attenuated, suggesting that most of their reported effect is mediated through weight change after diagnosis. In a stratified analysis, the effect was greater when the post-diagnosis weight gain occurred closer to diagnosis, when chemotherapy-related weight gain is of concern, compared to the time farther from diagnosis, however the estimates were not very precise. Additionally, weight gain after diagnosis appeared to be more deleterious for women who are heavier before diagnosis or who have gained weight as an adult before diagnosis.

Being physically active after diagnosis was associated with better survival, even with adjustments included for the effects of activity undertaken before diagnosis. In addition, the
inverse association with activity was noted regardless of the timing of the outcome relative to the date of diagnosis. Further, the effect may be more pronounced among women who are leaner before diagnosis, but again the estimates were imprecise.

The findings for both the change in weight and physical activity analyses are in general agreement with other recent reports (1-8), although the magnitude of the associations reported here are stronger. Recent reports of weight change and survival in similar populations have been mixed (5-7), although methodology has differed substantially across reports, with notable differences in study design, exposure assessment and definition of cutpoints for categorization of exposures. Nevertheless, two of these three previous studies report some degree of a positive association between weight gain and survival, even if only in subgroups of women (5, 7). The findings of this work generally agree with a U-shaped relationship between weight change and survival that have been described among the general population, with greater risk of mortality observed for those who lose or gain substantial amounts of weight, and the lowest risk of death observed for those who maintain, or gain little weight (9). Although weight gain and absolute obesity likely represent somewhat different exposures, as one can gain weight and not necessarily become obese, it is worth noting that in the general population, a similar pattern has been reported with excess deaths among the underweight (body mass index, BMI < 18.5) and obese (BMI ≥ 30.0), compared to those in the normal weight (BMI ≥ 18.5 and < 25) and overweight (BMI ≥ 25 and < 30) categories (10).

Recent analyses of physical activity and survival among women with breast cancer have consistently reported inverse associations (1-4, 8), although the effects reported here are
stronger. Possible reasons for the larger effect sizes found in the analyses undertaken here as compared to previous reports are discussed below.

**Missing Data**

A particular strength of this work was the development and application of a theoretically sound approach to the treatment of missing data. Missing covariate data is quite common in epidemiologic research, especially in longitudinal studies when subjects may be lost to follow-up, fail to remember, or simply refuse to answer certain questions. This issue is especially problematic when the variables of interest are sensitive in nature and so the likelihood that they are missing is dependent upon their actual (unobserved) value, in which case the data are not missing at random (NMAR). The text by Little and Rubin (11) provides a comprehensive overview of missing data in general and proposes the taxonomy of missing completely at random (MCAR), where the probability that data is missing is independent of observed or unobserved data, missing at random (MAR), when the probability that data is missing is dependent on observed data, and NMAR, when the probability that data is missing depends upon the unobserved missing values. Several authors have recently reviewed the issue of missing data in the context of epidemiologic and public health research (12, 13) including a discussion of proper and improper analytic approaches. Since most statistical software packages automatically exclude any subject with missing data on one or more variable such complete case analyses are common, although when there is a significant amount of missing data they result in a loss of statistical efficiency and under many circumstances can yield significantly biased estimates. The results presented in chapter 3 underscore the dangers associated with complete case analysis in particular: not only were the effect estimates for weight change attenuated from the more complete selection model results, but weight gain showed an inverse association, and ER-positive status showed a
positive association with breast cancer mortality, which are in direct contrast to associations found in the more comprehensive approach that was developed for this analysis as well as with clinical evidence. _Ad hoc_ methods are also widely employed, as many investigators still utilize some variant of the “missing indicator” method (12), where subjects with missing values are assigned to a unique category representing those with missing data, or improper imputation, where missing values are estimated from sample means, predicted from regression models or simply carried forward from previous observations (in longitudinal studies) and then treated as “true” data with no adjustment for the fact that such predicted values carry with them a degree of uncertainty. Both of these methodologies are to be avoided as they can induce significant bias and, in the case of improper imputation, underestimate standard errors (12, 14, 15).

The approach taken for this study was to extend the selection model proposed by Herring et al. (16) to allow for time-varying covariates, which to the best of my knowledge has not been addressed in the previous statistics literature. The method presented in chapter 3 is very flexible, can address both MAR and NMAR data and easily lends itself to sensitivity analyses to explore changes in these assumptions and model specifications. Sensitivity analysis is especially important in cases where NMAR is suspected since selection models can be quite sensitive to changes in specification (16) and missing data methods in general rely heavily on untestable assumptions.

Formal approaches to regression analysis with missing covariate data include maximum likelihood, multiple imputation, weighted estimating equations and Fully Bayesian (FB), which are described in general by Ibrahim et al. (17). The approach taken for this analysis was FB, which requires specification of models for the outcome, missing covariates
and the missing data mechanism (in the case of NMAR covariates), as well as prior distributions on each of the parameters for the aforementioned models. The Bayesian framework offers considerable flexibility since sensitivity analyses to assumptions of the missing data mechanism are easily conducted. Additionally, estimation is readily accomplished through use of the Gibbs sampler in WinBUGS (18) which is computationally quicker and requires considerably less coding than comparable likelihood-based techniques. Although the outcome for this study was survival, this method can easily be applied to any outcome, including binary data for the analysis of case-control studies. Bayesian methods are becoming more accepted in the analysis of epidemiologic data (19-24) and the technique outlined here should be accessible to many practicing epidemiologists with moderate quantitative training.

**Study Design**

An additional strength of this analysis is that it utilized data from a population-based cohort of women who were followed forward from the date of diagnosis of breast cancer. Most of the recent studies of both post-diagnosis weight change (5-7) and post-diagnosis physical activity (1-4) have been limited to designs where follow-up recruitment and exposure assessments were made several years after the date of diagnosis. In one study population that has yielded two reports (1, 7), single post-diagnosis exposure assessments were made an average of nearly 6 years, and as much as 16 years after diagnosis. According to recent SEER statistics, among women who were diagnosed with breast cancer in 1996, a total of 8.8% died during the first 2 years and 13% died in the first 5 years, while only an additional 6.2% (19.2% total) died in the second five years after diagnosis (25). Although absolute survival rates have improved over the years, mostly due to more effective treatments, this pattern remains consistent: those who survive past the first several years are more likely to
continue surviving. Such samples are unlikely to represent the general population of breast
cancer survivors.

An important issue surrounding previous study designs is the timing of exposure
assessments. Body size and physical activity were typically assessed only once, near the
time of enrollment (e.g. 2 or more years after diagnosis), and often not at uniform points
within a given study. Capturing the longitudinal patterns of both body size and physical
activity are important among breast cancer survivors as changes in body composition are
associated with length of time on treatment regimens, especially chemotherapy (26), and
weight gain usually peaks around the first year after diagnosis. Additionally, among women
with breast cancer, physical activity levels tend to decline during the first year and increase
thereafter, although only about half of the women have returned to their pre-diagnosis levels
by 3 years (27). Given the association between physical activity and body size, this could
also have implications not only for assessment of physical activity as an independent
prognostic factor, but also for weight change. If these factors are not measured consistently
and at relevant intervals from the date of diagnosis then important changes could be missed
and bias could result, especially when coupled with sampling issues discussed above. This
study is the first to utilize multiple assessments of both body size and physical activity, which
allowed the assessment of how weight change and physical activity levels at various times in
the survival experience affect prognosis.

**Exposure Assessment**

Self-reported exposures, such as body size and physical activity, are often of concern given
issues related to measurement error and recall bias. Sensitive questions such as weight are
often believed to be under-reported by heavier individuals and over-reported by lighter
individuals (28), however a study of women of similar age to this population showed that
self-reported anthropometric measures are highly accurate and comparable to measurements obtained in a clinic setting (29).

Validity of self-reported physical activity questionnaires are also often in question (30), however the instrument used for this study was developed for studies of physical activity and breast cancer (31) and its semi open-ended format offers an extremely comprehensive assessment of type, duration and frequency of recreational physical activity over broad periods of time. Although a prospective assessment of physical activity may have provided more accurate short-term assessments, the use of alternative assessments such as activity diaries or even objective measurement tools such as pedometers or accelerometers would not have been able to be utilized to the extent to assess long-term physical activity patterns in such a large study population.

**Study Limitations**

While this study has several notable strengths, its limitations should also be acknowledged. The study population consisted of primarily white women of generally higher socioeconomic status compared to the rest of the nation and thus their experiences may not be representative of the survival experience of all women who are diagnosed with breast cancer, in particular those of other racial or socioeconomic groups. Consequently, these results may not be generalizable to all women with breast cancer.

Due to sparseness of the data that made model convergence difficult, I was unable to examine the joint effects of physical activity and weight change. As physical activity and body size are linked, it would have been informative to examine each variable as a confounder, and more interestingly a potential effect modifier for the other. Although this was a limitation of this study, the analysis presented here illustrated the beneficial effect of physical activity regardless of baseline body size, which appears to be a marker for post-
diagnosis weight change. Furthermore, baseline body size was not a confounder for the analysis of post-diagnosis physical activity, nor was baseline physical activity a confounder for post-diagnosis weight change, which could be consistent with independent effects of each factor.

This analysis could have benefited from assessment of changes in more specific measures of adiposity, such as percent body fat, or changes in measures of fat patterning, such as waist-to-hip ratio, which are believed to be strongly associated with metabolically active central adiposity. To date, no study has examined the effect of post-diagnosis changes in fat patterning on survival after breast cancer, and so it is unclear if this, or weight change, is a more useful predictor of mortality. Although this would have yielded additional insight, this analysis did focus on weight change, which is also strongly associated with changes in central adiposity among adults (32).

This analysis could have also benefited from more complete and detailed stage and treatment data. Although I was able to adjust for tumor characteristics and gross measures of treatment regimen they did have significant amounts of missing data and the additional resolution afforded by TNM stage and more detailed treatment assessment could help reduce the potential for residual confounding by these variables in this analysis. While this additional information would have yielded more complete control of confounding, it is extremely unlikely that including detailed stage or treatment variables would have significantly attenuated the associations I observed.

**Future Directions**

The literature on the association of obesity and survival would benefit from additional work examining the effects of fat patterning and overall body composition, as it is believed that visceral fat is more detrimental than subcutaneous adipose tissue. Although adult weight
gain is often associated with an increase in visceral fat, treatment regimens and reductions in activity levels seen among breast cancer patients are also associated with changes in body composition such as loss of lean body mass in addition to a gain in adiposity (33). Increase in percent body fat due to reduction in lean body mass may not be fully captured by weight change, if total weight remains the same. A more detailed inspection of changes in body composition could further highlight areas for intervention.

Studies of physical activity and breast cancer survival have, to date, focused on general measures of activity, usually MET-hours per week or simply hours per week in various activities. Future work should assess the effects of different types of activities, including resistance and aerobic training. Additionally, complementary and alternative therapies are increasing in popularity among breast cancer patients (34), including therapies that incorporate physical activity, such as yoga and pilates (35-37). Each form of exercise tends to influence different physiological processes, with resistance training associated with increases in lean body mass (38, 39) and glycemic control (40, 41), while aerobic activity is associated with reduction in body size and favorable changes in risk factors for cardiovascular disease (42). Yoga and pilates have also shown potential for helping women regain mobility (37), and manage stress and fatigue (35, 36). The identification of differential effects of each type of activity could assist in designing post-diagnostic rehabilitation programs for breast cancer patients as well as public health interventions for these women in general.

Another area that warrants attention is the identification of a “healthy lifestyle pattern” that could be associated with improved survival. Diet, exercise and weight management all interact with each other in a manner that impacts the metabolic environment
affecting both cancer and cardiovascular disease, especially insulin resistance (42). There has been recent work on identifying dietary patterns that may be associated with survival among cancer patients (43, 44), however this approach could also be applied to identification of a general behavioral pattern among breast cancer survivors, such as a combined diet and physical activity pattern of those with the lowest mortality. This approach has been explored recently by one group that explored survival by a combination of fruit and vegetable intake and physical activity and they found that survival was highest among women with high fruit and vegetable intake and high physical activity than among any other combination of these behaviors (3). Unfortunately such interactions were not possible to explore in the study on which this dissertation is based as only 679 (45%) of the case women completed a follow-up food frequency questionnaire.

As intriguing as these dietary intake findings are, research focused on disentangling the effects of a healthy lifestyle on breast cancer survival would benefit from more rigorous methodologies for data reduction, such as factor analysis (44), reduced rank regression (42) or Bayesian hierarchical models (19, 45). This information could not only lend itself to targeted public health messages, but also enhance our understanding of the biology of survival after a diagnosis of breast cancer.

**Conclusion**

Findings from this population-based cohort of women with breast cancer, which utilized a statistical approach developed specifically for this study, indicate that survival is hampered among women who gain or lose weight after diagnosis, but enhanced among those who are physically active post-diagnosis even if they were overweight or inactive before diagnosis. Future studies should focus on determining which specific combination of characteristics of a healthy lifestyle after diagnosis, including consideration of changes in dietary intake,
increase survival among women with breast cancer. Intervention studies could also be undertaken to identify effective methods to assist women newly diagnosed with breast cancer to maintain their weight and engage in physical activity during the post-diagnosis period, including during the time period that they are receiving treatment for their disease.
References


APPENDIX
WinBUGS CODE

The software used to estimate the models for these analyses was WinBUGS version 1.4 (1); to invoke WinBUGS I used the R programming language as a front end to pass the data and parameters (http://www.stat.columbia.edu/~gelman/bugsR/) which offers a very convenient approach to using the WinBUGS program. Below is an example of the WinBUGS code used to estimate the non-ignorably missing data models discussed in chapter 3 (model 2 in table 3.2). The data elements for each subject are: dur (length of follow up for each subject), death (indicator of censoring or death), fu.years (time of final follow-up assessment), dpwt.fu (a vector of the three assessments of weight change), bmiref (BMI one year before diagnosis), wtchg20.cat (adult weight change), chemo (chemotherapy indicator), erstat (ER positive indicator), prstat (PR positive indicator), tumor2 (indicator tumor size >2cm), dxage (continuous age in years at diagnosis), income, educ (education level), postmenp (indicator of postmenopausal status). The parameters of the model are: N=1436, the total number of subjects; J=10, the number of intervals for the proportional hazards model and eps=10^-6 is a parameter that allows for definition of strict inequalities with the step function.

model select;
{
  # Partition time axis
  for (k in 1:J+1) {
    a[k] <- 10*(k-1)/J; # I chose 10, evenly spaced intervals
  }

  for (i in 1:N) {
    for (k in 1:J) {

      # Indicator if event-time in interval k
      d[i,k]<- death[i]*step(dur[i]-a[k]+eps)*step(a[k+1] - dur[i]);
    }
  }

# length of overlap of dur[i] with interval k
delta[i,k] <- (min(dur[i], a[k+1]) - a[k])*step(dur[i]-a[k]);

# Assign exposure to correct interval
dpwt[i,k] <- dpwt.fu[i,1]*step(1.0-a[k]-eps) +
   dpwt.fu[i,2]*(step(a[k]-1.0)*(step(fu.years[i],a[k]-eps) +
   equals(fu.years[i],1.0))) +
   dpwt.fu[i,3]*step(a[k]-fu.years[i])*step(fu.years[i]-1.0-eps);

# Assign dpwt to categories
dpwt0[i,k] <- step(-5 - dpwt[i,k]+eps); # > 5% loss in bodyweight
dpwt1[i,k] <- step(dpwt[i,k] + 5)*step(5 - dpwt[i,k]);
# change +/- 5% [Maintenance, REF]
dpwt2[i,k] <- step(dpwt[i,k] - 5 + eps)*step(10 - dpwt[i,k]);
# change in bodyweight > than 5% and <=10%
dpwt3[i,k] <- step(dpwt[i,k] - 10 + eps);
# change in bodyweight > 10%

# [T|X,Z]--model for time to event conditional upon observed
# and unobserved variables
theta[i,k] <- lambda[k]*exp(beta[1]*dpwt0[i,k] +
   beta[2]*dpwt2[i,k] + beta[3]*dpwt3[i,k] +
   beta[4]*step(bmiref[i]-25.0)*step(30.0-bmiref[i]-eps) +
   beta[5]*step(bmiref[i]-30.0) +
   beta[6]*equals(wtchg20.cat[i],0) +
   beta[7]*equals(wtchg20.cat[i],2) +
   beta[8]*chemo[i] + beta[9]*erstat[i] +
   beta[10]*prstat[i] + beta[11]*tumor2[i] +
   beta[12]*dxage[i]);

# define the likelihood
d[i,k] ~ dpois(mu[i,k]);
u[i,k] <- delta[i,k]*theta[i,k];

# MODEL 2: Models for missing covariates
logit(p.r.chemo[i]) <- phi.chemo[1] +
   phi.chemo[2]*dxage[i] +
   phi.chemo[3]*(equals(income[i],3)+equals(income[i],4)+
   equals(income[i],5)) +
   phi.chemo[4]*(equals(income[i],6)+equals(income[i],7)) +
   phi.chemo[5]*(equals(income[i],8)) +
   phi.chemo[6]*equals(educ[i],3) +
   phi.chemo[7]*equals(educ[i],4) +
   phi.chemo[8]*equals(educ[i],5);

chemo[i] ~ dbin(p.r.chemo[i],1);

logit(p.r.erstat[i]) <- phi.erstat[1] + phi.erstat[2]*dxage[i];
erstat[i] ~ dbin(p.r.erstat[i],1);

logit(p.r.prstat[i]) <- phi.prstat[1] + phi.prstat[2]*dxage[i];
prstat[i] ~ dbin(p.r.prstat[i],1);
logit(p.r.tumor2[i]) <- phi.tumor2[1] + 
phi.tumor2[2]*dxage[i] + 
phi.tumor2[3]*(equals(income[i],3)+equals(income[i],4)+
equals(income[i],5)) + 
phi.tumor2[4]*(equals(income[i],6)+equals(income[i],7)) + 
phi.tumor2[5]*(equals(income[i],8)) + 
phi.tumor2[6]*equals(educ[i],3) + 
phi.tumor2[7]*equals(educ[i],4) + 
phi.tumor2[8]*equals(educ[i],5) ;
tumor2[i] ~ dbin(p.r.tumor2[i],1);

# Models for missing weight change
# 3rd follow-up:
mu.dpwt[i,3] <- alpha3[1] + alpha3[2]*dxage[i] + 
alpha3[3]*chemo[i] + alpha3[4]*postmenp[i] + 
alpha3[5]*bmiref[i] + alpha3[6]*dpwt.fu[i,2] + 
alpha3[7]*dpwt.fu[i,1];
dpwt.fu[i,3] ~ dnorm(mu.dpwt[i,3],tau3.dpwt);

# 2nd follow-up:
alpha2[3]*chemo[i] + alpha2[4]*postmenp[i] + 
alpha2[5]*bmiref[i] + alpha2[6]*dpwt.fu[i,1];
dpwt.fu[i,2] ~ dnorm(mu.dpwt[i,2],tau2.dpwt);

# 2nd follow-up:
mu.dpwt[i,1] <- alpha1[1] + alpha1[2]*dxage[i] + 
alpha1[3]*chemo[i] + alpha1[4]*postmenp[i] + 
alpha1[5]*bmiref[i];
dpwt.fu[i,1] ~ dnorm(mu.dpwt[i,1],tau1.dpwt);

# Missingness models for each follow-up assessment:
# 3rd follow-up:
logit(p.r.fu[i,3]) <- gamma3[1] + gamma3[2]*dpwt.fu[i,3] + 
gamma3[3]*dxage[i] + gamma3[4]*fu.years[i] + 
gamma3[5]*p.r.fu[i,1] + gamma3[6]*p.r.fu[i,2];
r.fu[i,3] ~ dbin(p.r.fu[i,3],1);

# 2nd follow-up:
logit(p.r.fu[i,2]) <- gamma2[1] + gamma2[2]*dpwt.fu[i,2] + 
gamma2[3]*dxage[i] + gamma2[4]*p.r.fu[i,1];
r.fu[i,2] ~ dbin(p.r.fu[i,2],1);

# 1st follow-up:
logit(p.r.fu[i,1]) <- gammal[1] + gammal[2]*dpwt.fu[i,1] + 
gammal[3]*dxage[i];
r.fu[i,1] ~ dbin(p.r.fu[i,1],1);

} # PRIORS ON PARAMETERS
# Parameters for survival model
for (k in 1:J) { lambda[k] ~ dgamma(0.01, 0.01); }
for (l in 1:12) { beta[l] ~ dnorm(0, 0.00001); }

# Parameters for dpwt models
for (l in 1:7) { alpha3[l] ~ dnorm(0, 0.00001); }
for (l in 1:6) { alpha2[l] ~ dnorm(0, 0.00001); }
for (l in 1:5) { alpha1[l] ~ dnorm(0, 0.00001); }

# Parameters for missing data models
for (l in 1:8) { phi.chemo[l] ~ dnorm(0, 0.00001); }
for (l in 1:8) { phi.tumor2[l] ~ dnorm(0, 0.00001); }
for (l in 1:2) { phi.erstat[l] ~ dnorm(0, 0.00001); }
for (l in 1:2) { phi.prstat[l] ~ dnorm(0, 0.00001); }

tau3.dpwt ~ dgamma(0.01, 0.01);
tau2.dpwt ~ dgamma(0.01, 0.01);
tau1.dpwt ~ dgamma(0.01, 0.01);

# Parameters for missingness model
for (l in 1:6) { gamma3[l] ~ dnorm(0, 0.00001); }
for (l in 1:4) { gamma2[l] ~ dnorm(0, 0.00001); }
for (l in 1:3) { gamma1[l] ~ dnorm(0, 0.00001); }
}
References