The Relationship between and Geographic Distribution of Breast Cancer Statistics: Diagnosis, Survival, and Mortality in Selected Areas in the United States, 1973-2004[1]

Timothy Rooney

Dr. Charles Becker, *Advisor*

Dr. Michelle Connolly, *Seminar Advisor*

Duke University
Durham, North Carolina
2014

[1] Honors thesis submitted in partial fulfillment of the requirements for Graduation with Distinction in Economics in Trinity College of Duke University

Acknowledgements

The author would like to thank Dr. Charles Becker and Dr. Michelle Connolly for their invaluable help in completing this thesis. Additionally, the students of the Honors Seminar classes have provided helpful feedback and advice, thank you for all of your assistance.

Abstract

Using breast cancer registry data from the United States and regression models controlling for race, marital status, and county-level variation, this research analyzes the connections between these statistics and the geographic variation of each of them. In doing so, it determines that stage of diagnosis has a significant impact on survival likelihood and the likelihood of death due to breast cancer. It also determines that survival reduces mortality likelihood. Additionally, it determines that stage of diagnosis, survival, and mortality all vary geographically, postulating that the reason for this variation is due to lifestyle variation and uneven medical talent distribution.

JEL Classification: I1, I10, I19

Keywords: Health, Cancer, Diagnosis, Survival, Mortality

"*In the United States, one in three women and one in two men will develop cancer during their lifetime. A quarter of all American deaths, and about 15 percent of all deaths worldwide will be attributed to cancer."* – Siddhartha Mukherjee, *The Emperor of All Maladies: A Biography of Cancer*

I. Introduction

As a group of diseases, cancer has killed millions throughout time. The scourge of cancer is primarily centered in our minds in the 20th century. As our species lives longer and we have improved medical care, the prevalence of cancer has increased as more and more people are diagnosed and ultimately die of the disease. Throughout this time period, extensive research has been conducted to attempt to battle this scourge, to fight back against the inevitable, and to battle the very nature of our genome that causes the disease (Mukherjee, 2010; Lieberman, 2013). How and where have we succeeded?

Previous literature has examined this question from different viewpoints. Initial studies of cancer mortality show that while people diagnosed with cancer are surviving longer, they are still dying of cancer (Bailar and Smith, 1986; Bailar and Gornik, 1997; Welch at al., 2000). However, later research shows that increasing five-year survival rates contributes to decreasing cancer mortality (Lichtenberg, 2010). Analysis has also shown the contributions of drugs, treatments, and early detection to increasing survival and decreasing mortality (Sun et al., 2010; Lichtenberg, 2004; Lichtenberg, 2010). By comparing treatment costs to social benefits, further research has shown that this increase is economically beneficial (Lakdwalla et al., 2010). Other research has determined that it is possible for life expectancy of those diagnosed with certain cancers to equal that of the general population (Yin et al., 2012; Gamborti-Passerini et al., 2011). Additionally, significant research has shown the influence of race and socioeconomic status on diagnosis and survival (Amey et al, 1997; Bradley et al, 2001; Bradley et al 2002; Cross et al, 2002;

Meliker et al, 2009; Booth et al, 2010; Rauh-Hain et al, 2013). Research has also shown the influence of distance from treatment centers (Scoggins et al, 2001; Huang et al, 2009). Finally, research has shown that there is some evidence of geographic and spatial variation in cancer survival (Philipson et al, 2012; Wang et al, 2008).

Since there has already been significant research analyzing cancer, this research complements the existing research by focusing on breast cancer and examining two different facets of the disease. The two facets of the disease are the connection between cancer statistics—stage of diagnosis, survival, and mortality—and the geographic variation of these statistics. To address these questions, breast cancer diagnoses from specific locals in the United States from 1973-2004 are analyzed.

The focus is on breast cancer because it is common, high profile, and not tied to specific carcinogens. Stage of diagnosis, survival rates, and mortality are all analyzed because each alone does not provide a complete picture of cancer. Stage of diagnosis alone says nothing about survival and survival rates are biased without understanding stage of diagnosis—detecting cancer earlier but not improving life expectancy will still result in an increase in survival. Mortality, which is unbiased, shows the ultimate toll of cancer.

The first facet, the connection between these statistics, shows the relative importance of stage of diagnosis on survival, the relative importance of stage of diagnosis on mortality, and the relative importance of survival on mortality. Thus, it shows the validity of these statistics to capture the condition and to analyze it.

The second facet, the geographic variation of these statistics, shows the geographic variation of breast cancer as a whole by examining the geographic variation in late stage diagnosis, survival rates, and mortality rates. Various lifestyle factors are expected to

influence this geographic variation, and this analysis shows the effects of these factors on the different statistics of the disease.

Combined, these two facets show the connection between these statistics and the geographic variation of them. The connection between these statistics and their geographic variation provides a complete understanding of them, and this understanding translates into an understanding of breast cancer in the population.

II. Background

To fully understand populations with cancer, cancer statistics, stages of cancer diagnosis, and cancer treatments must be understood. There are three main cancer statistics: incidence, survival, and mortality. The incidence statistic is the rate of cancer diagnosis in the population (for a given year). It is reported as the number of diagnoses per 100,000 people. Survival is the proportion of people diagnosed with cancer who survive a given number of years (reported for a given year or given set of years). Two types of survival statistics are reported: absolute and relative. Absolute survival is the proportion of people alive after a certain number of years. Relative survival is adjusted for the general survival rate of the population. For this research, one-year absolute survival and five-year absolute survival rates are used. These rates report the proportion of patients alive one-year and five-years after diagnosis. The mortality statistic is the rate of cancer deaths in the population (for a given year). It is reported as the number of deaths per 100,000 people (SEER, 2010).

Methods of determining stage of cancer diagnosis have changed throughout the last thirty years. There are many different criteria for determining the stage at diagnosis. More recent diagnoses are classified on a more descriptive scale than was available for older

diagnoses. Because this research begins with diagnoses from the 1970s, a more historic categorization of stage of diagnosis will be used. On this scale, there are four different stages: in situ, localized, regional, and distant. Each stage has a specific definition referring to the extent of cancer propagation. The *in situ* stage refers to a "noninvasive neoplasm," which is a tumor that has not expanded past the epithelial tissue or base membrane. The *localized* stage refers to an "invasive neoplasm" that is completely "confined to the organ of origin." The *regional* stage refers to an "invasive neoplasm" that has expanded beyond the organ in which it originated into the nearby organs/tissues, the regional lymph nodes, or both. The *distant* stage refers to a neoplasm that has expanded well beyond the nearby organs/tissue or regional lymph nodes; it has expanded to remote areas of the body. Although the precise definitions of these stages are defined differently, depending on the source, these are the definitions from the data source used in this research (SEER, 2010).

Treatment options for cancer have evolved throughout the time period of this research. The standard techniques of surgery, radiation, and chemotherapy have been expanded beyond their original purview, as they have been combined and improved upon since their inception (Mukherjee, 2010). Each of these treatment techniques has its strengths and weaknesses. In the last 15 years, a new type of therapy—targeted therapy—has also been developed. These therapies directly attack the underlying cancer-causing genetic mutations and can extend life beyond the power of conventional treatments (Gamborti-Passerini et al., 2011; Hudis, 2007; Yin et al, 2012). The intricacies of treatment are not important for the purposes of this research. These treatments are relevant to this research because it is the change in treatment technology that drives survival and mortality changes and the change in diagnosis technology that drives stage of diagnosis changes.

While these treatments have been shown to improve cancer survival and mortality, this research addresses the relationship between the statistics and whether there is geographic variation to such changes—not the treatments or the technologies themselves.

III. Literature Review

There are multiple trends in the literature: analysis of changing cancer mortality, analysis of the causes of changes in cancer survival, evaluations of the costs and benefits of treatment, and analysis of the non-treatment factors that contribute to diagnosis, survival, and mortality.

Bailar and Smith (1986) and Bailar and Gornik (1997) both analyze changing incidence and mortality, utilizing data from the National Center for Health Statistics. They conclude that focusing "on improving treatment must be judged a qualified failure" (Bailar and Smith, 1986). They show the importance of incidence and conclude that focusing on treatment alone is unsuccessful.

Welch et al (2000) presents another pessimistic evaluation. This paper determines that, while five-year survival rates increased for all 20 tumor types analyzed, mortality only declined for 12 tumor types while increasing for the remaining 8 types. This paper utilizes data from the NCI's Surveillance, Epidemiology, and End Results Program (SEER) and calculates that increasing five-year survival rates are correlated with increasing incidence but not mortality. Lichtenberg (2010a) revisits Welch et al, concluding that there is a partial correlation between increasing five-year survival and decreasing mortality. Lichtenberg controls for incidence while calculating the effect of five-year survival on mortality, using SEER and Australian data.

Lichtenberg (2004) analyzes the effects of drugs on cancer mortality, determining that the "increase in the stock of drugs account[s] for about 50-60% of the increase in age-adjusted survival rates in the first 6 years after diagnosis" and increases the life expectance of people diagnosed with cancer by approximately a year. This paper uses SEER data and concludes that drugs have significantly extended cancer survival (Lichtenberg, 2004).

Lichtenberg (2010b) analyzes the changes in cancer mortality—using SEER data. This study concludes that, of the 17.2% decline in cancer mortality between 1991 and 2006, 7% is due to the decline in incidence, 27% is due to drug innovation, and 40% is due to imaging innovation resulting in better diagnosis. The model tests the effects of new drug procedures, advanced imaging procedures, and the age-adjusted incidence rate on the mortality rate to reach these conclusions (Lichtenberg, 2010b).

Sun et al (2010) analyzes the relative impacts of innovation on cancer survival. They utilize data from SEER for combined cancers, colorectal cancer, lung cancer, pancreatic cancer, breast cancer, and non-Hodgkin's lymphoma. Sun et al calculate the change in detection, change in cancer life expectancy, and the share of survival gains due to treatment. This calculation leads to the conclusion that improvements in treatment are primarily responsible for increases in survival (Sun et al, 2010).

Lakdawalla et al (2012) determine that the relative gains from the war on cancer are 23 million life-years and $1.9 trillion of social value for the years 1988-2000. This paper also determines that average life expectancy for cancer patients increased 3.9 years from 1988 to 2000. This work builds on Sun et al's 2010 paper and focuses on calculating the aggregate social value. The paper concludes that cancer spending has produced consumer surplus (Lakdawalla et al, 2010).

Research has also been conducted into the various factors that influence diagnosis and survival. Numerous papers have covered these topics in various settings, so a selected portion is presented here. The papers that analyze non-treatment factors are grouped into three categories of analysis: demographic factors, distance, and geographic variation.

It is clear that demographic factors have an influence on cancer survival and diagnosis. Previous research suggests that socioeconomic status, race, and location all influence stage of diagnosis and survival prospects. Amey et al (1997) shows that both race and location of residence influence the stage at which breast cancer is diagnosed, using Florida registry data from 1981 to 1989. Minorities and geographically remote people are more likely to be diagnosed at a later stage, a result consistent with previous research. This paper shows that the influence of race also interacts with location, so rural black women are diagnosed with more advanced breast cancer. However, this paper does not control for socioeconomic status, so the effect of it is captured by the location and race variables. Bradley et al (2002) demonstrates that, by controlling for socioeconomic status in addition to race, the effect that is attributed to race is actually attributable to socioeconomic status. This paper uses data from SEER and analyzes both stage of diagnosis and survival, concluding that race does not exert a statistically significant influence on "unfavorable breast cancer outcomes" but low socioeconomic status is "associated with late-stage breast cancer at diagnosis, type of treatment received, and death." Cross et al (2002) confirms the result that socioeconomic status is the explanatory variable, not race. Booth et al (2010) provides further evidence by examining a Canadian population. The paper shows that there are small differences in stage of diagnosis that are attributable to socioeconomic status. Additionally, this paper shows that socioeconomic status is associated with survival.

Bradley et al (2001) produces similar results, showing that there is a disparity between high and low socioeconomic status for stage of diagnosis and survival. Meliker et al (2009) show that the apparent influence of race on survival disappears in smaller geographic areas, suggesting that "modifiable societal factors are responsible for apparent racial disparities." This result is consistent with the literature showing the importance of socioeconomic status. Rauh-Hain et al (2013) shows that there is racial disparity, but it disappears after 1995. This paper did not control for socioeconomic status, but it did control for SEER registry. Combined, these papers show that race appears to influence cancer outcomes, but they also show that this effect is truly ascribable to socioeconomic status.

It is also clear that distance from a healthcare provider matters. Scoggins et al (2011) shows that later stage diagnoses are "associated with travel burden." However, conversely, the paper does not present evidence that driving time is associated with later stage diagnoses or worse treatment. This suggests that some of these effects could be due to the population characteristics of the rural communities, as the regressions do not control for socioeconomic status (they do control for race). Huang et al (2009) control for socioeconomic status, race, and age and show that "longer travel distance also adversely affects early detection for rural population." This finding confirms the suggestion of the previous research. The distance from treatment centers appears to decrease the likelihood of early stage diagnosis and good treatment.

Geographic and spatial variation also exists. Philipson et al (2012) studies the differences between cancer survival gains (and their value) in the US and Europe from 1983 to 1999. The study determines that survival is higher in the United States and that,

while the cost of care in the United States is higher, the social value of the care is higher as well. Wang et al (2008) investigates the spatial variation of late-stage breast cancer diagnosis in Illinois. The paper shows that "poor geographical access to primary health care significantly increases the risk of late diagnosis." This result confirms the earlier results analyzing distance and rural populations and shows that there is variation within states inside the US.

These papers show the changing cancer landscape over time and the connection between the statistics. The disparity in the findings provides an opportunity for this research to verify the true relationship. Additionally, this literature shows that many factors influence cancer diagnosis and survival. The literature also indicates that there is geographic variation. The variables tested in the literature are controlled for in this research, and the question of geographic variation is addressed.

IV. Theoretical Framework

The theoretical basis for the first facet of this analysis lies in the assumption of connections between stage of diagnosis, survival (one-year and five-year), and mortality. Later stages of diagnosis are expected to decrease survival likelihood and increase mortality likelihood. This is because later stage diagnoses are tumors that have progressed further and are thus more dangerous—they are more extensive, larger, and have already exacted a greater toll upon the body. Additionally, tumors that are only diagnosed later could be more virulent or occur in a population of people that both are more likely to be diagnosed later and are less likely to follow through with treatment options. Survival is expected to decrease the likelihood of mortality. This expectation is drawn from the literature. If treatments, which are measured as they improve survival, have a lasting

biological effect, then they should reduce the likelihood of morality as well. Thus, these assumptions expect these statistics to be linked.

The theoretical basis for the second facet this analysis lies in the assumption that there will be geographic variation in stage of diagnosis, survival, and mortality. The basis for this assumption is due to two underlying premises—variation in populations, and variation in the distribution of medical talent and facilities throughout the US.

Populations across the US are dissimilar. Some areas have clear demographic differences, such as the percentage of the population that is of Hispanic origin, and some have less clear differences, such as lifestyles. These differences will lead to differences in survival and stage of diagnosis because lifestyles influence medical outcomes. There are clear examples, such as exercise and diet, and there are less clear examples, such as how individuals value health and how often individuals see a doctor. These factors could conceivably alter the health of populations and thus alter medical outcomes. In this research, geographic differences in lifestyle will alter the stage of diagnosis and efficacy of treatment since those who see doctors will likely be diagnosed earlier and those who adhere to treatment regimes will be more likely to survive, as shown in one-year or five-years. Additionally, this effect is expected in mortality as well. Those who follow treatment instructions should be less likely to die of cancer if the treatments work. Although lifestyle presents a valid mechanism for geographic variation in cancer, this effect is expected to be secondary to that of medical talent distribution.

Medical talent and facility technology are not likely to be evenly distributed throughout the US—both the skill of practitioners and the tools available to practitioners will likely cluster in high profile hospitals and urban centers (such as specially designated

comprehensive cancer care centers that are specifically designed to cater to cancer patients). If medical talent is unevenly distributed throughout the US then one would expect geographic variation in stage of diagnosis, survival rates, and mortality. Specifically, those patients near these centers should have better outcomes. Patients in those areas with better practitioners should undergo earlier detection, survive longer (be more likely to survive one-year or five-years), and have lower mortality rates (be less likely to die of breast cancer.

Since these two premises will cause geographic variation, it is not possible to disaggregate the individual effects of variation in lifestyle and medical talent/facilities—while these variations are likely distributed differently, those patterns are not discernable here. Population demographics will be controlled to the greatest extent possible. Thus geographic variation will be attributed to medical talent and lifestyle considerations independent of demographic considerations.

V. Data

The data source for this analysis will be the Surveillance, Epidemiology, and End Results Program (SEER) of the National Cancer Institute, a division of the National Institutes of Health. The SEER data are collected from registries throughout the US and contain information about individual tumor cases. Depending on the registry, there are data available from 1973-2010. [2]

The SEER dataset contains many variables, a subset of which is used for this analysis. The included variables are registry ID, race, Hispanic origin, sex, marital status,

[2] To analyze the greatest number of years possible, data beginning in 1973 is used, this limits the data to nine registries: Connecticut, Detroit, San Francisco-Oakland, Hawaii, Iowa, New Mexico, Utah, Atlanta, and Seattle-Puget Sound.

age at diagnosis, year of diagnosis, historic stage, radiation, and survival months. Each of these variables is reported as a discontinuous variable and is modified into dummy variables for the analysis. The registry ID variable is converted into nine dummy variables, one for each registry. The race variable is converted into nine dummy variables—white, black, American Indian, Chinese, Japanese, Filipino, Hawaiian, other race, and unknown race. The Hispanic origin variable is converted into three dummy variables—Hispanic, not Hispanic, and unknown Hispanic origin. The marital status variable is converted into seven dummy variables—single, married, separated, divorced, widowed, unmarried, and unknown status. The age of diagnosis variable is included as reported. The year of diagnosis variable is converted into a vector from 0-37, with a value of 0 indicating diagnosis in 1973, a value of 1 indicating diagnosis in 1974, etc.[3] The historic stage variable is converted into five dummy variables—I (in situ), II (localized), III (regional), IV (distant), and unstaged.[4] The radiation variable is converted into four variables—radiation, no radiation, unknown radiation, and refused radiation. Patient data for which the race, marital status, Hispanic origin, stage of diagnosis, or radiation treatment is unknown is excluded.[5] Additionally, only female patients are considered. The survival months variable is converted into two different survival rate dummy variables—five-year survival and one-year survival. Each of these dummy variables is positive if the patient survived 60 (12) or more months.

---

[3] 2004 is the last year that allows for calculation for five-year survival rates, so data from the years 2005-2010 is excluded.

[4] The SEER Program refers to this variable as historic stage because other stage variables are reported for various years of the dataset. For this analysis, it represents Stage of Diagnosis, as referred to in the Background section.

[5] Patients who are listed as other race are also excluded. Patients who refused radiation are also excluded, as those who refuse radiation are not a random group.

There are several limitations to this dataset. There are no socioeconomic status data included and there are limited treatment data available. The closest data to socioeconomic status in the data are the counties of residence, so models using county fixed effects are included in the analysis. However, models without county fixed effects are run as well in order to see the variation both with and without county level controls. Lack of socioeconomic data will present some limitations in interpretations of the data—socioeconomic status will bias certain race, marital status, and registry variables. Still, this will not corrupt the analysis, as the purpose is to identify the geographic variation, which will still appear. There are limited treatment data available in this dataset, namely there are only radiation treatment data and no surgery or chemotherapy treatment data. This will not cause problems because this analysis does not concern individual treatments or the success of individual treatments. However, this will lower the R-squared values for the survival and mortality models, since they will not include variation due to specific treatments.

Despite these limitations, this dataset is the best option for analyzing long-term cancer trends. It is extensive and covers a broad swath of the US. There are over 650,000 individual cases in the breast cancer dataset. Additionally, this dataset covers a wide range of years with consistent data reporting and variables. Significant research has used this dataset because of these strengths. The SEER dataset provides the best available option for analyzing cancer throughout this time period.[6]

VI. Empirical Specification

[6] An extended discussion of the limitations of this dataset, along with summary statistics for the data, can be found in the Appendix.

Multiple regressions models, analyzing the two facets of this research, are run in two groups. The first group analyzes the first facet: the connection between stage of diagnosis, survival (both one-year and five-year), and mortality (death due to breast cancer). The second group analyzes second facet: the geographic distributions of the stage of diagnosis, survival, and mortality.

In the first group, three sets of regressions encapsulate the different relationships: effect of stage of diagnosis on survival, effect of stage of diagnosis on mortality, and effect of survival on mortality. Within each set, two different regressions are run: one without county-level fixed effects and one with county-level fixed effects.[7] In the regressions involving survival, one-year survival and five-year survival are addressed separately. The equation for the first regression of this group follows. All the regression equations in this group are similar to the following equation. They differ by the explanatory variable of interest and the dependent variable—both of which are highlighted below.[8]

$$\boldsymbol{Five\ Year\ Survival}$$

$$= \alpha + \beta_1(\boldsymbol{Stage\ of\ Diagnosis}) + \beta_2(Race\ Dummies)$$
$$+ \beta_3(Marital\ Status\ Dummies) + \beta_4(Hispanic\ Dummies)$$
$$+ \beta_5(Age\ at\ Diagnosis) + \beta_6(Age\ Squared) + \beta_7(Year\ Vector) + \varepsilon$$

In the second group, three sets of regressions test the geographic distribution of each statistic: stage of diagnosis, survival (one-year and five-year), and mortality. To test geographic distribution, the dummy variables that represent the registry the case is from are used to represent geographic areas. Like in the first group, each regression is run with

[7] These county-level fixed effects apply to the county of residence of the patient.
[8] Although only one regression equation is reported here, a complete record of the equations used can be found in the Appendix.

and without county fixed effects.[9] Within each set of regressions, two types of regressions are reported: variation without race and marital controls and variation with race and marital controls along with interaction terms included, in order to observe variation with and without the explanatory power of the controls. This means that four regressions are reported for stage of diagnosis, one-year survival, five-year survival, and death due to breast cancer (mortality). The stage of diagnosis variable used here is *Late Stage,* which includes stages III and IV. All the regression equations in this group are similar to the following equation.[10] The differences between the sets of regressions concern the dependent variable—which is highlighted below.[11]

$$\boldsymbol{Late\ Stage\ Diagnosis}$$
$$= \alpha + \beta_1(Registry\ Dummies) + \beta_2(Race\ Dummies)$$
$$+ \beta_3(Marital\ Status\ Dummies) + \beta_4(Hispanic\ Dummies)$$
$$+ \beta_5(Age\ at\ Diagnosis) + \beta_6(Age\ Squared) + \beta_7(Year\ Vector)$$
$$+ \beta_8(Stage\ of\ Diagnosis\ Dummies) + \beta_9(Radiation)$$
$$+ \beta_{10}(Age\ x\ Stage\ Interactions) + \beta_{11}(Age\ x\ Stage\ x\ Year\ Interactions)$$
$$+ \varepsilon$$

All of these models contain a modified variable—*Age Squared.* This variable is the age of the patient at the time of diagnosis squared. It is included because the age of the patient is expected to have a non-linear effect on stage of diagnosis, survival, and mortality. The models analyzing the geographic distribution of one-year survival, five-year survival,

[9] One county from each registry is excluded in the regressions, a complete discussion can be found in the Appendix.
[10] The interaction terms are not included in the stage of diagnosis regression because they include the effects of stage of diagnosis.
[11] Although only one regression equation is reported here, a complete discussion of the equations used can be found in the Appendix.

and mortality all include interaction terms because the effects of these variables are expected to be compounded through their interaction. These interaction variables are divided into two categories: the (Age at Diagnosis)x(Stage of Diagnosis) interactions and the (Year of Diagnosis)x(Age at Diagnosis)x(Stage of Diagnosis) interactions. The (Age at Diagnosis)x(Stage of Diagnosis) interactions are created by multiplying the age of diagnosis with each stage dummy variable (I, II, III, and IV), resulting in four interaction variables: (Age at Diagnosis)x(Stage I), (Age at Diagnosis)x(Stage II), (Age at Diagnosis)x(Stage III), and (Age at Diagnosis)x(Stage IV). The (Year of Diagnosis)x(Age at Diagnosis)x(Stage of Diagnosis) interaction variables are created by multiply the (Age at Diagnosis)x(Stage of Diagnosis) variables by the year of diagnosis vector, resulting in four interaction variables: (Year of Diagnosis)x(Age at Diagnosis)x(Stage I), (Year of Diagnosis)x(Age at Diagnosis)x(Stage II), (Year of Diagnosis)x(Age at Diagnosis)x(Stage III), and (Year of Diagnosis)x(Age at Diagnosis)x(Stage IV). These terms account for the interaction effects of age at diagnosis and stage of diagnosis and the interaction effects of age of diagnosis, stage of diagnosis, and year of diagnosis. In all models the interactions involving stage I are excluded to avoid perfect multicollinearity.

VII. Findings

The results from the models are reported in the following tables. Although each regression contains many controls, the data presented here highlight the key variables from each regression. A complete discussion of the controls and the complete results tables can be found in Appendix IV. Because the dataset is composed of individual cases, the mortality regressions are reported with *BC Death* as the independent variable. This refers to the likelihood of death due to breast cancer, conditional upon diagnosis.

Table 1: Stage of Diagnosis and Survival

VARIABLES	(1) FiveYears	(2) FiveYears	(3) OneYear	(4) OneYear
Stage II	0.00537***	0.00511***	0.0217***	0.0216***
	(0.00157)	(0.00157)	(0.000860)	(0.000861)
Stage III	-0.155***	-0.154***	-0.00555***	-0.00539***
	(0.00173)	(0.00173)	(0.000945)	(0.000946)
Stage IV	-0.617***	-0.615***	-0.330***	-0.329***
	(0.00272)	(0.00272)	(0.00149)	(0.00149)
Fixed Effects	No	Yes	No	Yes
Constant	0.121***	0.122***	0.757***	0.756***
	(0.00875)	(0.00907)	(0.00479)	(0.00497)
Observations	502,467	502,467	502,467	502,467
R-squared	0.213	0.214	0.154	0.155

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1 shows the effect that stage of diagnosis has on survival. The stage of diagnosis clearly has a significant effect on the survival of the patient. Stage IV diagnosis decreases the probability of five-year survival by over 60%. Stage II has a small positive effect on survival. Stage III decreases the probability of survival by about 15%. Although the effects are less dramatic for stages II and III on one-year survival, stage IV still has a large effect, decreasing the probability of one-year survival by over 30%. The differences between the models with and without county-level fixed effects are minor.

Table 2: Stage of Diagnosis and Mortality

VARIABLES	(1) BC Death	(2) BC Death
Stage II	0.0122***	0.0123***
	(0.00155)	(0.00156)
Stage III	0.232***	0.232***
	(0.00171)	(0.00171)
Stage IV	0.621***	0.620***
	(0.00269)	(0.00269)
Fixed Effects	No	Yes
Constant	0.460***	0.453***
	(0.00869)	(0.00900)

Observations	502,467	502,467
R-squared	0.196	0.197

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2 shows that stage of diagnosis, in addition to altering survival prospects, alters the likelihood that a patient dies due to breast cancer. The constant term shows that patients have a high likelihood of death due to breast cancer (over 45% chance without including the effect of the patient's age), but the stage terms show that as the stage of diagnosis increases the likelihood of death due to breast cancer increases as well. Stage II increases the likelihood by about 1%, stage III by about 23%, and stage IV by about 62%. The values of the coefficients are not qualitatively different when fixed effects are included.

Table 3: Survival and Mortality

VARIABLES	(1) BC Death	(2) BC Death	(3) BC Death	(4) BC Death
Five Year Survival	-0.455***	-0.455***		
	(0.00126)	(0.00126)		
One Year Survival			-0.328***	-0.327***
			(0.00256)	(0.00256)
Fixed Effects	No	Yes	No	Yes
Constant	0.629***	0.623***	0.874***	0.864***
	(0.00826)	(0.00857)	(0.00932)	(0.00965)
Observations	502,467	502,467	502,467	502,467
R-squared	0.252	0.253	0.088	0.090

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3 demonstrates a clear connection between survival and mortality. BC Death indicates a death due to breast cancer—this is mortality. While neither one-year survival nor five-year survival perfectly track mortality, there is a significant decrease in the likelihood of death due to breast cancer with patients who survive one-year or five-years

after diagnosis. One-year survival reduces the likelihood of death due to breast cancer by about 33%, and five-year survival reduces the likelihood of death due to breast cancer by about 46%. Like the previous regressions, these results are not significantly different when comparing the models without county fixed effects to those with them.

Table 4: Geographic Variation in Tumor Stage Diagnosis

VARIABLES	(1) Late Stage	(2) Late Stage	(3) Late Stage	(4) Late Stage
Atlanta	-0.00199	0.0178*	-0.0111***	0.00990
	(0.00286)	(0.00984)	(0.00291)	(0.00984)
Detroit	0.0166***	0.0166***	0.0114***	0.0253***
	(0.00227)	(0.00484)	(0.00231)	(0.00484)
San Francisco	-0.00949***	0.00521	-0.0105***	0.000592
	(0.00229)	(0.00429)	(0.00234)	(0.00431)
Hawaii	-0.0434***	-0.0121	-0.0304***	-0.00657
	(0.00364)	(0.0177)	(0.00477)	(0.0180)
Iowa	0.000802	0.0249	0.00958***	0.0364
	(0.00245)	(0.0348)	(0.00248)	(0.0348)
New Mexico	0.0177***	0.000835	0.0139***	-0.00109
	(0.00336)	(0.00572)	(0.00349)	(0.00577)
Seattle	-0.0105***	-0.0369***	-0.00352	-0.0273**
	(0.00238)	(0.0119)	(0.00240)	(0.0119)
Utah	0.0163***	-0.00311	0.0250***	0.00776
	(0.00340)	(0.0554)	(0.00341)	(0.0554)
Race/Marital Controls	No	No	Yes	Yes
Fixed Effects	No	Yes	No	Yes
Constant[12]	0.722***	0.726***	0.670***	0.675***
	(0.0106)	(0.0109)	(0.0107)	(0.0110)
Observations	502,467	502,467	502,467	502,467
R-squared	0.023	0.025	0.026	0.027

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 4 shows that there is clear geographic variation in the stage of diagnosis of breast cancer. While not every registry is significant, this indicates that not all registries as

[12] The constant includes the effects of a married, white woman from Connecticut.

substantively different from the Connecticut registry.[13] Hawaii and Seattle have statistically significantly lower rates of later diagnoses (conditional upon diagnosis) than Connecticut, while New Mexico, Utah, and Detroit have higher rates. There is significant variation between the models that do not include the marital and race controls and those that do, and there is significant variation between the models that include county fixed effects and those that do not. It is important to note here that the R-squared values for all four regressions presented here are small, showing that these models geographic variation do not account for a significant amount of the variation in the stage of breast cancer diagnosis.

Table 5: Geographic Variation in One-Year Survival

VARIABLES	(1) OneYear	(2) OneYear	(3) OneYear	(4) OneYear
Atlanta	0.000749	0.00178	0.00438***	0.00527
	(0.00128)	(0.00439)	(0.00129)	(0.00437)
Detroit	-0.00815***	-0.000167	-0.00553***	-0.00204
	(0.00101)	(0.00216)	(0.00103)	(0.00215)
San Francisco	0.00285***	0.00136	0.00333***	0.00292
	(0.00102)	(0.00191)	(0.00104)	(0.00192)
Hawaii	0.00440***	-0.0105	0.00338	-0.00889
	(0.00162)	(0.00791)	(0.00212)	(0.00798)
Iowa	0.00426***	-0.00298	0.00201*	-0.00621
	(0.00109)	(0.0155)	(0.00110)	(0.0154)
New Mexico	0.000400	0.00792***	-0.000199	0.00744***
	(0.00150)	(0.00255)	(0.00155)	(0.00257)
Seattle	0.00668***	0.00913*	0.00477***	0.00592
	(0.00106)	(0.00530)	(0.00107)	(0.00528)
Utah	-0.000611	0.0178	-0.00345**	0.0125
	(0.00151)	(0.0247)	(0.00151)	(0.0246)
Marital/Race Controls	No	No	Yes	Yes
Fixed Effects	No	Yes	No	Yes
Interactions	No	No	Yes	Yes
Constant[14]	0.738***	0.736***	0.806***	0.803***
	(0.00478)	(0.00493)	(0.00585)	(0.00597)
Observations	502,467	502,467	502,467	502,467

[13] To avoid perfect multicollinearity, the Connecticut variable is excluded from the regression (and all geographic variation regressions). Thus, the constant includes the effect of Connecticut and all other registry variables report differences relative to Connecticut.
[14] The constant includes the effects of a married, white woman from Connecticut.

R-squared	0.153	0.153	0.160	0.161

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The four regressions in Table 5 show the geographic variation in one-year survival. The first regression shows that there is variation due to geography, without controlling for race or marital status—but the differences are reported as less than 1% in every case. The second regression shows that, when controlling for county-level fixed effects, only New Mexico is statistically significantly different from Connecticut.—but its difference is less than 1% The third regression shows that, when race and marital status controls are added with the interaction terms, there is almost no geographic variation. All of the differences are less than 1%. The fourth regression, when including county-level fixed effects, shows that only New Mexico is statistically significantly different from Connecticut and this effect is, yet again, less than 1%.

Table 6: Geographic Variation in Five-Year Survival

VARIABLES	(1) FiveYears	(2) FiveYears	(3) FiveYears	(4) FiveYears
Atlanta	-0.00336	-0.0335***	0.0108***	-0.0203**
	(0.00233)	(0.00802)	(0.00237)	(0.00800)
Detroit	-0.0205***	-0.0108***	-0.00999***	-0.0155***
	(0.00185)	(0.00394)	(0.00188)	(0.00394)
San Francisco	0.0127***	0.00204	0.0169***	0.0108***
	(0.00187)	(0.00350)	(0.00190)	(0.00351)
Hawaii	0.0275***	-0.0344**	0.0173***	-0.0381***
	(0.00297)	(0.0145)	(0.00388)	(0.0146)
Iowa	0.0125***	0.0128	0.00700***	0.00334
	(0.00200)	(0.0284)	(0.00201)	(0.0283)
New Mexico	-0.00485*	0.0148***	-0.00349	0.0157***
	(0.00274)	(0.00467)	(0.00284)	(0.00469)
Seattle	0.0219***	0.00663	0.0183***	0.000723
	(0.00195)	(0.00969)	(0.00196)	(0.00965)
Utah	0.00573**	0.00618	-0.000413	-0.00235
	(0.00277)	(0.0452)	(0.00277)	(0.0450)
Marital/Race Controls	No	No	Yes	Yes
Fixed Effects	No	Yes	No	Yes

Interactions	No	No	Yes	Yes
Constant[15]	0.0637***	0.0624***	0.201***	0.199***
	(0.00874)	(0.00902)	(0.0107)	(0.0109)
Observations	502,467	502,467	502,467	502,467
R-squared	0.209	0.210	0.215	0.216

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 6 shows the geographic variation in five-year survival and the four regressions combine to show the different levels of variation. The first regression shows the variation in five-year survival without race or marital status controls, and the second regression shows this variation with county-level fixed effects. In these regressions, it is apparent that there is geographic variation both with and without county controls. Unlike one-year survival, these differences are palpable. Additionally, while these effects change while accounting for county-level fixed effects, they do not disappear. The third regression shows the variation while controlling for race and marital status, and the fourth regression shows similar variation with county-level fixed effects. These two regressions show five-year survival variation exists independent of the controls. Like the previous two regressions, there is variation both with and without county-level fixed effects.

Table 7: Geographic Variation in Mortality

VARIABLES	(1) BC Death	(2) BC Death	(3) BC Death	(4) BC Death
Atlanta	0.00286	0.0324***	-0.00717***	0.0222***
	(0.00232)	(0.00796)	(0.00235)	(0.00794)
Detroit	0.0179***	0.0264***	0.00929***	0.0277***
	(0.00183)	(0.00391)	(0.00187)	(0.00390)
San Francisco	-0.0105***	0.00483	-0.0148***	-0.00261
	(0.00185)	(0.00347)	(0.00189)	(0.00348)
Hawaii	-0.0306***	0.00387	-0.0222***	0.00967
	(0.00294)	(0.0143)	(0.00385)	(0.0145)
Iowa	0.00714***	0.0640**	0.00979***	0.0704**

[15] The constant includes the effects of a married, white woman from Connecticut.

	(0.00198)	(0.0281)	(0.00200)	(0.0280)
New Mexico	0.0166***	0.0114**	0.0158***	0.0110**
	(0.00272)	(0.00463)	(0.00281)	(0.00465)
Seattle	-0.0138***	-0.000280	-0.0124***	0.00157
	(0.00193)	(0.00961)	(0.00194)	(0.00957)
Utah	0.00973***	0.0436	0.0120***	0.0489
	(0.00275)	(0.0448)	(0.00275)	(0.0446)
Marital/Race Controls	No	No	Yes	Yes
Fixed Effects	No	Yes	No	Yes
Interactions	No	No	Yes	Yes
Constant[16]	0.485***	0.480***	0.324***	0.319***
	(0.00867)	(0.00894)	-0.0106	-0.0108
Observations	502,467	502,467	502,467	502,467
R-squared	0.195	0.196	0.202	0.203

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 7 shows the geographic variation in breast cancer mortality.[17] The first and second regressions show the distribution without including race or marital status controls or interactions. In these regressions, it is evident that there is variation both with and without county-level fixed effects. The third and fourth regressions show the distribution with these controls and interactions included. Similarly to the first two regressions, they show that variation exists both with and without controlling for county-level effects. These regressions demonstrate the geographic distribution in breast cancer mortality.

In addition to these regressions analyzing mortality, maps showing the variation in the proportion of cases that are diagnosed at a late stage, the proportion of cases that survive one year, the proportion of cases that survive five years, and the proportion of cases that die due to breast cancer have been tabulated. These maps are included in the appendix.

[16] The constant includes the effects of a married, white woman from Connecticut.
[17] In this probit model, that is the probability of death caused by breast cancer, conditional upon breast cancer diagnosis.

The findings presented here show the relationship between stage of diagnosis, survival, and mortality. They also show the geographic distribution of late stage diagnosis, one-year survival, five-year survival, and mortality.

VIII. Discussion

These regressions produce interesting results. The first group shows the connection between the various cancer statistics that define this disease, while the second group shows the geographic variation of those statistics within the United States.

The results of the first group show that stage of diagnosis impacts both survival and mortality. The link between later stages of diagnosis and a higher likelihood of death due to breast cancer seems to be self-evident. But, it is not always clear that earlier detection is better. This evidence shows that throughout the years surveyed here, patients with earlier stage diagnoses have been more likely to live longer and less likely to die of breast cancer. This is an important fact, and it shows that, throughout time, the stage of diagnosis has a palpable influence upon the likelihood of death due to breast cancer. The first implication of this result is that earlier detection is better. However, there are some limitations in the inferences that can be drawn due to the nature of this analysis. This evidence simply shows that those tumors diagnosed at earlier stages are less likely to lead to death than those diagnosed at later stages—the reason for this effect cannot be determined. It is important to note that this reason could be due to the biological nature of those tumors or population characteristics. However, previous literature suggests that diagnosing tumors earlier results in longer survival and a lower likelihood of death due to breast cancer (Sun et al, 2010; Lichtenberg, 2010b; Lichtenberg 2004). These results mean that early detection is imperative to improved breast cancer survival and reduced mortality. However, these

results do not clearly indicate that more efforts should be made to increase earlier detection presently. While that may be the case, this research simply shows that those tumors that are diagnosed earlier are less likely to result in death.

The results also indicate that survival is linked to stage of diagnosis. Late stage diagnoses are less likely to survive one-year and less likely to survive five-years, relative to earlier stage diagnoses. This result is in line with the reduced likelihood of breast cancer mortality. If those diagnosed with breast cancer are less likely to die of breast cancer than they are also more likely to survive. The survival aspect simply confirms the mortality analysis and produces the same results, with the same issues. This data shows the nature of these tumor cases, and while this suggests that increasing early detection would increase survival and decrease mortality, these suppositions must be qualified by noting that this data is descriptive.

The results also show that one-year survival and five-year survival both have a negative impact on the likelihood of breast cancer death. That is, those cases that live longer—whether it be one-year five-years—are less likely to die due to breast cancer than those who do not live that long (there are obvious differences in the effect of the survival statistic and mortality, with five-year survival having a larger effect). Various literature sources (Bailar and Gornik, 1986; Bailar and Smith, 1997; Welch et al, 2000) have questioned the link between survival and mortality in the past. But this evidence, which encompasses some of the time periods, shows that these claims, while true in the past, are no longer true. Those patients that survive are less likely to die of breast cancer. This statement seems almost definitional, but it is important to delineate the distinction between the two. These rates encompass the probability of death for a limited time after

diagnosis, while mortality encompasses the total probability of death by breast cancer, conditional on diagnosis. The literature on cancer research uses five-year survival to quantify benefits of treatments, and, here, it is evident that there is validity, at the population level, to using survival to proxy for mortality because this connection is large for both one-year survival and five-year survival: -0.33 and -0.36, respectively. Thus, this evidence shows that there is clearly a connection between survival and death due to breast cancer. Those cases that survive longer are less likely to die due to breast cancer.

There is little variation, if any, between the regressions with and without county-level fixed effects in this group. This shows that the difference between counties does not change the relationships between stage of diagnosis, survival, and mortality. While there is variation between counties, the connection between these statistics is independent of this variation.

The results from the first group show the importance of stage of diagnosis to survival and mortality, and they show the link between survival and decreased mortality. In showing these links, these results illuminate the evidence that stage of diagnosis, survival, and mortality are intimately linked.

The results of the second group show that late stage diagnosis, one-year survival, five-year survival, and probability of breast cancer mortality all vary geographically.[18] These results are important because the results from the first group of regressions show the importance of each of these statistics.

The geographic variation of late stage diagnosis produces the most limited results of this group. These regressions all have low R-squared values, meaning that very little of the

[18] All statistics are conditional upon diagnosis with breast cancer

variation in stage of diagnosis is explained by them. This means that, while there is variation in the likelihood of late stage diagnosis across the registries, these effects are not important when compared to other factors that influence stage of diagnosis. The most important result from this set of regressions is not the specific registry variations, but that there are variations that account for a small proportion of the variation. The high constants (all are above 0.67) in these regressions show that there is a high probability of late stage diagnosis. The variation due to geographic location is highest for Hawaii (-0.04, without controls; -0.03 with controls). So, while there is variation, geographic differences are not the driving factor. In fact, when controlling for county-level fixed effects and controls only Detroit and Seattle are statistically significantly different from Connecticut. This suggests that socioeconomic status and related lifestyle variables account for much of the variation in stage of diagnosis, but the controls and county-level fixed effects capture these effects. While there is variation with no controls, only county-level fixed effects, and only controls are used, these effects almost all disappear in the final regression. The final result, combined with the low R-squared values for all of the regressions, suggests that there are limits to the explanatory power of the geographic variation. The limitations of the explanatory power limits the interpretations of the results—no substantive conclusions can be made about the geographic variation in lifestyles or medical talent/facilities. Stage of diagnosis appears to be dependent, primarily, on factors not included here.

The geographic variation in one-year and five-year survival presents more interesting results. These regressions have higher R-squared values (about 0.15 and about 0.21, respectively), which shows that the variation captured is important. For one-year survival, most of the statistically significant variation disappeared when including county-

level fixed effects. This phenomenon suggests that much of the geographic variation is county-specific, which could include socioeconomic factors that are independent of marital status and race. Additionally, even though there is variation in these regressions, the sizes of the coefficients are small, all are reported as less than 0.01. This suggests that the effect of geographic variation is small. The explanatory power of these regressions is due more to the controls than the registry variables. This result is not unexpected, one-year survival, as shown in the first group of regressions, is highly dependent upon stage of diagnosis. It is likely that the biological nature of the tumors is primarily important for this survival statistic. For five-year survival, the statistically significant results remained when controlling for marital status and race and adding county-level fixed effects. The R-squared values for the five-year survival regressions were higher as well. Five-year survival varies geographically more than one-year survival, as expected. Five-year survival is likely affected more by lifestyle factors because treatment is dependent upon lifestyle factors and survival is dependent upon treatment. When including all controls, the regressions with and without county level fixed-effects differ, suggesting that some spatial variation is geographic (large scale/registry dependent) and some is at the county-level. This evidence supports the supposition that geographic variation in lifestyle, medical talent, or some other factor influences the likelihood of survival.

The geographic variation in mortality also suggests variation in lifestyle and medical talent. However, it is expected that the medical talent factor dominates here—while lifestyle can influence health in many ways, it should not influence the ultimate outcome (death due to breast cancer) unless it is affecting the ability of a patient to pay for procedures. Lifestyle has a limited effect on the ultimate success of treatment, even strict

adherence to a treatment protocol can result in death. Medical talent, however, better explains the differences in mortality. These results show that the geographic distribution in mortality is generally small. But, when controlling for all factors and interactions, the registries with statistically significant differences in mortality—Atlanta, Detroit, Iowa, and New Mexico—all have higher mortality rates than Connecticut, while San Francisco, Hawaii, Seattle, and Utah are all insignificantly different from Connecticut. Iowa had the largest difference (0.07), suggesting a large difference between Iowa and Connecticut. This data suggests that medical talent clusters in the latter five as opposed to the former four, and this suggestion holds up qualitatively. San Francisco, Connecticut, and Seattle all have highly rated research hospitals and cancer centers. Hawaii and Utah contain unique populations, showing that population lifestyles or other characteristics could have an influence. This does not suggest that urbanization has an effect, but it does suggest that there are differences between the two groups that have palpable effects. One conclusion is due to the distribution in medical talent and quality of treatment. These suppositions from the data are not clear. It is clear that there are geographic differences in mortality, some are explained by the race and marital status controls, some are explained by the county fixed effects, and some are not explained by either.

Combined, these geographic results show the influence of lifestyle and medical talent variation. The one-year survival analysis produces the weakest results because these factors are the least important here, since one-year survival is dominated by tumor biology. Five-year survival and mortality have significant variation. these factors are influenced more by lifestyle and medical talent, and this result shows lifestyle and medical talent distribution.

These geographic regressions combine to show that breast cancer varies in diagnosis, survival, and mortality throughout the United States. While the variation in stage of diagnosis does not seem important, the variation in survival and mortality are clearly important. The connections shown between these statistics indicate that any variation is important, the best standard of care, if it improves diagnosis or survival, will improve mortality and save lives. The maps in the appendix visually demonstrate geographic variation.

IX. Conclusion

The connections between these statistics show that stage of diagnosis, survival, and death are intimately linked. It is clear that the stage of diagnosis has a measurable influence on the likelihood of five-year survival, one-year survival, or mortality. It is also clear that those patients that survive longer are less likely to die due to breast cancer.

The geographic variation in these statistics suggests that the practical effects of medical treatment vary throughout the United States. Some of this variation is undoubtedly due to lifestyle factors and the uneven distribution of medical talent. It is evident that while all these factors vary, five-year survival and mortality have greater variation than one-year survival, while geographic variation explains little of the variation in stage of diagnosis.

Taken together, these results show the linkage between and geographic distribution of diagnosis, survival, and mortality. They show that all three statistics are important to understanding cancer and that they vary throughout the United States.

X. Works Cited

Amey, Cheryl H., Michael K. Miller, and Stan L. Albrecht. "The Role of Race and Residence in Determining Stage at Diagnosis of Breast Cancer," *The Journal of Rural Health* 13 (1997): 99-108. onlinelibrary.wiley.com.

Bailar III, John C., and Elaine M. Smith. "Progress Against Cancer?" *The New England Journal Of Medicine* 314 (1986):1226-1232. http://www.nejm.org.

Bailar III, John C., and Heather L. Gornik. "Cancer Undefeated," *The New England Journal Of Medicine* 336 (1997): 1569-1574. http://www.nejm.org.

Booth, Christopher M., Gavin Li, Jina Zhang-Salomons, and William J. Mackillop. "The Impact of Socioeconomic Status on Stage of Cancer at Diagnosis and Survival," *Cancer* 116 (2010):4160-7. doi: 10.1002/cncr.25427.

Bradley, Cathy J., Charles W. Given, and Caralee Roberts. "Disparities in Cancer Diagnosis and Survival," *Cancer* 91 (2001):178-88. www.cancer.org.

Bradley, Cathy J., Charles W. Given, and Caralee Roberts. "Race, Socioeconomic Status, and Breast Cancer Treatment and Survival," *Journal of the National Cancer Institute* 94 (2002):490-6. jnci.oxfordjournals.org.

Cross, Chaundre K., Jay Harris, and Abram Recht. "Race, Socioeconomic Status, and Breast Carcinoma in the U.S.: What Have We Learned from Clinical Studies?" *Cancer* 95 (2002):1988-99. doi:10.1002/cncr.10830.

Gambacorti-Passerini, Carlo, Laura Antolini, Francois-Xavier Mahon, Francois Guilhot, Michael Deiniger, Carmen Fava, Arnon Nagler, Chiara Maria Della Casa, Enrica Morra, Elisabetta Abruzzese, Anna D'Emilio, Sarit Assouoline, Muheez A. Durosinmi, Onno Leeksma, Enrico Maria Pogliani, Miriam Puttini, Eunjung Jang, Josy Reiffers, Maria Grazia Valsecchi, and Dong-Wook Kim. "Multicenter Independent Assessment of Outcomes in Myeloid Leukemia Patients Treated with Imatinib," *Journal of the National Cancer Institute* 103 (2011): 553-561. http://jnci.oxfordjournals.org.

Huang, Bin, Mark Dignan, Daikwon Han, and Owen Johnson. "Does Distance Matter? Distance to Mammography Facilities and Stage at Diagnosis of Breast Cancer in Kentucky," *The Journal of Rural Health* 25 (2009):366-371. doi: 10.1111/j.1748-0361.2009.00245.

Hudis, Clifford A. "Trastuzumab — Mechanism of Action and Use in Clinical Practice," *The New England Journal of Medicine* 357 (2007): 39-51. http://www.nejm.org.

Lakdawalla, Darius N., Eric C. Sun, Anupam B. Jena, Carolina M. Reyes, Dana P. Goldman, and Tomas J. Philipson. "An Economic Evaluation of the War on Cancer," *Journal of Health Economics* 29 (2010):333-346. doi: 10.1016/j.jealeco.2010.02.006.

Lichtenberg, Frank. "Are Increasing 5-Year Survival Rates Evidence of Success Against Cancer? A Reexamination Using Data From the U.S. and Australia," National Bureau of Economic Research Working Paper No. 16051, 2010. http://www.nber.org/papers/w16051.

Lichtenberg, Frank. "Has Medical Innovation Reduced Cancer Mortality?" National Bureau of Economic Research Working Paper No. 15880, 2010. http://www.nber.org/papers/w15880.

Lichtenberg, Frank. "The Expanding Pharmaceutical Arsenal in the War on Cancer," National Bureau of Economic Research Working Paper No. 10328, 2004. http://www.nber.org/papers/w10328.

Lieberman, Daniel. *The Story of the Human Body: Evolution, Health, and Disease*. New York: Knopf Doubleday Publishing Group, 2013.

Meliker, Jayme R., Pierre Goovaerts, Geoffrey M. Jacquez, Gilliam A. AvRuskin, and Glenn Copeland. "Breast and Prostate Cancer Survival in Michigan," *Cancer* 115 (2009):2212-21. doi: 10.1002/cnr.24251.

Mukherjee, Siddhartha. *The Emperor of All Maladies: A Biography of Cancer*. New York: Scribner, 2010.

Philipson, Tomas, Michael Eber, Darius N. Lakdawalla, Mitra Corral, Rena Conti, and Dana P. Goldman. "An Analysis of Whether Higher Health Care Spending in The United States Versus Europe Is 'Worth It' In the Case Of Cancer." *Health Affairs* 4 (2012): 667-675. Accessed February 24, 2014, doi:10.1377/hlthaff.2011.1298.

Rauh-Hain, J. Alejandro, Joel T. Clemmer, Leslie S. Bradford, Rachel M. Clark, Whitfield B. Growdon, Annekathryn Goodman, David M. Boruta II, John O. Schorge, and Marcela G. del Carmen. "Racial Disparities in Cervical Cancer Over Time," *Cancer* 119 (2013):3644-52. doi: 10.1002/cncr.28261.

Scoggins, John F., Catherine R. Fedorenko, Sara M. A. Donahue, Dedra Buchwald, David K. Blough, and Scott D. Ramsey. "Is Distance to Provider a Barrier to Care for Medicaid Patients with Breast, Colorectal, or Lung Cancer?" *The Journal of Rural Health* 28 (2012):54-62. doi: 10.1111/j.1748-0361.2011.00371.

Sun, Eric, Anupam B. Jena, Darius Lakdawalla, Carolina Reyes, Tomas J. Philipson, and Dan Goldman. "The Contributions of Improved Therapy and Earlier Detection to Cancer Survival Gains, 1988-2000," *Forum for Health Economics & Policy* 13 (2010): 1-20. doi:10.2202/1558-9544.1195.

Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2010), National Cancer Institute, DCCPS, Surveillance

Research Program, Surveillance Systems Branch, released April 2013, based on the November 2012 submission.

Wang, Fahui, Sara McLafferty, Veronic Escamilla, and Lan Luo. "Late-Stage Breast Cancer Diagnosis and Health Care Access in Illinois," *Prof Geogr* 60 (2008):54-69. doi: 10.1080/00330120701724087.

Welch, H. Gilbert, Lisa M. Schwartz, and Steven Woloshin. "Are Increasing 5-Year Survival Rates Evidence of Success Against Cancer?" *Journal of the American Medical Association* 283 (2000): 2975-2978. http://jamma.jamnetwork.com.

Yin, Wesley, John R. Penrod, J. Ross Maclean, Darius N. Lakdawalla, and Tomas Philipson. "Value of Survival Gains in Chronic Myeloid Leukemia," *The American Journal of Managed Care* 18 (2012): 257-264. http://www.ajmc.com.

XI. Appendix I - Maps

This appendix includes the maps that show the geographic distribution of late stage diagnosis, one-year survival, five-year survival, and mortality. Each map uses a color scale where the darkest color is the highest number and the lightest color is the lowest number. Because the data is dispersed throughout the United States most of the map appears blank. They are reported in the same order as the data is in the text.

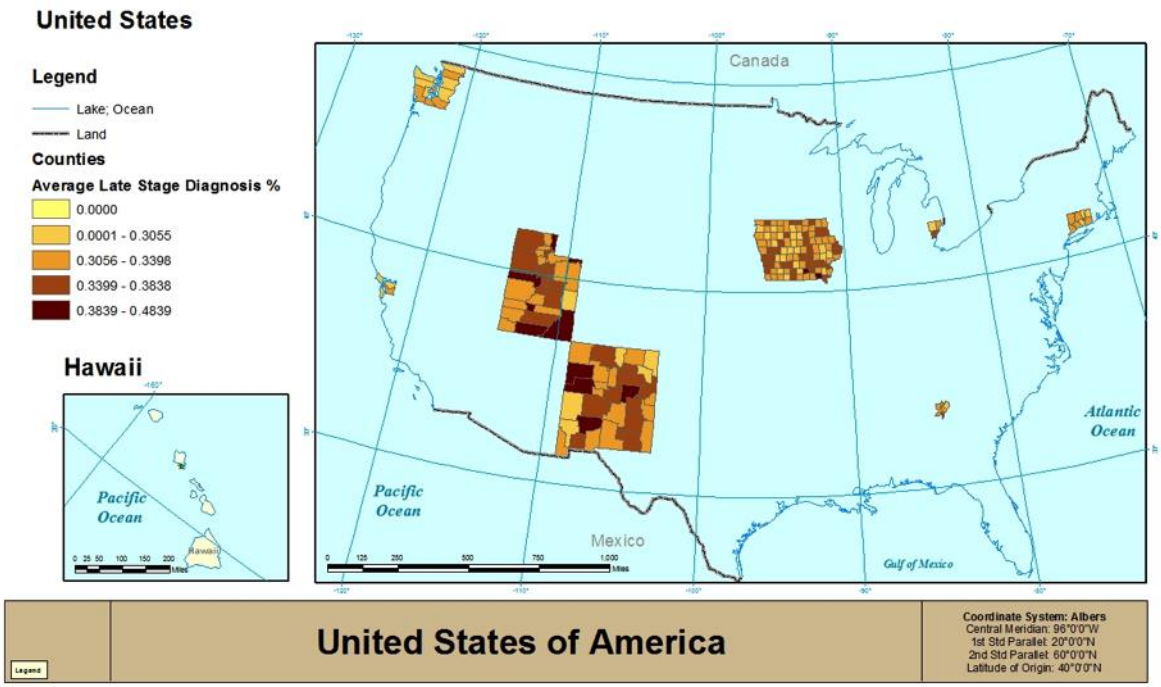

Figure 1: Geographic Variation in Late Stage Diagnosis Map

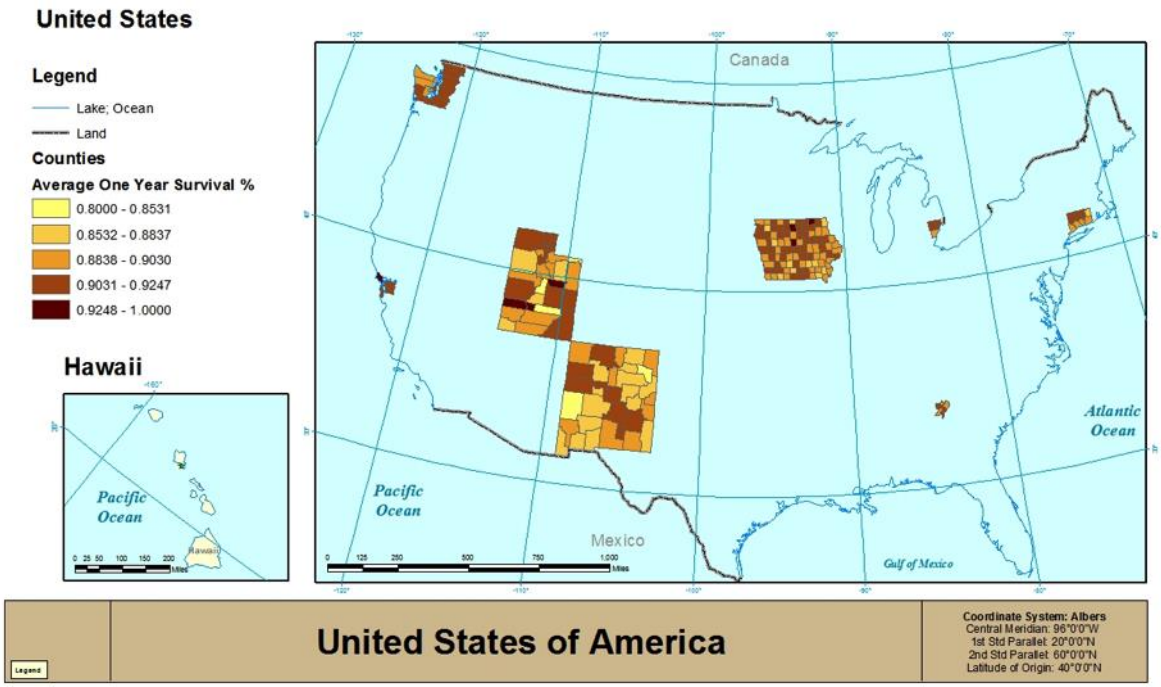

Figure 2: Geographic Variation in One-Year Survival Map

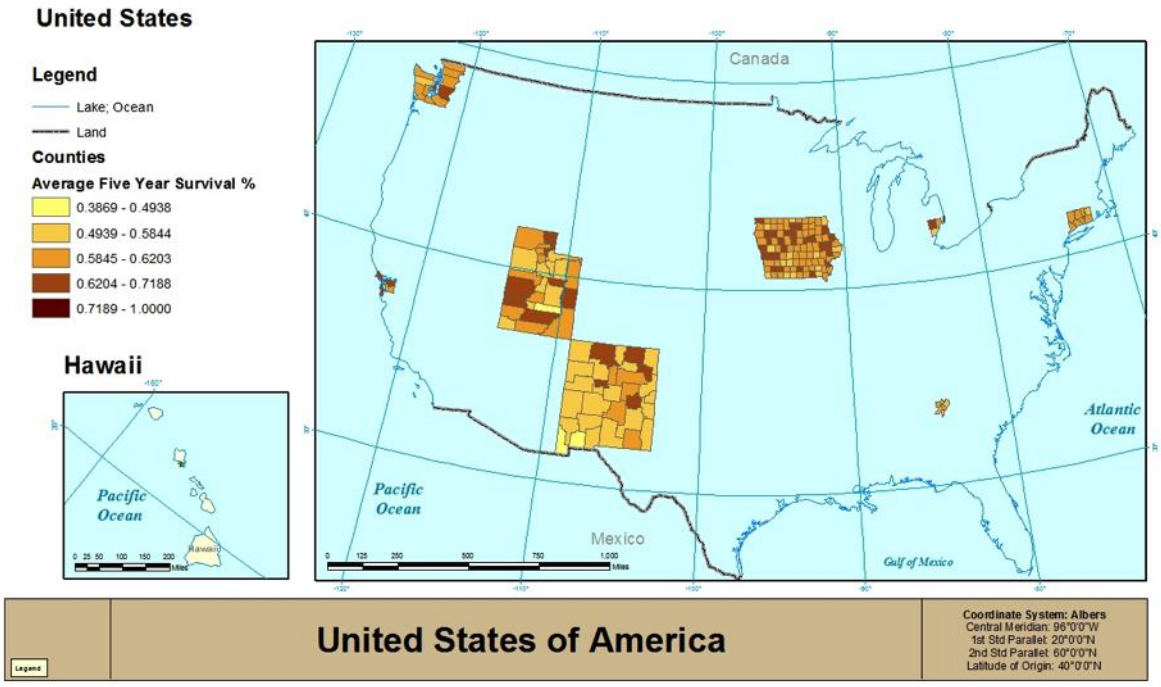

Figure 3: Geographic Variation in Five-Year Survival Map

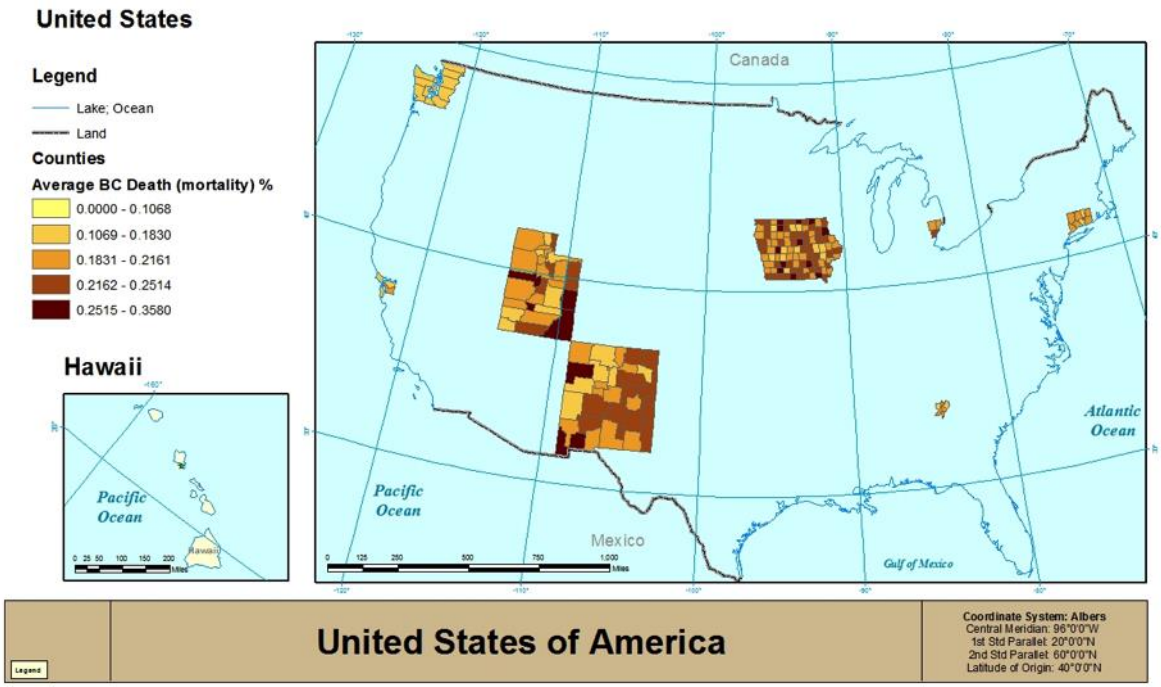

Figure 4: Geographic Variation in Mortality Map

XII. Appendix II – Data

Although a thorough discussion of the dataset appears in the text, several additional points are discussed here. This appendix has two sections. The first section includes this extended discussion and the second section includes the summary statistics from the dataset.

Section 1 – A Discussion of Data Limitations

The dataset used in this research is an excellent source of data. It is extensive and covers a large number of years. However, in addition to the socioeconomic limitations mentioned earlier, there are other limitations with the data. Cancer registries are designed to include all of the cancer cases from their geographic area—regardless of treatment location. To accomplish this, modern cancer registries share data to ensure complete coverage. It is not clear if the early data is as complete. However, this research assumes that coverage issues are minimal and will not bias the analysis.

Additionally, the comparison of data across such a large swath of years presents some potential issues. The main issue is the possibility that the data is not consistent across the years, even though the stage variable used in this analysis is a historic variable—meant to account for this possibility. Along with this specific issue with stage of diagnosis, the data's most likely potential issue stems from non-random differences across the years, such as changes in reporting methods or the consistency thereof. The SEER program seems to have taken steps to ensure data accuracy; there is significant evidence of recodes of data in the variable set. So, this effect is expected to be minimal and is assumed to not influence this research.

Finally, as will be seen in the summary statistics reported below, the dataset, while being representative across the years and stages, presents a non-random population derived from the areas in which these registries are located. Because of this, perfect interpretations of and extrapolations from the data are not possible. However, despite these limitations the analysis reaches useful conclusions.

Section 2 – Summary Statistics

The summary statistics for the data are presented here. Six tables are presented showing the number of observations for each variable.

Table 8: Geographic Distribution of Cases

Registry	Number of Cases	%
Connecticut	110903	16.2%
Detroit	115809	16.9%
San Francisco	114152	16.7%
Hawaii	28616	4.2%
Iowa	84072	12.3%
New Mexico	35735	5.2%
Seattle	103395	15.1%
Utah	34244	5.0%
Atlanta	57479	8.4%

Table 9: Marital Status Distribution of Cases

Marital Status	Number of Cases	%
Single	72671	10.6%
Married	679967	99.4%
Separated	9349	1.4%
Divorced	60266	8.8%
Widowed	135840	19.8%
Unmarried	12	0.0%
Unknown	25938	3.8%

Table 10: Racial Distribution of Cases

Race	Number of Cases	%
White	580371	84.8%
Black	57181	8.4%
American Indian	2767	0.4%
Chinese	8614	1.3%
Japanese	11688	1.7%
Filipino	8129	1.2%
Hawaiian	4968	0.7%
Other	825	0.1%
Unknown	2115	0.3%
Hispanic Status		
Hispanic	25046	3.7%
Unknown	24649	3.6%
Not Hispanic	634699	92.7%

Table 11: Yearly Distribution of Cases

Year	Number of Cases	%
1973	7562	1.1%
1974	10017	1.5%
1975	10237	1.5%
1976	10024	1.5%
1977	9964	1.5%
1978	10128	1.5%
1979	10520	1.5%
1980	10743	1.6%
1981	11331	1.7%
1982	11535	1.7%
1983	12322	1.8%
1984	13248	1.9%
1985	14640	2.1%
1986	15419	2.3%
1987	16964	2.5%
1988	16919	2.5%
1989	16589	2.4%
1990	17681	2.6%
1991	18283	2.7%
1992	18663	2.7%
1993	18579	2.7%
1994	19302	2.8%

1995	20158	2.9%
1996	20707	3.0%
1997	22060	3.2%
1998	23513	3.4%
1999	23871	3.5%
2000	23588	3.4%
2001	24415	3.6%
2002	24395	3.6%
2003	23312	3.4%
2004	23943	3.5%
2005	24025	3.5%
2006	24509	3.6%
2007	25517	3.7%
2008	26216	3.8%
2009	27131	4.0%
2010	26364	3.9%

Table 12: Distribution of Cases by Stage of Diagnosis

Stage	Number of Cases	%
1	101568	14.8%
2	344207	50.3%
3	182256	26.6%
4	36126	5.3%

Table 13: Percentage of Cases for Each Dependent Variable

Statistic	Number of Cases	%
Late Stage	218382	31.9%
Five Year Survival	417817	61.0%
One Year Survival	619224	90.5%
Mortality	131267	19.2%

XIII. Appendix III – Empirical Specification

In the text of this thesis, two regression equations are reported but the findings from 26 individual regressions are reported. In this appendix, the equations for those 26 regressions are reported in the order in which they appeared in the text.

Equation 1 – The Effect of Stage of Diagnosis on Five-Year Survival

$$Five\ Year\ Survival = \alpha + \beta_1(Stage\ of\ Diagnosis) + \beta_2(Race\ Dummies) + \beta_3(Marital\ Status\ Dummies) + \beta_4(Hispanic\ Dummies) + \beta_5(Age\ of\ Diagnosis) + \beta_6(Age\ Squared) + \beta_7(Year\ Vector) + \beta_8(Radiation) + \varepsilon$$

Equation 2 – The Effect of Stage of Diagnosis on Five-Year Survival, with County FE

$$Five\ Year\ Survival = \alpha + \beta_1(Stage\ of\ Diagnosis) + \beta_2(Race\ Dummies) + \beta_3(Marital\ Status\ Dummies) + \beta_4(Hispanic\ Dummies) + \beta_5(Age\ of\ Diagnosis) + \beta_6(Age\ Squared) + \beta_7(Year\ Vector) + \beta_8(Radiation) + \beta_9(County\ FE) + \varepsilon$$

Equation 3 – The Effect of Stage of Diagnosis on One-Year Survival

$$One\ Year\ Survival = \alpha + \beta_1(Stage\ of\ Diagnosis) + \beta_2(Race\ Dummies) + \beta_3(Marital\ Status\ Dummies) + \beta_4(Hispanic\ Dummies) + \beta_5(Age\ of\ Diagnosis) + \beta_6(Age\ Squared) + \beta_7(Year\ Vector) + \beta_8(Radiation) + \varepsilon$$

Equation 4 – The Effect of Stage of Diagnosis on One-Year Survival, with County FE

$$One\ Year\ Survival$$
$$= \alpha + \beta_1(Stage\ of\ Diagnosis) + \beta_2(Race\ Dummies)$$
$$+ \beta_3(Marital\ Status\ Dummies) + \beta_4(Hispanic\ Dummies)$$
$$+ \beta_5(Age\ of\ Diagnosis) + \beta_6(Age\ Squared)$$
$$+ \beta_7(Year\ Vector) + \beta_8(Radiation) + \beta_9(County\ FE) + \varepsilon$$

Equation 5 – The Effect of Stage of Diagnosis on Mortality

$$Mortality\ (BC\ Death)$$
$$= \alpha + \beta_1(Stage\ of\ Diagnosis) + \beta_2(Race\ Dummies)$$
$$+ \beta_3(Marital\ Status\ Dummies) + \beta_4(Hispanic\ Dummies)$$
$$+ \beta_5(Age\ of\ Diagnosis) + \beta_6(Age\ Squared) + \beta_7(Year\ Vector) + \varepsilon$$

Equation 6 – The Effect of Stage of Diagnosis on Mortality, with County FE

$$Mortality\ (BC\ Death)$$
$$= \alpha + \beta_1(Stage\ of\ Diagnosis) + \beta_2(Race\ Dummies)$$
$$+ \beta_3(Marital\ Status\ Dummies) + \beta_4(Hispanic\ Dummies)$$
$$+ \beta_5(Age\ of\ Diagnosis) + \beta_6(Age\ Squared) + \beta_7(Year\ Vector)$$
$$+ \beta_8(County\ FE) + \varepsilon$$

Equation 7 – The Effect of Five-Year Survival on Mortality

$$Mortality\ (BC\ Death)$$
$$= \alpha + \beta_1(Five\ Year\ Survival) + \beta_2(Race\ Dummies)$$
$$+ \beta_3(Marital\ Status\ Dummies) + \beta_4(Hispanic\ Dummies)$$
$$+ \beta_5(Age\ of\ Diagnosis) + \beta_6(Age\ Squared) + \beta_7(Year\ Vector) + \varepsilon$$

Equation 8 – The Effect of Five-Year Survival on Mortality, with County FE

$$Mortality\ (BC\ Death)$$
$$= \alpha + \beta_1(Five\ Year\ Survival) + \beta_2(Race\ Dummies)$$
$$+ \beta_3(Marital\ Status\ Dummies) + \beta_4(Hispanic\ Dummies)$$
$$+ \beta_5(Age\ of\ Diagnosis) + \beta_6(Age\ Squared) + \beta_7(Year\ Vector)$$
$$+ \beta_8(County\ FE) + \varepsilon$$

Equation 9 – The Effect of One-Year Survival on Mortality

$$Mortality\ (BC\ Death)$$
$$= \alpha + \beta_1(One\ Year\ Survival) + \beta_2(Race\ Dummies)$$
$$+ \beta_3(Marital\ Status\ Dummies) + \beta_4(Hispanic\ Dummies)$$
$$+ \beta_5(Age\ of\ Diagnosis) + \beta_6(Age\ Squared) + \beta_7(Year\ Vector) + \varepsilon$$

Equation 10 – The Effect of Stage of One-Year Survival on Mortality, with County FE

$$Mortality\ (BC\ Death)$$
$$= \alpha + \beta_1(One\ Year\ Survival) + \beta_2(Race\ Dummies)$$
$$+ \beta_3(Marital\ Status\ Dummies) + \beta_4(Hispanic\ Dummies)$$
$$+ \beta_5(Age\ of\ Diagnosis) + \beta_6(Age\ Squared) + \beta_7(Year\ Vector)$$
$$+ \beta_8(County\ FE) + \varepsilon$$

Equation 11 – The Geographic Variation of Late Stage Diagnosis

$$Stage\ III\ or\ IV\ Diagnosis$$
$$= \alpha + \beta_1(Registry\ Dummies) + \beta_2(Age\ of\ Diagnosis) + \beta_3(Age\ Squared)$$
$$+ \beta_4(Year\ Vector) + \varepsilon$$

Equation 12 – The Geographic Variation of Late Stage Diagnosis, with County FE

$$Stage\ III\ or\ IV\ Diagnosis$$
$$= \alpha + \beta_1(Registry\ Dummies) + \beta_2(Age\ of\ Diagnosis) + \beta_3(Age\ Squared)$$
$$+ \beta_4(Year\ Vector) + \beta_5(County\ FE) + \varepsilon$$

Equation 13 – The Geographic Variation of Late Stage Diagnosis, with Full Controls

$$Stage\ III\ or\ IV\ Diagnosis$$
$$= \alpha + \beta_1(Registry\ Dummies) + \beta_2(Race\ Dummies)$$
$$+ \beta_3(Marital\ Status\ Dummies) + \beta_4(Hispanic\ Dummies)$$
$$+ \beta_5(Age\ of\ Diagnosis) + \beta_6(Age\ Squared) + \beta_7(Year\ Vector) + \varepsilon$$

Equation 14 – The Geographic Variation of Late Stage Diagnosis, with Full Controls and County FE

$$Stage\ III\ or\ IV\ Diagnosis$$
$$= \alpha + \beta_1(Registry\ Dummies) + \beta_2(Race\ Dummies)$$
$$+ \beta_3(Marital\ Status\ Dummies) + \beta_4(Hispanic\ Dummies)$$
$$+ \beta_5(Age\ of\ Diagnosis) + \beta_6(Age\ Squared)$$
$$+ \beta_7(Year\ Vector) + \beta_8(County\ FE) + \varepsilon$$

Equation 15 – The Geographic Variation of One-Year Survival

$$One\ Year\ Survival$$
$$= \alpha + \beta_1(Registry\ Dummies) + \beta_2(Age\ of\ Diagnosis) + \beta_3(Age\ Squared)$$
$$+ \beta_4(Year\ Vector) + \beta_5(Radiation) + \beta_6(Stage\ of\ Diagnosis\ Dummies)$$
$$+ \varepsilon$$

Equation 16 – The Geographic Variation of One-Year Survival, with County FE

$$One\ Year\ Survival$$
$$= \alpha + \beta_1(Registry\ Dummies) + \beta_2(Age\ of\ Diagnosis) + \beta_3(Age\ Squared) + \beta_4(Year\ Vector) + \beta_5(Radiation) + \beta_6(Stage\ of\ Diagnosis\ Dummies) + \beta_7(County\ FE) + \varepsilon$$

Equation 17 – The Geographic Variation of One-Year Survival, with Full Controls and Interactions

$$One\ Year\ Survival$$
$$= \alpha + \beta_1(Registry\ Dummies) + \beta_2(Race\ Dummies) + \beta_3(Marital\ Status\ Dummies) + \beta_4(Hispanic\ Dummies) + \beta_5(Age\ of\ Diagnosis) + \beta_6(Age\ Squared) + \beta_7(Year\ Vector) + \beta_8(Radiation) + \beta_9(Stage\ of\ Diagnosis\ Dummies) + \beta_{10}(Age\ x\ Stage\ Interations) + \beta_{11}(Age\ x\ Stage\ x\ Year\ Interactions) + \varepsilon$$

Equation 18 – The Geographic Variation of One-Year Survival, with Full Controls, Interactions, and County FE

$$One\ Year\ Survival$$
$$= \alpha + \beta_1(Registry\ Dummies) + \beta_2(Race\ Dummies) + \beta_3(Marital\ Status\ Dummies) + \beta_4(Hispanic\ Dummies) + \beta_5(Age\ of\ Diagnosis) + \beta_6(Age\ Squared) + \beta_7(Year\ Vector) + \beta_8(Radiation) + \beta_9(Stage\ of\ Diagnosis\ Dummies) + \beta_{10}(Age\ x\ Stage\ Interations) + \beta_{11}(Age\ x\ Stage\ x\ Year\ Interactions) + \beta_{12}(County\ FE) + \varepsilon$$

Equation 19 – The Geographic Variation of Five-Year Survival

$$Five\ Year\ Survival$$
$$= \alpha + \beta_1(Registry\ Dummies) + \beta_2(Age\ of\ Diagnosis) + \beta_3(Age\ Squared)$$
$$+ \beta_4(Year\ Vector) + \beta_5(Radiation) + \beta_6(Stage\ of\ Diagnosis\ Dummies)$$
$$+ \varepsilon$$

Equation 20 – The Geographic Variation of Five-Year Survival, with County FE

$$Five\ Year\ Survival$$
$$= \alpha + \beta_1(Registry\ Dummies) + \beta_2(Age\ of\ Diagnosis) + \beta_3(Age\ Squared)$$
$$+ \beta_4(Year\ Vector)$$
$$+ \beta_5(Radiation) + \beta_6(Stage\ of\ Diagnosis\ Dummies) + \beta_7(County\ FE) + \varepsilon$$

Equation 21 – The Geographic Variation of Five-Year Survival, with Full Controls and Interactions

$$Five\ Year\ Survival$$
$$= \alpha + \beta_1(Registry\ Dummies) + \beta_2(Race\ Dummies)$$
$$+ \beta_3(Marital\ Status\ Dummies) + \beta_4(Hispanic\ Dummies)$$
$$+ \beta_5(Age\ of\ Diagnosis) + \beta_6(Age\ Squared) + \beta_7(Year\ Vector)$$
$$+ \beta_8(Radiation) + \beta_9(Stage\ of\ Diagnosis\ Dummies)$$
$$+ \beta_{10}(Age\ x\ Stage\ Interations) + \beta_{11}(Age\ x\ Stage\ x\ Year\ Interactions)$$
$$+ \varepsilon$$

Equation 22 – The Geographic Variation of Five-Year Survival, with Full Controls, Interactions, and County FE

$$Five\ Year\ Survival$$

$$= \alpha + \beta_1(Registry\ Dummies) + \beta_2(Race\ Dummies)$$
$$+ \beta_3(Marital\ Status\ Dummies) + \beta_4(Hispanic\ Dummies)$$
$$+ \beta_5(Age\ of\ Diagnosis) + \beta_6(Age\ Squared) + \beta_7(Year\ Vector)$$
$$+ \beta_8(Radiation) + \beta_9(Stage\ of\ Diagnosis\ Dummies)$$
$$+ \beta_{10}(Age\ x\ Stage\ Interations)$$
$$+ \beta_{11}(Age\ x\ Stage\ x\ Year\ Interactions) + \beta_{12}(County\ FE) + \varepsilon$$

Equation 23 – The Geographic Variation of Mortality

$$Mortality\ (BC\ Death)$$

$$= \alpha + \beta_1(Registry\ Dummies) + \beta_2(Age\ of\ Diagnosis) + \beta_3(Age\ Squared)$$
$$+ \beta_4(Year\ Vector) + \beta_5(Radiation) + \beta_6(Stage\ of\ Diagnosis\ Dummies)$$
$$+ \varepsilon$$

Equation 24 – The Geographic Variation of Mortality, with County FE

$$Mortality\ (BC\ Death)$$

$$= \alpha + \beta_1(Registry\ Dummies) + \beta_2(Age\ of\ Diagnosis) + \beta_3(Age\ Squared)$$
$$+ \beta_4(Year\ Vector)$$
$$+ \beta_5(Radiation) + \beta_6(Stage\ of\ Diagnosis\ Dummies) + \beta_7(County\ FE) + \varepsilon$$

Equation 25 – The Geographic Variation of Mortality, with Full Controls and Interactions

$$Mortality\ (BC\ Death)$$

$$= \alpha + \beta_1(Registry\ Dummies) + \beta_2(Race\ Dummies)$$

$$+ \beta_3\ (Marital\ Status\ Dummies) + \beta_4(Hispanic\ Dummies)$$

$$+ \beta_5(Age\ of\ Diagnosis) + \beta_6(Age\ Squared) + \beta_7(Year\ Vector)$$

$$+ \beta_8(Radiation) + \beta_9(Stage\ of\ Diagnosis\ Dummies)$$

$$+ \beta_{10}(Age\ x\ Stage\ Interations) + \beta_{11}(Age\ x\ Stage\ x\ Year\ Interactions)$$

$$+ \varepsilon$$

Equation 26 – The Geographic Variation of Mortality, with Full Controls, Interactions, and County FE

$$Mortality\ (BC\ Death) = \alpha + \beta_1(Registry\ Dummies) + \beta_2(Race\ Dummies) +$$

$$\beta_3\ (Marital\ Status\ Dummies) + \beta_4(Hispanic\ Dummies) + \beta_5(Age\ of\ Diagnosis) +$$

$$\beta_6(Age\ Squared) + \beta_7(Year\ Vector) +$$

$$\beta_8(Radiation) + \beta_9(Stage\ of\ Diagnosis\ Dummies) + \beta_{10}(Age\ x\ Stage\ Interations) +$$

$$\beta_{11}(Age\ x\ Stage\ x\ Year\ Interactions) + \beta_{12}(County\ FE) + \varepsilon$$

XIV. Appendix IV – Findings

In the text of this thesis, seven abbreviated tables are reported with no discussion of the controls used. This appendix contains three different sections. In the first section, the controls from one regression are reported and analyzed. In the second section, a further analysis of the four variables of interest is included. In the third section, the complete tables (without county-level fixed effects) are reported.

Section 1 – A Discussion of Controls

To enable a salient discussion of the controls and interactions used throughout this thesis, the controls from Regression 3 in Table 7 are reported below:

Table 14: Control Variables for the Geographic Variation in Mortality, Regression 3

VARIABLES	(3) BC Death
Year of Diagnosis	-0.0122***
	(0.000146)
Age at Diagnosis	-0.00120***
	(0.000302)
Age at Diagnosis Squared	2.57e-05***
	(2.35e-06)
Stage II	0.190***
	(0.00681)
Stage III	0.478***
	(0.00731)
Stage IV	0.885***
	(0.0120)
Radiation	0.000927
	(0.00118)
Single	0.0177***
	(0.00184)
Separated	0.00700
	(0.00442)
Divorced	0.0171***
	(0.00195)
Widowed	0.0107***
	(0.00156)
Black	0.0550***
	(0.00211)
American Indian	0.0199**
	(0.00913)
Chinese	-0.000459
	(0.00518)

Japanese	-0.0356***
	(0.00495)
Filipino	0.0113**
	(0.00549)
Hawaiian	0.0209***
	(0.00722)
Hispanic	0.00544*
	(0.00314)
Age x Stage II	-0.00465***
	(0.000120)
Age x Stage III	-0.00471***
	(0.000128)
Age x Stage IV	-0.00702***
	(0.000201)
Year x Age x Stage II	8.31e-05***
	(2.63e-06)
Year x Age x Stage III	2.14e-05***
	(2.94e-06)
Year x Age x Stage IV	0.000145***
	(4.52e-06)
Fixed Effects	No
Constant	0.324***
	-0.0106
Observations	502,467
R-squared	0.202

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In this regression, there are two classes of controls used: marital/racial status controls and normal controls. Additionally, this regression includes the two types of interaction variables: Age x Stage and Year x Age x Stage. The normal controls in this regression are *Year of Diagnosis*, *Age at Diagnosis,* and *Age Squared.* Additionally, this regression includes the stage of diagnosis variables. These variables are included in the survival and mortality geographic distribution regressions. In the following paragraphs, the control variables are described and interpreted (as will the interaction variables). Similar interpretations, as noted, follow for all of the regressions. As a whole, the normal controls catch the effect of age and technology upon the dependent variables (mortality it here), the marital/racial status controls capture the effects of socioeconomic status that are

associated with either race or marital status. Most of socioeconomic status will be associated with these variables, and the interaction terms will show the relationships they contain. These results are consistent with previous literature examining race and socioeconomic status (see the Literature Review)

The normal controls in this section, which include the stage of diagnosis, show their effects. The *Year of Diagnosis* vector shows the effect of each additional year after 1973, in effect it demonstrates the technological/imaging/treatment change over the course of time. In this case, this variable shows that the odds of breast cancer death decrease about 1% each year. The *Age of Diagnosis* variable shows the effect of age and the *Age Squared* variable shows that effect squared. Collectively, they capture the effect of age. The negative coefficient for *Age of Diagnosis* shows how increasing age is consistent with decreasing mortality, while the small, positive coefficient on *Age Squared* shows that this effect becomes less negative and eventually becomes positive with old age. This means that for most of a woman's life the older she gets the less likely she is to die of breast cancer (conditional on diagnosis). This effect is likely two fold. The first reason is that early onset breast cancer, which is genetic, is considered to be extremely virulent. These genetic tumors, caused by mutations in the BRCA1 and BRCA2 genes, are more dangerous than later onset tumors. The second reason is that older women have less "survival time" left as it is, so the odds they will die of their breast cancer and not something else (like old age or comorbidities) are decreased. This effect disappears at extremely high ages, likely indicating the debilitating effects of tumors at such an advanced age. The *Radiation* variable shows the effect that having radiation treatment has on mortality, the coefficient is slightly positive but not statistically significant. This effect could come about for a variety of

reasons. Radiation therapy is used on a non-random population of breast tumors, so this variable does not solely represent the effect of radiation—this population of tumors is likely more virulent, larger tumors that cannot be operated on. Additionally, radiation therapy could only have therapeutic effects in the short run, if radiation only delays the tumor's growth then it could come back and kill the patient. The *Stage of Diagnosis* variables show the effect that being diagnosed at a later stage than stage I (II, III, or IV) on the dependent variables (mortality in this case). These variables are obviously not included in regressions with stage of diagnosis as the dependent variable. The effects are consistent, later stages of diagnosis decrease the likelihood of survival and increase the likelihood of mortality. These normal controls are used in most of the regressions (see Appendix III – Empirical Specification).

The marital status and racial controls are used in all of the sets of regressions. They were selected because they are the closest variables included in the dataset to socioeconomic status. Many socioeconomic and associated lifestyle factors are expected to correlate with some of these variables (see the Literature Review). There is clear omitted variable bias in the lack of socioeconomic data. Socioeconomic status controls the access a patient has to the best care and newest medicines. Higher socioeconomic status is also associated with healthier lifestyles and diet. So, while all these patients have been diagnosed with cancer, those with higher socioeconomic status would be the best prepared to survive the longest. There are clear connections with socioeconomic status and health. The marital status variables show that married people are the least likely to die of cancer (conditional upon diagnosis), but these variables are better interpreted as collectively catching socioeconomic factors—the difference between married and single people,

separated and divorced, etc. The race variables also capture socioeconomic status and lifestyle correlates, as can be seen here. Combined, these two sets of variables capture a significant portion of socioeconomic status and lifestyle factors—those associated with marital status and race.

In addition to control variables, this regression includes interaction terms. These terms capture the interactions between different variables. The first set, which captures the interaction between stage and age, shows the effects of the different stages that are compounded by age. So, for instance, the coefficients for these three interactions are negative. This means that as age at diagnosis increases, patients are less likely to die of breast cancer after being diagnosed. This is consistent with the previous observations of lower virulence and increased risks (comorbidities/old age) at older ages. Thus, this interaction captures the interaction as designed. The second set, which captures the interaction between year and stage and age, shows how the effect of the first set changes over time. In this reaction, these are all slightly positive. This means that as the years progress the previous effect is diminished—those diagnosed at later stages at later ages are more likely to die from breast cancer after being diagnosed. This is expected. This is a result of the improving detection technology. The discussion in the findings section related to the idea that detection technology has improved and helped save lives. Here, the removal of less virulent tumors (that are now caught earlier) from the later stage diagnoses means that the tumors that are diagnosed at that stage are more virulent. So this interaction effect is expected.

Together, these control variables and interaction terms help to provide greater explanatory power to the model and remove potential omitted variable bias. The weakness

of these control variables is that they are all covariates with other omitted variables that could not be included—they are approximates for socioeconomic and lifestyle factors. Although this is a weakness to these controls, it does not hurt the analysis. The point of this research is to understand the cancer statistics, their relationship, and their geographic variation. It is not possible to determine the causes behind the geographic variation from this research, and it is clear that some of this variation is likely due to socioeconomic status factors that track geographically and not in one of the ways controlled for here. But, that is perfectly acceptable. Additionally, this fact only hurts the first group of regressions if cancer is diagnosed in an abnormal population distribution. This is not expected to be the case because cancer is a disease inherent in our genetic structure, but, except 2-10% of cases brought on at an early age, it is not heritable. So, the lack of socioeconomic controls limits the explanatory power of the models but does not corrupt results or the interpretation.

Finally, some of the regressions include county-level fixed effects. The previous literature suggests county level effects due to community variation, socioeconomic status, or distance from treatment centers (Meliker et al, 2009; Huang et al, 2009). Thus, county-level fixed effects account for the variation by county and allow a clear inspection of the geographic variation to be made.

Section 2 – A Discussion of the Dependent Variables

There are four primary dependent variables included in this analysis: late stage diagnosis, one-year survival, five-year survival, and death due to breast cancer (mortality). Each of these variables was selected to elucidate a different aspect of breast cancer. The late stage diagnosis variable shows diagnosis practices, the survival variables show the

length of time patients live, and the mortality/BC Death variable shows whether or not patients die of breast cancer. These variables are analyzed independently to show geographic variation and together to show their connections, but this analysis does not include change over time. However, the evolution of each variable over time can be deduced from the data. The tables containing the data referenced in this section are in the following section of this appendix.

From the regressions testing the effect of diagnosis on survival, it can be deduced that survival is increasing over the years for breast cancer. the year variable has a positive coefficient. However, the practical effects are small, each additional year increases five-year survival likelihood by 0.004 and one-year survival likelihood by 0.0098. This indicates that while there is increasing trends in survival over this time period. The following tables also indicate that mortality is decreasing over time, but, yet again, these effects are small. The odds of dying of breast cancer decrease by less than 1% each year. These effects are small in individual years but large in aggregate, the yearly effect over the entirety of the dataset—32 years—increases five-year survival likelihood by 12.8% and decreases mortality likelihood by 28%, conditional upon diagnosis. Both of these statistics confirm the previous literature showing that survival is increasing over time and mortality is decreasing over time (Philipson et al, 2012). They also confirm the intuition that treatments have been improving over time.

Stage of Diagnosis undergoes a similar pattern. The data for its geographic distribution show suggest that the likelihood of later diagnosis (conditional upon diagnosis) decreases 23.2% during the course of the dataset. This also parallels previous research suggesting that diagnosis techniques have improved.

The effects of the year variable reported here show that improvements in all three statistics have been made. This idea re-enforces the connection between the statistics and provides evidence for success against cancer. It is impossible to speculate accurately on the success of alternative histories but here we see that there has been success.

Because of this success, this research focuses instead upon the connection between the statistics and the geographic distribution of them. These four statistics combine to capture a complete image of breast cancer at any point in time. This research shows the variation and connections that remain in this changing cancer landscape.

Section 3 – Full Results Tables

This section includes the tables for the full regressions (excluding county-level fixed effects) for all the regression included in this thesis. There are seven tables (15-21) following the same order as the tables in the body of the thesis (1-7). They are numbered in the same manner as well.

Table 15: Effect of Diagnosis on Survival

VARIABLES	(1) FiveYears	(2) FiveYears	(3) OneYear	(4) OneYear
Stage II	0.00537***	0.00511***	0.0217***	0.0216***
	(0.00157)	(0.00157)	(0.000860)	(0.000861)
Stage III	-0.155***	-0.154***	-0.00555***	-0.00539***
	(0.00173)	(0.00173)	(0.000945)	(0.000946)
Stage IV	-0.617***	-0.615***	-0.330***	-0.329***
	(0.00272)	(0.00272)	(0.00149)	(0.00149)
Year of Diagnosis	0.00432***	0.00435***	0.000976***	0.000977***
	(6.47e-05)	(6.57e-05)	(3.54e-05)	(3.60e-05)
Age at Diagnosis	0.0290***	0.0291***	0.00858***	0.00863***
	(0.000288)	(0.000288)	(0.000157)	(0.000158)
Age at Diagnosis Squared	-0.000286***	-0.000287***	-8.70e-05***	-8.74e-05***
	(2.36e-06)	(2.36e-06)	(1.29e-06)	(1.29e-06)
Single	-0.0267***	-0.0275***	-0.0110***	-0.0110***
	(0.00185)	(0.00186)	(0.00101)	(0.00102)
Separated	-0.0509***	-0.0481***	-0.0192***	-0.0188***
	(0.00440)	(0.00446)	(0.00241)	(0.00244)
Divorced	-0.0342***	-0.0355***	-0.00980***	-0.0101***

	(0.00196)	(0.00197)	(0.00108)	(0.00108)
Widowed	-0.0331***	-0.0330***	-0.0111***	-0.0110***
	(0.00157)	(0.00158)	(0.000861)	(0.000863)
Black	-0.0837***	-0.0772***	-0.0229***	-0.0203***
	(0.00205)	(0.00221)	(0.00112)	(0.00121)
American Indian	-0.0602***	-0.0582***	-0.00291	-0.00425
	(0.00917)	(0.00946)	(0.00502)	(0.00518)
Chinese	0.0197***	0.00867	0.00381	0.00118
	(0.00510)	(0.00530)	(0.00279)	(0.00290)
Japanese	0.0548***	0.0435***	0.00933***	0.00659**
	(0.00411)	(0.00501)	(0.00225)	(0.00274)
Filipino	-0.0117**	-0.0212***	-0.00277	-0.00497
	(0.00535)	(0.00556)	(0.00293)	(0.00304)
Hawaiian	-0.0347***	-0.0442***	-0.0135***	-0.0153***
	(0.00643)	(0.00731)	(0.00352)	(0.00400)
Hispanic	-0.0192***	-0.0193***	-0.00496***	-0.00506***
	(0.00304)	(0.00321)	(0.00166)	(0.00176)
Radiation	0.0243***	0.0225***	0.0138***	0.0133***
	(0.00118)	(0.00119)	(0.000646)	(0.000650)
Fixed Effects	No	Yes	No	Yes
Constant	0.121***	0.122***	0.757***	0.756***
	(0.00875)	(0.00907)	(0.00479)	(0.00497)
Observations	502,467	502,467	502,467	502,467
R-squared	0.213	0.214	0.154	0.155

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 16: Effect of Diagnosis on Mortality

VARIABLES	(1) BC Death	(2) BC Death
Stage II	0.0122***	0.0123***
	(0.00155)	(0.00156)
Stage III	0.232***	0.232***
	(0.00171)	(0.00171)
Stage IV	0.621***	0.620***
	(0.00269)	(0.00269)
Year of Diagnosis	-0.00875***	-0.00875***
	(6.22e-05)	(6.33e-05)
Age at Diagnosis	-0.00516***	-0.00513***
	(0.000285)	(0.000285)
Age at Diagnosis Squared	3.34e-05***	3.29e-05***
	(2.34e-06)	(2.34e-06)
Single	0.0138***	0.0160***
	(0.00184)	(0.00185)
Separated	-0.00404	-0.00276
	(0.00437)	(0.00443)
Divorced	0.0164***	0.0187***
	(0.00195)	(0.00195)

Widowed	0.00974***	0.0100***
	(0.00156)	(0.00156)
Black	0.0571***	0.0563***
	(0.00204)	(0.00220)
American Indian	0.0243***	0.0119
	(0.00911)	(0.00939)
Chinese	-0.0167***	0.00102
	(0.00507)	(0.00526)
Japanese	-0.0524***	-0.0320***
	(0.00408)	(0.00498)
Filipino	-0.00434	0.0110**
	(0.00532)	(0.00552)
Hawaiian	0.000508	0.0202***
	(0.00639)	(0.00726)
Hispanic	0.00826***	0.00615*
	(0.00302)	(0.00319)
Fixed Effects	No	Yes
Constant	0.460***	0.453***
	(0.00869)	(0.00900)
Observations	502,467	502,467
R-squared	0.196	0.197

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 17: Effect of Survival on Mortality

VARIABLES	(1) BC Death	(2) BC Death	(3) BC Death	(4) BC Death
Five Year Survival	-0.455***	-0.455***		
	(0.00126)	(0.00126)		
One Year Survival			-0.328***	-0.327***
			(0.00256)	(0.00256)
Year of Diagnosis	-0.00797***	-0.00797***	-0.0103***	-0.0103***
	(5.98e-05)	(6.09e-05)	(6.56e-05)	(6.67e-05)
Age at Diagnosis	0.00764***	0.00767***	-0.00310***	-0.00308***
	(0.000277)	(0.000278)	(0.000304)	(0.000304)
Age at Diagnosis Squared	-9.58e-05***	-9.63e-05***	8.30e-06***	7.75e-06***
	(2.29e-06)	(2.29e-06)	(2.50e-06)	(2.50e-06)
Single	0.00928***	0.0115***	0.0220***	0.0247***
	(0.00177)	(0.00178)	(0.00196)	(0.00197)
Separated	-0.00656	-0.00366	0.0198***	0.0216***
	(0.00421)	(0.00427)	(0.00465)	(0.00471)
Divorced	0.00653***	0.00861***	0.0216***	0.0244***
	(0.00188)	(0.00189)	(0.00208)	(0.00208)
Widowed	0.000181	0.000540	0.0146***	0.0149***
	(0.00151)	(0.00151)	(0.00166)	(0.00167)
Black	0.0338***	0.0355***	0.0718***	0.0709***
	(0.00197)	(0.00212)	(0.00217)	(0.00234)
American Indian	0.0124	-0.00184	0.0460***	0.0292***

	(0.00879)	(0.00906)	(0.00970)	(0.0100)
Chinese	-0.0150***	0.000259	-0.0263***	-0.00514
	(0.00489)	(0.00508)	(0.00540)	(0.00560)
Japanese	-0.0413***	-0.0222***	-0.0700***	-0.0445***
	(0.00394)	(0.00480)	(0.00435)	(0.00530)
Filipino	-0.00878*	0.00493	-0.00378	0.0152***
	(0.00513)	(0.00532)	(0.00566)	(0.00588)
Hawaiian	-0.00926	0.0101	0.00475	0.0300***
	(0.00616)	(0.00700)	(0.00680)	(0.00773)
Hispanic	0.00807***	0.00540*	0.0191***	0.0163***
	(0.00291)	(0.00308)	(0.00321)	(0.00340)
Fixed Effects	No	Yes	No	Yes
Constant	0.629***	0.623***	0.874***	0.864***
	(0.00826)	(0.00857)	(0.00932)	(0.00965)
Observations	502,467	502,467	502,467	502,467
R-squared	0.252	0.253	0.088	0.090

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 18: Geographic Distribution of Late Stage Diagnosis

VARIABLES	(1) Late Stage	(2) Late Stage	(3) Late Stage	(4) Late Stage
Atlanta	-0.00199	0.0178*	-0.0111***	0.00990
	(0.00286)	(0.00984)	(0.00291)	(0.00984)
Detroit	0.0166***	0.0166***	0.0114***	0.0253***
	(0.00227)	(0.00484)	(0.00231)	(0.00484)
San Francisco	-0.00949***	0.00521	-0.0105***	0.000592
	(0.00229)	(0.00429)	(0.00234)	(0.00431)
Hawaii	-0.0434***	-0.0121	-0.0304***	-0.00657
	(0.00364)	(0.0177)	(0.00477)	(0.0180)
Iowa	0.000802	0.0249	0.00958***	0.0364
	(0.00245)	(0.0348)	(0.00248)	(0.0348)
New Mexico	0.0177***	0.000835	0.0139***	-0.00109
	(0.00336)	(0.00572)	(0.00349)	(0.00577)
Seattle	-0.0105***	-0.0369***	-0.00352	-0.0273**
	(0.00238)	(0.0119)	(0.00240)	(0.0119)
Utah	0.0163***	-0.00311	0.0250***	0.00776
	(0.00340)	(0.0554)	(0.00341)	(0.0554)
Year of Diagnosis	-0.00724***	-0.00725***	-0.00727***	-0.00731***
	(7.54e-05)	(7.63e-05)	(7.63e-05)	(7.73e-05)
Age at Diagnosis	-0.00655***	-0.00656***	-0.00525***	-0.00530***
	(0.000348)	(0.000348)	(0.000352)	(0.000352)
Age at Diagnosis Squared	3.74e-05***	3.73e-05***	2.59e-05***	2.61e-05***
	(2.82e-06)	(2.82e-06)	(2.89e-06)	(2.89e-06)
Single			0.0262***	0.0263***
			(0.00228)	(0.00229)
Separated			0.0925***	0.0920***
			(0.00547)	(0.00547)
Divorced			0.0226***	0.0230***

			(0.00241)	(0.00242)
Widowed			0.0213***	0.0211***
			(0.00193)	(0.00193)
Black			0.0695***	0.0660***
			(0.00261)	(0.00271)
American Indian			0.0608***	0.0502***
			(0.0113)	(0.0116)
Chinese			-0.0175***	-0.0183***
			(0.00643)	(0.00650)
Japanese			-0.0459***	-0.0458***
			(0.00613)	(0.00615)
Filipino			0.0248***	0.0231***
			(0.00680)	(0.00682)
Hawaiian			0.0541***	0.0516***
			(0.00896)	(0.00897)
Hispanic			0.0440***	0.0425***
			(0.00389)	(0.00394)
Fixed Effects	No	Yes	No	Yes
Constant	0.722***	0.726***	0.670***	0.675***
	(0.0106)	(0.0109)	(0.0107)	(0.0110)
Observations	502,467	502,467	502,467	502,467
R-squared	0.023	0.025	0.026	0.027

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 19: Geographic Distribution of One-Year Survival

VARIABLES	(1) OneYear	(2) OneYear	(3) OneYear	(4) OneYear
Atlanta	0.000749	0.00178	0.00438***	0.00527
	(0.00128)	(0.00439)	(0.00129)	(0.00437)
Detroit	-0.00815***	-0.000167	-0.00553***	-0.00204
	(0.00101)	(0.00216)	(0.00103)	(0.00215)
San Francisco	0.00285***	0.00136	0.00333***	0.00292
	(0.00102)	(0.00191)	(0.00104)	(0.00192)
Hawaii	0.00440***	-0.0105	0.00338	-0.00889
	(0.00162)	(0.00791)	(0.00212)	(0.00798)
Iowa	0.00426***	-0.00298	0.00201*	-0.00621
	(0.00109)	(0.0155)	(0.00110)	(0.0154)
New Mexico	0.000400	0.00792***	-0.000199	0.00744***
	(0.00150)	(0.00255)	(0.00155)	(0.00257)
Seattle	0.00668***	0.00913*	0.00477***	0.00592
	(0.00106)	(0.00530)	(0.00107)	(0.00528)
Utah	-0.000611	0.0178	-0.00345**	0.0125
	(0.00151)	(0.0247)	(0.00151)	(0.0246)
Year of Diagnosis	0.000937***	0.000937***	0.00114***	0.00116***
	(3.51e-05)	(3.55e-05)	(8.06e-05)	(8.09e-05)
Age at Diagnosis	0.00911***	0.00911***	0.00766***	0.00768***
	(0.000155)	(0.000155)	(0.000166)	(0.000166)

Age at Diagnosis Squared	-9.20e-05***	-9.20e-05***	-8.66e-05***	-8.67e-05***
	(1.26e-06)	(1.26e-06)	(1.30e-06)	(1.30e-06)
Stage II	0.0216***	0.0216***	-0.0800***	-0.0797***
	(0.000861)	(0.000862)	(0.00375)	(0.00376)
Stage III	-0.00625***	-0.00601***	-0.0488***	-0.0485***
	(0.000946)	(0.000946)	(0.00403)	(0.00403)
Stage IV	-0.332***	-0.331***	-0.119***	-0.118***
	(0.00149)	(0.00149)	(0.00660)	(0.00660)
Radiation	0.0137***	0.0136***	0.0148***	0.0147***
	(0.000649)	(0.000651)	(0.000649)	(0.000651)
Single			-0.0115***	-0.0115***
			(0.00101)	(0.00102)
Separated			-0.0182***	-0.0181***
			(0.00244)	(0.00244)
Divorced			-0.0101***	-0.0101***
			(0.00107)	(0.00107)
Widowed			-0.0112***	-0.0112***
			(0.000859)	(0.000860)
Black			-0.0219***	-0.0207***
			(0.00116)	(0.00121)
American Indian			-0.00511	-0.00514
			(0.00503)	(0.00516)
Chinese			0.00183	0.00159
			(0.00286)	(0.00289)
Japanese			0.00723***	0.00686**
			(0.00273)	(0.00273)
Filipino			-0.00499*	-0.00476
			(0.00302)	(0.00303)
Hawaiian			-0.0156***	-0.0151***
			(0.00398)	(0.00399)
Hispanic			-0.00520***	-0.00514***
			(0.00173)	(0.00175)
Age + Stage II			0.00186***	0.00186***
			(6.61e-05)	(6.61e-05)
Age + Stage III			0.000775***	0.000776***
			(7.04e-05)	(7.05e-05)
Age + Stage IV			-0.00391***	-0.00391***
			(0.000111)	(0.000111)
Year + Age + Stage II			-1.07e-05***	-1.08e-05***
			(1.45e-06)	(1.45e-06)
Year + Age + Stage III			-2.70e-06*	-2.87e-06*
			(1.62e-06)	(1.62e-06)
Year + Age + Stage IV			3.58e-05***	3.55e-05***
			(2.49e-06)	(2.49e-06)
Fixed Effects	No	Yes	No	Yes
Constant	0.738***	0.736***	0.806***	0.803***
	(0.00478)	(0.00493)	(0.00585)	(0.00597)
Observations	502,467	502,467	502,467	502,467
R-squared	0.153	0.153	0.160	0.161

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 20: Geographic Distribution of Five-Year Survival

VARIABLES	(1) FiveYears	(2) FiveYears	(3) FiveYears	(4) FiveYears
Atlanta	-0.00336	-0.0335***	0.0108***	-0.0203**
	(0.00233)	(0.00802)	(0.00237)	(0.00800)
Detroit	-0.0205***	-0.0108***	-0.00999***	-0.0155***
	(0.00185)	(0.00394)	(0.00188)	(0.00394)
San Francisco	0.0127***	0.00204	0.0169***	0.0108***
	(0.00187)	(0.00350)	(0.00190)	(0.00351)
Hawaii	0.0275***	-0.0344**	0.0173***	-0.0381***
	(0.00297)	(0.0145)	(0.00388)	(0.0146)
Iowa	0.0125***	0.0128	0.00700***	0.00334
	(0.00200)	(0.0284)	(0.00201)	(0.0283)
New Mexico	-0.00485*	0.0148***	-0.00349	0.0157***
	(0.00274)	(0.00467)	(0.00284)	(0.00469)
Seattle	0.0219***	0.00663	0.0183***	0.000723
	(0.00195)	(0.00969)	(0.00196)	(0.00965)
Utah	0.00573**	0.00618	-0.000413	-0.00235
	(0.00277)	(0.0452)	(0.00277)	(0.0450)
Year of Diagnosis	0.00418***	0.00419***	0.00537***	0.00541***
	(6.42e-05)	(6.49e-05)	(0.000148)	(0.000148)
Age at Diagnosis	0.0306***	0.0306***	0.0272***	0.0272***
	(0.000284)	(0.000284)	(0.000304)	(0.000304)
Age at Diagnosis Squared	-0.000301***	-0.000301***	-0.000286***	-0.000286***
	(2.31e-06)	(2.30e-06)	(2.37e-06)	(2.37e-06)
Stage II	0.00505***	0.00521***	-0.108***	-0.107***
	(0.00158)	(0.00158)	(0.00687)	(0.00687)
Stage III	-0.157***	-0.157***	-0.269***	-0.268***
	(0.00173)	(0.00173)	(0.00737)	(0.00737)
Stage IV	-0.624***	-0.622***	-0.945***	-0.943***
	(0.00272)	(0.00272)	(0.0121)	(0.0121)
Radiation	0.0238***	0.0234***	0.0251***	0.0247***
	(0.00119)	(0.00119)	(0.00119)	(0.00119)
Single			-0.0279***	-0.0282***
			(0.00185)	(0.00186)
Separated			-0.0511***	-0.0505***
			(0.00446)	(0.00446)
Divorced			-0.0349***	-0.0351***
			(0.00196)	(0.00197)
Widowed			-0.0337***	-0.0335***
			(0.00157)	(0.00157)
Black			-0.0790***	-0.0760***
			(0.00213)	(0.00221)
American Indian			-0.0599***	-0.0571***
			(0.00921)	(0.00944)
Chinese			0.00882*	0.00761
			(0.00523)	(0.00529)

Japanese			0.0449***	0.0440***
			(0.00499)	(0.00500)
Filipino			-0.0220***	-0.0216***
			(0.00553)	(0.00555)
Hawaiian			-0.0456***	-0.0443***
			(0.00729)	(0.00730)
Hispanic			-0.0183***	-0.0194***
			(0.00317)	(0.00321)
Age + Stage II			0.00264***	0.00262***
			(0.000121)	(0.000121)
Age + Stage III			0.00179***	0.00179***
			(0.000129)	(0.000129)
Age + Stage IV			0.00584***	0.00583***
			(0.000203)	(0.000203)
Year + Age + Stage II			-3.90e-05***	-3.89e-05***
			(2.66e-06)	(2.66e-06)
Year + Age + Stage III			1.09e-05***	1.08e-05***
			(2.96e-06)	(2.97e-06)
Year + Age + Stage IV			-2.78e-05***	-2.82e-05***
			(4.56e-06)	(4.56e-06)
Fixed Effects	No	Yes	No	Yes
Constant	0.0637***	0.0624***	0.201***	0.199***
	(0.00874)	(0.00902)	(0.0107)	(0.0109)
Observations	502,467	502,467	502,467	502,467
R-squared	0.209	0.210	0.215	0.216

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 21: Geographic Distribution of Breast Cancer Mortality

VARIABLES	(1) mortality	(2) mortality	(3) mortality	(4) mortality
Atlanta	0.00286	0.0324***	-0.00717***	0.0222***
	(0.00232)	(0.00796)	(0.00235)	(0.00794)
Detroit	0.0179***	0.0264***	0.00929***	0.0277***
	(0.00183)	(0.00391)	(0.00187)	(0.00390)
San Francisco	-0.0105***	0.00483	-0.0148***	-0.00261
	(0.00185)	(0.00347)	(0.00189)	(0.00348)
Hawaii	-0.0306***	0.00387	-0.0222***	0.00967
	(0.00294)	(0.0143)	(0.00385)	(0.0145)
Iowa	0.00714***	0.0640**	0.00979***	0.0704**
	(0.00198)	(0.0281)	(0.00200)	(0.0280)
New Mexico	0.0166***	0.0114**	0.0158***	0.0110**
	(0.00272)	(0.00463)	(0.00281)	(0.00465)
Seattle	-0.0138***	-0.000280	-0.0124***	0.00157
	(0.00193)	(0.00961)	(0.00194)	(0.00957)
Utah	0.00973***	0.0436	0.0120***	0.0489

	(0.00275)	(0.0448)	(0.00275)	(0.0446)
Year of Diagnosis	-0.00863***	-0.00863***	-0.0122***	-0.0122***
	(6.37e-05)	(6.44e-05)	(0.000146)	(0.000147)
Age at Diagnosis	-0.00583***	-0.00583***	-0.00120***	-0.00122***
	(0.000282)	(0.000282)	(0.000302)	(0.000302)
Age at Diagnosis Squared	3.88e-05***	3.87e-05***	2.57e-05***	2.56e-05***
	(2.29e-06)	(2.29e-06)	(2.35e-06)	(2.35e-06)
Stage II	0.0117***	0.0117***	0.190***	0.190***
	(0.00156)	(0.00156)	(0.00681)	(0.00681)
Stage III	0.233***	0.232***	0.478***	0.477***
	(0.00172)	(0.00172)	(0.00731)	(0.00731)
Stage IV	0.624***	0.623***	0.885***	0.883***
	(0.00270)	(0.00270)	(0.0120)	(0.0120)
Radiation	0.00231*	0.00300**	0.000927	0.00157
	(0.00118)	(0.00118)	(0.00118)	(0.00118)
Single			0.0177***	0.0186***
			(0.00184)	(0.00185)
Separated			0.00700	0.00686
			(0.00442)	(0.00442)
Divorced			0.0171***	0.0178***
			(0.00195)	(0.00195)
Widowed			0.0107***	0.0108***
			(0.00156)	(0.00156)
Black			0.0550***	0.0544***
			(0.00211)	(0.00219)
American Indian			0.0199**	0.00994
			(0.00913)	(0.00936)
Chinese			-0.000459	0.00347
			(0.00518)	(0.00524)
Japanese			-0.0356***	-0.0333***
			(0.00495)	(0.00496)
Filipino			0.0113**	0.0122**
			(0.00549)	(0.00550)
Hawaiian			0.0209***	0.0208***
			(0.00722)	(0.00723)
Hispanic			0.00544*	0.00601*
			(0.00314)	(0.00318)
Age + Stage II			-0.00465***	-0.00463***
			(0.000120)	(0.000120)
Age + Stage III			-0.00471***	-0.00469***
			(0.000128)	(0.000128)
Age + Stage IV			-0.00702***	-0.00698***
			(0.000201)	(0.000201)
Year + Age + Stage II			8.31e-05***	8.26e-05***
			(2.63e-06)	(2.63e-06)
Year + Age + Stage III			2.14e-05***	2.10e-05***
			(2.94e-06)	(2.94e-06)
Year + Age + Stage IV			0.000145***	0.000145***
			(4.52e-06)	(4.52e-06)
Fixed Effects	No	Yes	No	Yes
Constant	0.485***	0.480***	0.324***	0.319***
	(0.00867)	(0.00894)	-0.0106	-0.0108
Observations	502,467	502,467	502,467	502,467

R-squared	0.195	0.196	0.202	0.203

*** p<0.01, ** p<0.05, * p<0.1

XV. County-Level Fixed Effects Reporting

A complete discussion and reporting of the county-level fixed effects from the models used in this research is available as a technical appendix. This appendix is available by request.