The Professor and the Coal Miner: The effect of socioeconomic and geographical factors on breast cancer diagnosis and survival outcome

Shelley Chen

Dr. Charles Becker, Faculty Advisor

Dr. Kent Kimbrough, Seminar Advisor

Honors thesis submitted in partial fulfillment of the requirements for Graduation with Distinction in Economics in the Trinity College at Duke University.

Duke University

Durham, NC 27708

2015

Abstract: Previous studies reported that patients who live farther from cancer centers do not necessarily experience delayed cancer detection and shortened survival. However, the results are biased because of the incomplete observation of patient survival, which cannot be properly accounted for with the multivariable regression model. In this thesis, I isolated the effect of the breast cancer patient's distance to a comprehensive cancer center on the stage of diagnosis and survival using the Cox Proportional Hazards model. I linked data from the Kentucky Surveillance, Epidemiology, and End Results 18, the Kentucky Life Tables, and the Kentucky Area Health Resource Files and identified 37654 patients diagnosed with breast cancer. I estimated the effect of distance on marginal probability of cancer mortality, controlling for non-cancer related death, socioeconomic status, and demographic factors in patients. After controlling for covariates, travel distance between the patient and the nearest comprehensive cancer center was statistically significantly on the breast cancer mortality probability, but not on the stage of diagnosis. In the Kentucky population, patients who were located farther from comprehensive cancer centers experience an increased marginal probability of mortality (proportional hazard = 1.004; 95% CI: [1.000502 1.007311]). The linkage of SEER 18 and AHRF data provided more comprehensive information on the socioeconomic risk factors of cancer mortality than past study datasets. For the stage of diagnosis, a low physician to population ratio and high county-level Medicaid coverage were associated with more advanced stages of diagnosis. In turn, a more advanced stage of diagnosis, lower physician to population ratio, and identification as African American increased the marginal probabilities of mortality.

Acknowledgements

I would like to thank Dr. Charles Becker for his continued encouragement and support for my research and his invaluable insight into the methods of health economics research. I would also like to thank Dr. Kent Kimbrough for his guidance and comments, without which this thesis would not have been possible. Finally, I would like to thank Joel Herndon and Mark Thomas from the Duke Data and Visualization Lab for their immense help with acquiring data and utilizing GIS.

I. Introduction

Cancer has been a topic of intense interest for economic research due to its impact on healthcare costs and economic ramifications on society. It is considered a disease of old age, where the biological mechanisms that monitor cell growth gradually deteriorate and certain cells begin to divide at an abnormal rate. Today, cancer is the second leading cause of mortality in the United States. As more people survive to old ages due to advances in science and medicine, the prevalence of cancer is expected to increase. Globally, the number of new cases of cancer is expected to increase by 57% worldwide in the next two decades, from a current incidence of 14 million annual cases to 22 million (WHO). In the same period, the annual mortality rate is expected to increase from 8.2 million to 13 million.

Previous public health research showed that from 1962- 1991, the cancer incidence and mortality rate of all cancers had both increased steadily (Bailar and Smith, 1986). Data from the *Surveillance, Epidemiology, and End Results National Registry 9* (SEER) showed that the mortality rate for all cancers began to decrease in 1991 until 2006, but incidence rates continued to increase (Lichtenberg 2010). The improvements in cancer survival were primarily attributed to advanced cancer imaging diagnostics and better treatment, but came with increasing cost to the healthcare system (Lichtenberg, 2010; Sun, 2010).

The economic cost of cancer has been significant and increasing. The National Institutes of Health estimates the costs of cancer in the U.S. to be $216.6 billion in 2009, representing an increase from a cost of $172 billion in 2002 after adjusting for inflation *(Cancer Facts and Figures 2014)*. A further breakdown of the cost from 2009 shows that $86.6 billion are incurred from direct healthcare expenditures while $130 billion are due to indirect costs – loss of productivity due to premature mortality. The indirect cost of cancer represents a 1.5 fold addition to the direct cost of

cancer. In light of the economic costs of cancer, and the upwards trend of cancer incidence, researchers have conducted numerous studies to evaluate the success and shortcomings in our approach towards cancer.

Despite the improvement in cancer survival outcomes since 1991, not all patients benefits equally. Research shows that worse cancer outcomes can be clustered by race, health insurance status, and other socioeconomic factors (Ward, 2004). Furthermore, physical distance from cancer centers and travel time also affect cancer survival. It has been shown that a relationship exists between spatial location and cancer diagnosis and survival, but that relationship remains inconclusive (Gill, 2002; Gunderson 2013; Massarweh 2014; Wang et al, 2008).

While the influence of socioeconomic status on cancer outcomes has been widely studied, research into the influence of spatial location has just started to gain momentum. I am interested in how cancer outcomes vary with the patient's distance from accessing high-quality treatment at comprehensive cancer centers. I link the Kentucky *Surveillance, Epidemiology, and End Results National Registry 18* (SEER 18), which contains longitudinal data from 2000 – 2010 with the 2014 Area Health Resource Files (AHRF) to study this relationship. SEER 18 contains cancer patient survival and the 2014 Area Health Resource Files contain county-level demographic and health resource data from 2005-2012. I hope to elucidate: 1) the relationship between distance and the patient's stage of diagnosis for breast cancer, 2) the effect of a patient's distance from the nearest cancer center on their cancer mortality rate conditional on the stage of diagnosis, and 3) the effect of race and socioeconomic factors on breast cancer mortality conditional on all other factors.

I draw conclusions from the results from breast cancer in Kentucky to similar types of cancers in other geographical areas. Understanding these relationships will allow non-profit organizations, healthcare professionals, and policymakers to be more cognizant of the impact of

different factors when implementing solutions to improve cancer outcomes. The result of my research hopefully will contribute to the field of research of health economics and access to care and motivate changes for improving welfare for cancer patients of all populations. In the rest of the thesis paper, previous research on healthcare access and cancer outcome are discussed in section III, the theoretical framework of spatial factors and healthcare outcomes are discussed in section IV, data collection and utilization are discussed in section V. Section VI discusses the methodology; section VII and VIII present the results and discussion.

II. Background

"To a large extent, factors such as where we live, the state of our environment, genetics, our income and education level, and our relationships … have considerable impacts on health, whereas the more commonly considered factors such as access and use of health care services often have less of an impact" (World Health Organization 2014). These factors are considered social determinants of physical health, which are closely related to socioeconomic risk factors of health. Low education levels are linked with higher stress and poor health. Cutler (2009) found that higher levels of education correlated to greater cognitive ability and greater health seeking behavior. Lange (2010) found that educated women are more willingly adopt preventative behaviors. Physical environment consists of resources such as safe water and clean air, safe communities, healthy workplaces, and ease of healthcare access. Social support networks, including families, friends and communities, are linked to better health. Genetics plays a key part in determining lifespan, healthiness and the likelihood of developing certain illnesses. Culture, customs, and tradition all influence health as well.

Patients with more risk factors can expect to experience worse outcomes than the average population. Additionally, they may not receive an optimal amount of prevention, diagnostics, and

treatment due to financial and geographical constraints. They may be more likely to present at an advanced stage of cancer at an emergency room or cancer center when symptoms finally arise and require more complex and costly treatments. Furthermore, as cancer progresses to more advanced stages, the patient unnecessarily suffers greater pain and debilitation. Ward (2004) found that five-year survival for all cancers combined is 10 percentage points lower among persons who live in poorer than in more affluent census tracts. Bradley (2002) and Massarweh et al. (2014) found that the lack of insurance is correlated with lower survival for cancer patients. Interestingly, both found no association between race or ethnicity and cancer survival. Finally, longer distance of travel to get care from a cancer center is also correlated with lower survival rates (Gill et al. 2002).

Health disparity is a term that characterizes health outcomes of certain populations lagging behind the improvements experienced by the general population. The elimination of disparities is a goal of the American Cancer Society and many other health organizations, and is defined as "a reduction in cancer incidence and mortality and an increase in cancer survival among socioeconomically disadvantaged people to levels comparable to those in the general population" (Ward et al., 2004).

The goal of my research is to analyze the effect of distance on cancer survival and stage of diagnosis for breast cancer. I analyze stage of diagnosis and survival time for 37,645 breast cancer patients residing in Kentucky diagnosed after 2000. Given the results of previous research, it is important to control for variation in state characteristics, insurance coverage, socioeconomic, and demographic factors in my research. The SEER 18 Registry was chosen because it is the most complete and recent registry for stage of diagnosis and survival data. Breast cancer is chosen because it is one of the leading causes of death in women in the United States and occurs in people of all ages (SEER 18), but at the same time is not an irrevocable death sentence. Breast cancer has a

national 5-year survival rate of 89%, but this survival is highly dependent on diagnosis and treatment. Of particular importance, there is a lot of variability in survival, which is advantageous for analysis. It has very clear stages of diagnosis, significant progress in screening and treatment, and there is a wide collection of literature regarding survival rate and various socioeconomic factors.

Kentucky is an ideal state because of the relative racial homogeneity (the majority of the population is white or black) and high degree of economic homogeneity within counties, which allows us to be closer to using iid samples. An initial analysis of the SEER 18 data reflects the racial homogeneity of Kentucky, where 92.98% were white, 6.61% were black, and close to 0% were Asian or American Indian. The racial homogeneity can control for genetic variation in breast cancer risk, which varies by race (Gomez, 2014). Its counties are considered economically homogeneous due to the largely rural characteristics and, in many cases, reliance on the mining industry (Strickland 1999). Kentucky also has a small number of comprehensive cancer centers and concentrates medical talent in its few metropolitan areas. This helps control for variation among cancer centers and provides adequate variability in the measured distances from the patient to the closest cancer center.

For my research, breast cancer cases diagnosed between 2000 and 2006 are included to have the greatest number of observations for 5-year survival and to control for year- specific effects like more advanced technology. In addition, diagnoses from earlier years give the greatest timespan for analysis and allow us to observe much more complete mortality data, because the average length of survival is 8.8 years after diagnosis (*SEER 18).* The investigation of distance to care on cancer will allow us to see where improvement remains to be made and motivate future policies that promote equity in healthcare in both rural and urban populations.

Cancer Statistics Definitions

Cancer is frequently evaluated by the rate of incidence, absolute or relative 5-year survival rate, and mortality rate. They give us meaningful information related to an individual or a population's risk of developing certain types of cancer or one's life expectancy after diagnosis and can be used to compare disparity across populations.

Cancer incidence rate is the number of new cases of cancer diagnosis out of the total population for a given year, reported as cases per 100,000 people:

$$\textbf{Incidence rate} = (\textbf{New cancers} / \textbf{Population}) \times 100{,}000 \quad (1)$$

The relative 5-year survival rate is the estimated percentage of patients who are still alive 5 years after their date of diagnosis relative to the general population, reported as cases per 100,000.

$$\textbf{5 year survival} = \%\ \textbf{diagnosed surviving} / \%\ \textbf{General Population surviving} \times 100{,}000 \quad (2)$$

The rate of mortality is the total number of cancer related death out of the total population in a given year, usually reported as number of people in 100,000:

$$\textbf{Mortality Rate} = (\textbf{Cancer Deaths} / \textbf{Population}) \times 100{,}000 \quad (3)$$

Bailar et al. (1986) reasoned that absolute mortality rate is the most reasonable measure of success in cancer outcome rather than 5 year survival rates, which are confounded by false diagnoses. Patients with harmless lesions who are diagnosed as serious cancer cases may inflate the 5 year survival rate. To investigate the problems introduced by changing patterns of diagnosis on survival rates, Lichtenberg (2010) and Asadzadeh (2011) further examined the relationship between five-year survival rate and mortality rate using data from NCI's Surveillance, Epidemiology, and End Results Program (SEER), controlling for the stage of diagnosis, and found that there is a positive

correlation between 5-year survival rates and decreasing mortality. For future research, Lichtenberg (2010) recommended that the 5-year survival rate be combined with other measures of success for more a more accurate picture. Thus, I will look at the 5-year survival rate as a measure of health outcome and as well as cancer mortality as a continuous measure.

III. Literature Review

Due to the growing economic burden of cancer in the United States in the face of limited resources, there is an increased need to understand how various factors affect the welfare of cancer patients and how the effect varies by population so that better policies and programs can be put into place. Using data from the National Center for Health Statistics, Bailar and Smith (1986) found that from 1962 to 1982, the crude incidence rate of all cancers combined had an upward trend. They hypothesized that this was either due to failure of prevention or due to changing standards of diagnostics. Similarly, the mortality rate for all cancer types combined increased by 25 percent despite rapid growth in cancer research and treatment delivery over the 20 years. The results suggested that the improved treatment alone was not sufficient in lowering the disease burden of cancer, but rather that better prevention is needed.

More recent research has been conducted to dissect the causes of the improvement in cancer outcomes. Using SEER data, Lichtenberg (2010) noted that after 15 years of steady increase in the mortality of all cancer types combined, mortality declined for the first time, by 17.2% between 1991 and 2006. Using a difference-in-difference model, the study found that the greatest factor in the decline of cancer was innovation in imaging that led to better diagnosis and earlier detection. Another 34% of the 17.2 percentage point decline is attributed to a decline in the incidence of cancer and drug innovation.

Ideally, the impact of society's innovation in pharmaceutics and biotechnology should be distributed evenly to all cancer patients so that everyone has an increased chance of overcoming mortality from cancer. In reality, however, different populations have varying levels of access to these services. This may be based on their average levels of education, socioeconomic status, spatial factors, and medical talent available in the area. Thus, my analysis will try to control for these variations.

Research on the relationship between physical distance from care and cancer survival reflects disparity in access to cancer care. Gill et al. (2002) conducted a study on the correlation between distance to major cancer centers and gastrointestinal cancer survival in New Zealand patients. Using the domicile code associated with each patient, the distance of the center of each domicile from the major cancer centers was measured and categorized into groups by distance. Socioeconomic status variables, including income, living space, qualifications, employment, and transport, were combined into a proxy called the NZDep96 deprivation score. Using a multivariate regression, they found that only increasing age was associated with poorer prognosis of gastrointestinal cancer, rather than distance from major cancer center or socioeconomic factors. Geographical isolation seems to have no effect on survival from GI cancer. They reasoned that it could be because patients in remote areas acknowledge their distance and have arrangements for direct referral pathways to cancer centers. Although a correlation does not exist for GI cancer, a correlation between distance and survival may be present in other types of cancer.

Contrary to Gill et al (2002)'s findings, Kim (2013) studied the correlation between the distance women traveled to a clinic and having an abnormal mammogram for women living in poverty. Kim (2013) used data from the Illinois Breast and Cervical Cancer Program (IBCCP) that provides free breast cancer screening services to women without health insurance to control for

much of the variation in socioeconomic and demographic factors. It was found that travel distance to a clinic was positively correlated with an abnormal mammogram. Once distance to a clinic is added as an explanatory variable, the effect of racial ethnicity disappeared. In consensus with Kim (2013), Gunderson et al. (2013) also found that farther travel was not correlated with survival rate among patients with cervical cancer. He evaluated the impact of distance from residence to treatment center on the survival and recurrence rate of cervical cancer. Furthermore, he found that non-Caucasians were less willing to travel more than 30 minutes compared to Caucasians.

The relationship between physical distance to clinics and stage of diagnosis has been examined as well. Massarweh et al. (2014) studied the correlation between cancer stage diagnosis and travel time to healthcare for colon cancer patients across the United States. They also explored whether the relationship between the travel distance and the stage of diagnosis relationship was a national phenomenon or a regional one. Using stage of diagnosis data from the SEER-Medicare Registry, Census 2000 data, and a multivariable two-level hierarchical regression, they found that patients with longer travel distance were more likely to present with more advanced diagnosis. Ethnicity also played a role in a later stage of diagnosis - black patients were more likely to present at later stages of diagnosis than white patients. Finally, those who are uninsured or using Medicaid had higher odds of presenting with a more advanced diagnosis than those who were privately insured, (Bradley, 2002; Massarweh, 2014). Their results show that age, lower education level, rural residence, and lower income are correlated with a later stage of diagnosis. The sample size in this study was small, so my research will improve upon this to determine the effect of distance to cancer centers on the stage of diagnosis more accurately.

There is evidence that travel time to mammography facilities can affect the stage of diagnosis outcome, which is relevant for my regression equation. Mammography facilities are FDA approved

facilities that conduct screening and diagnosis, which includes hospitals, outpatient departments, clinics, radiology practices, or physicians' offices. The cost of mammography is usually completely covered by national awareness programs. Huang (2009) found that in the state of Kentucky, advanced diagnoses are correlated with longer average travel distances than early stage diagnoses after controlling for age, race, insurance, and education, but not income. The odds of a later stage of diagnosis are significantly higher for women who lived 15 miles from a facility compared to those who lived only 5 miles from a facility. The nearness to mammography facilities promote earlier detection through screening and diagnostic imaging, and is expected to correlate to an earlier stage of diagnosis and better survival probability. I will control for the effect of distance to mammography facilities by considering the access to mammography facilities within a reasonable distance of 30 miles (Wang and Leu 2004). I also plan to improve upon Huang (2009) by including income data and other relevant medical resource data such as physician availability in my regression model.

The previous literature shows that numerous factors such as the distance to mammography facilities, distance to cancer centers, and distance to clinics all need to be controlled for. Given the contradicting results from previous literature, my research will give me the opportunity to verify the relationship between physical distance and health outcomes. Distance to care and its influence on the stage of diagnosis and cancer survival is a growing subject of research and still needs to be more fully addressed by including all appropriate control variables. Previous research on the relationship between distance and care has attempted to include cancer centers near the patient, but did not include all the facilities that affect the patients. I will study the effect of distance to the closest mammography facilities, medical schools, and cancer centers on the stage of diagnosis and cancer mortality for the patient.

Finally, I plan on using a more comprehensive dataset and a multiple equation methodology that parses out the indirect effect of stage of diagnosis on survival rather than using one regression of physical distance on survival alone. Because later stage of diagnosis correlates with a higher mortality rate, I examine how my controls and independent variables affect the stage of diagnosis across different geographical regions and populations. Moreover, it is necessary to establish the effect of demographic, socioeconomic factors, and physical distance on the stage of diagnosis. Then, this indirect effect on cancer survival will be taken into account when the relationship between survival and physical distance and stage of cancer is studied.

For the dependent variable, I will use the 5-year survival rate, which is the most commonly used measure for survival. I will also use the length of time of survival after diagnoses, which provides more detailed information. Using the most recent dataset from the Surveillance, Epidemiology, and End Results 18 (SEER 18) registry and focusing on the state of Kentucky, I will compare my analysis results on the effect of physical distance to the analysis in previous literature after controlling for socioeconomic status, income, and distance to cancer related facilities.

IV. Theoretical Framework

The framework for this research lies in the connection between stage of diagnosis[1] and survival as well as their relationship to the ease of accessing comprehensive cancer centers. Based on the initial data analysis from the SEER 18, the stage of diagnosis and months of survival have a correlation of -0.3979. A preliminary regression on survival and stage of diagnosis shows a reduction

[1] SEER defines 4 stages of diagnosis for breast cancer: *in situ*, localized, regional, and distant stages. *in situ* stage is characterized by a tumor that is not yet malignant. The localized stage is characterized by a tumor that has expanded past the epithelial tissue from which the first cancerous cells arose, but is limited to a specific organ. The regional stage is characterized by a tumor that has spread beyond the organ into the lymph nodes. Finally, the distant stage refers to a neoplasm that has expanded to numerous areas of the body including organs and lymph nodes.

of survival by 20 months for each later stage of diagnosis (*SEER 18*). It is also intuitive that the earlier and more accurately an abnormality can be identified, the better the chance at receiving treatment and surviving. Thus, a later stage of diagnosis is expected to reduce survival because of the elevated penetration of the cancerous tumors.

The stage of diagnosis and the survival probability are both related to the patient's spatial location and ability to access services at comprehensive cancer centers. It is known that NCI-designated comprehensive cancer centers wield a greater selection of clinical trials and treatments, highly skilled oncologists, and cutting-edge technology. Studies have shown that survival outcomes are better at such centers (cancer.gov). Medical schools and newly trained medical talents tend to be closely located to these comprehensive cancer centers, which may lead to greater advancements of medical knowledge around these centers and a greater availability of physicians. It is possible that patients located in proximity to metropolitan areas enjoy both a higher availability and higher quality of physicians, leading to better survival outcomes.

For Kentucky, the only comprehensive cancer center is located in the urban area of Lexington, and the two major medical schools are in Lexington and Louisville. Using data from the ARHF, I calculated the physician to population ratio for each county and found that the most urban areas, Lexington and Jefferson, have the highest density of physicians (*ARHF 2014*). Interestingly, and most surprisingly, the county with the highest physician to population ratio also contains the state's only comprehensive cancer center.

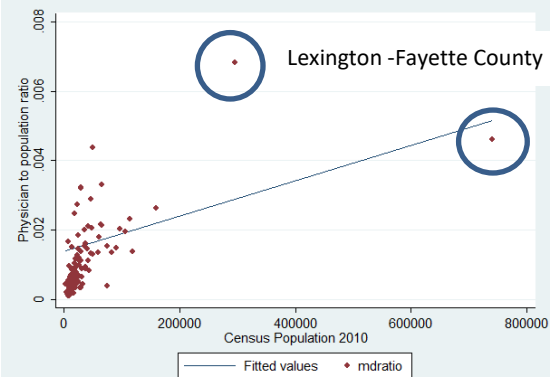


Figure 1. Scatterplot of physician density *vs.* county population. Average physician density is 2 per thousand.

The distance from the patient's residence to the NCI-designated comprehensive cancer centers varies significantly, which can affect the cost of getting treatment, healthcare-seeking behavior, and health outcome. To visualize the variation in distances, cancer centers surrounding Kentucky and county population centers are geocoded and plotted using GIS and ArcMap, and straight line distances from the each county center to the closest cancer center were measured. The mean distance from a cancer center was 65.19 miles, and the standard deviation is 30.9176 miles.

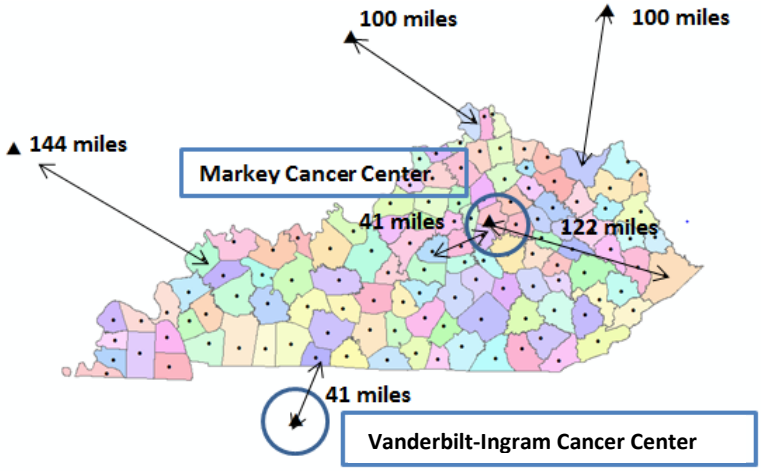


Figure 2. Map of Kentucky showing variations in the distance between patient location and nearest comprehensive cancer centers (represented by black triangles). Average distance to the cancer centers is 65 miles.

As seen, out of the 5 total NCI-designated cancer centers (shown by black triangles) in the vicinity of Kentucky, only 2 are within reasonable travel distance of Kentucky residents – the Markey Cancer Center in Lexington and the Vanderbilt-Ingram Cancer Center in Nashville, TN. In the examples, residents from a county that is 41 miles away from Nashville may choose to go to Vanderbilt to seek care because it is closer to them than the Markey Cancer Center. On the other hand, residents who are 41 miles from Lexington will instead choose the Markey Cancer Center.

However, distance is an impediment to getting treatment from an NCI designated cancer center when a county is as far as 122 miles away from the closest place, which is least 4 hours of travel time for each roundtrip visit. For these residents, there may exist the possibility of relocating closer to the cancer center for treatment and will be considered as I move forward with the research.

The variation in the availability and quality of medical resources to the population and the patients' ability to seek care are expected to affect the stage of diagnosis and survival rates of cancer. The hypothesis is that patients who are closer to cancer centers will have better survival outcomes due to medical resources and accessibility.

V. Data

This research will draw on data from the Surveillance, Epidemiology, and End Results Program 18 (SEER 18) from the National Cancer Institute. The SEER 18 Registry is one of the several national cancer registries under the SEER program, with others being SEER 9, SEER 11, SEER 13, and so on. The SEER 18 Registry is the most updated and contains the widest geographical range of stage of diagnosis and survival data on individual cancer cases across the U.S. from 1973-2010. It covers the states of Kentucky, California, New Jersey, and Alaska, in addition to states from all previous SEER registries (New Mexico, Connecticut, Iowa, Washington, Georgia, Hawaii). Each cancer case is matched with a county of residence, so every patient from the same county will share the same physical distance to healthcare facilities. The registry contains 149 variables, and important variables are found in the table below. Summary statistics for pertinent variables are found in the appendix.

Table 1. Variable Names and Descriptions

Variable	Description
Distance_cancer	Distance from nearest comprehensive cancer center
Distance_mammography	Distance from nearest mammography center
stage	Stage of diagnosis: 0= in situ or stage 1, 1=localized or stage 2, 2=regional or stage 2, 3=distant or stage 4
%high school educated	Proxy for education level by county
median income	Median household income by county
%privateinsurance	% of privately insured patients by county
%medicaid	% of Medicaid insured patients by county
Physician-population ratio	Physician to population ratio
age_dx	Age at diagnosis
non-cancer death	Probability of non-cancer related death at each age
sex	Gender: 1=female
year	Year of diagnosis: dummy variables for year
African-American	African-American race
Other_race	Other ethnicities
mar_stat	Marriage Status

Age at Diagnosis

The age at diagnosis ranges from 19 to 102 years old, with a mean of 60.

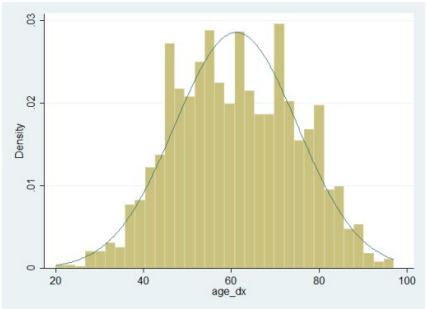


Figure 3. Distribution of age at diagnosis from 2000-2011 for Kentucky

Stage of Diagnosis Variable

The stage of diagnosis is reported as discrete values, each representing a stage of diagnosis. The average stage of diagnosis is 1, which falls between the in situ and localized stage. The localized stage, which is the second of the four defined stages, is the most frequent stage of diagnosis.

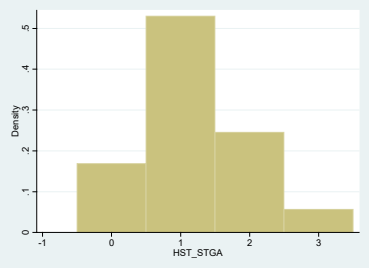


Figure 4. Distribution of stage of diagnosis from 2000-2011 for Kentucky (1: in situ stage; 2: localized stage; 3: regional stage; 4: invasive stage)

Year of Diagnosis Variable

The year of diagnosis ranges from 2000 to 2010 and is converted into dummy variables. A graph of number of new diagnosis each year in Kentucky shows an increasing incidence, in contrast to the stable level of incidence for the whole country. Given the population for Kentucky in 2011, which is about 4,339,357, the incidence rate for breast cancer as of 2011 is calculated to be 88.2 per 100,000. This is less than the national average incidence rate, which is 124.6 per 100,000. This may be due to the fact that not everyone who has cancer is being tracked by the SEER registry.

Survival Variable

The survival variable is reported as survival months in the registry. It is also converted into a dummy variable for 5 year survival - 1 if the patient survived more than 60 months (5 years) after their diagnosis and 0 otherwise. From 2000-2010, the one year survival rate was 95.07%. From 2000-2006, the 5 year survival rate was 79.20%, which includes mortality from cancer and non-cancer

related causes. This is exactly 10% lower than the national average for breast cancer, which suggests that health disparities exist in Kentucky compared to the rest of the U.S.

Sample data from 3000 patients diagnosed in 2000 show the variability in cancer survival. The average length of survival is 105.2 months, or 8.8 year after diagnosis (SEER 18). 77.21% of the patients survived past 5 years and 62% of the patients survived for more than 10 years, meaning that they are currently alive (SEER 18). The survival length and survival rate imply that breast cancer is highly treatable, but can be fatal due to delayed diagnosis and treatment (SEER Facts).

Figure 5. Distribution of survival time among breast cancer patients diagnosed in 2000

However, as with many public health datasets, the SEER 18 Registry lacks information on the socioeconomic status of the participants. However, this can be addressed, albeit imperfectly, by linking patient data to demographic data by the FIPS county code, which is a five-digit Federal Information Processing Standard (FIPS) code that uniquely identifies counties and county equivalents. This is chosen because it is the most specific level of geographical identification in the SEER registry and the AHRF dataset.

To be able to control for socioeconomic factors and other demographic factors in the analysis, demographic data and insurance coverage data from the Area Health Resource Files (an expansion upon the Census data) can be linked to the SEER 18 patient data based on the FIPS

county codes. The Area Health Resource Files include 6,848 variables on socioeconomic data such as income, education, insurance status, Medicaid enrollment, physician to population ratio, number of hospitals, and so on.

Another set of variables not in the SEER 18 data are physical distances between the patient and cancer screening facilities and cancer centers. The U.S. Food and Drug Administration regulates 9400 mammography facilities across the U.S., with 162 in Kentucky. Similar to Wang (2010), I obtained a comprehensive database of cancer screening facilities, clinics, and cancer centers. Using the Geographical Information Systems (GIS) technology, I geocoded the physical addresses of all hospitals and cancer centers in Kentucky and those near the borders of Kentucky and mapped them. Once this layer of map is completed, I added a second layer with the county population centroids, which is a proxy for the patient's location because SEER 18 only identifies each patient's county of residence. Finally, I measured the physical distance between each patient and the nearest cancer facilities to obtain distance independent variables and appended this to the existing SEER 18 registry and AHRF data by patient ID number.

Finally, I recognized patient mortality can result from both cancer and non-cancer related problems depending on the patient's age. At a very old age, patients are more likely to have mortality hazards from other causes, such as physical injuries or metabolic problems. To control for non-cancer related death, I linked data on the probabilities of death from all other causes with the SEER-AHRF data based on the age variable. The probabilities were obtained from the Census Life Tables.

Principle components analysis is used to reduce dimensionality in the data by finding the vectors that explain the greatest variation in the data. In this data set, PCA identifies three main components. The first component consists of the socioeconomic level of the county as represented by median income and education level, the second component was human capital as represented by

population and physician to population ratio, and the third component consists of the healthcare service access as represented by the distance from the nearest cancer center and the stage of diagnosis at which the patient presents (Appendix).

One of the limitations of this final merged dataset is the accuracy of the FIPS County codes as geographical divisions, which is the finest granularity provided by the SEER 18 Registry data. Ideally, it would be better to use data on the block-group or census tract level. Despite this limitation, however, this dataset will provide the most comprehensive data given what is available. It provides the best source of longitudinal data for cancer progression in patients. It is intensive in coverage of breast cancer cases as well as socioeconomic and demographic characteristics of each breast cancer case. Additionally, many previous studies utilized the method of linking various SEER datasets with other demographic data at the county level in this manner because of the advantages of the SEER dataset.

VI. Empirical Specification

Cox proportional hazards and probit regression models are used to determine how travel distance and socioeconomic factors drive the individual survival probability from breast cancer. Ordered probit regressions are used to measure how socioeconomic factors drive the patient's stage of diagnosis for breast cancer, which in turn affects survival probability. In probit regressions, survival is a discrete, binary variable, where 1 indicates the patient is alive in the 5^{th} year after diagnosis:

Five Year Survival = $a + \beta_1$ *(Stage of Diagnosis)* $+ \beta_2$ *(Spatial Factors)* $+ \beta_3$ *(Demographic and Socioeconomic Factors)*$+ \beta_4$ *(Year of Diagnosis Dummies) + interaction terms* $+ e$

Spatial factors are the distance to mammography facilities or to comprehensive cancer centers, the socioeconomic indicators include the physician to population ratio, household income, education level, and insurance status. The demographic factors include age, race, marital status, and sex. Interaction terms include age interactions with distance and race.

To study survival time as a continuous variable, the Cox proportional hazards model is more desirable because it accounts for the censoring of survival data among patients diagnosed at different times and varying probabilities of death at various time points after diagnosis. The Cox model uses a maximum likelihood estimation to determine $h_0(t)$, the baseline hazard function representing the probability of death if all variables are equal to zero. The hazard h(t), which gives the marginal probability that an individual will experience death with a change of 1 in the covariate, is regressed on the stage of diagnosis and other explanatory variables. The Cox proportional hazards regression is represented by:

$$\mathbf{h(t)} = h_0(t)\ [\exp(\beta_1\ (\textit{Stage of Diagnosis}) + \beta_2\ (\textit{Spatial Factors}) + \beta_3\ (\textit{Demographic and Socioeconomic Factors}) + \beta_4\ (\textit{Year of Diagnosis Dummies}) + \textit{interaction terms} + e)]$$

In ordered probit regressions, the stage of diagnosis is an ordinal variable from 0 to 3 to represent more advanced stages:

$$\textit{Stage of Diagnosis} = a + \beta_1(\textit{Stage of Diagnosis}) + \beta_2\ (\textit{Spatial Factors}) + \beta_3\ (\textit{Demographic and Socioeconomic Factors}) + \beta_4\ (\textit{Year of Diagnosis Dummies}) + \textit{interaction terms} + e$$

VII. Results

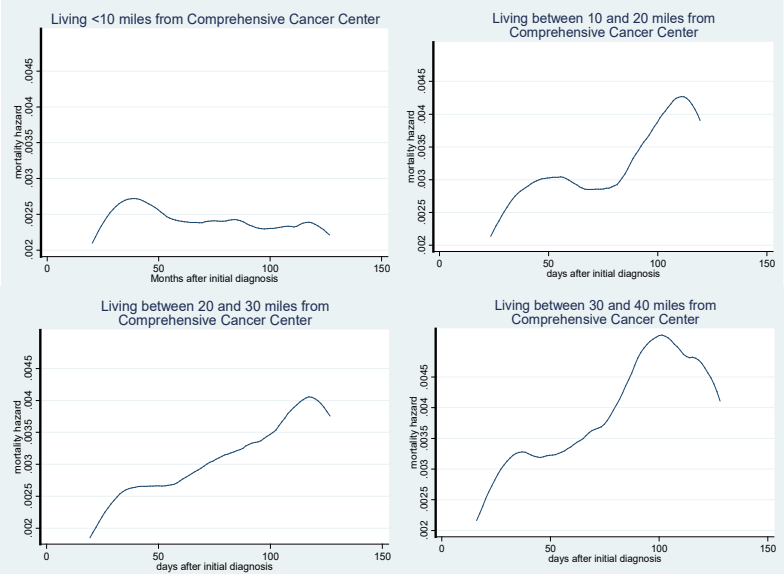


Fig 6. The smoothed mortality curves show the instantaneous mortality probability for each month after the patient is diagnosed, spanning 150 months. The cumulative mortality probability is reflected by the area under the curve. Non-cancer related death is controlled for.

The initial survival curve analysis of the effect of distance gives astounding results that show disparity in survival based on distance. When I group the patients by their distances from the nearest comprehensive cancer centers and visualize the instantaneous mortality probabilities, I see huge variations in the mortality risk and the progression of mortality probability (Fig. 6). Comparing the groups shows that the cumulative probability of death is higher for populations who live farther from cancer centers, and the instantaneous probability of death at any month after diagnosis is also higher. For patients who live within 10 miles, the pattern is that the mortality probability is the greatest in the first two years after diagnosis, but gradually decreases and stabilizes afterwards. The instantaneous probabilities of death range from 0.20% to 0.27%, which is equivalent to the probability of natural death for a 43 year-old. However, for those who live farther than 10 miles away, their initial instantaneous probabilities of death range from 0.20% to 0.55%， and has a

pattern of rapid increase after diagnosis. For patients who live 10-30 miles from a cancer center, the peak value is around 0.40%. For those who live more than 30 miles away, the peak value is 0.55%, which is equivalent to the probability of natural death for a 52 year-old.

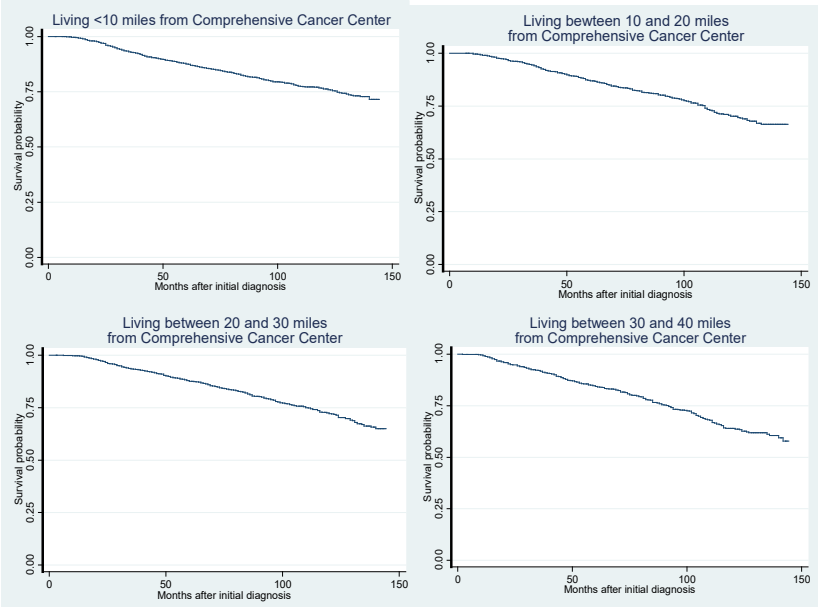


Fig 7. Kaplan Meier survival estimate graphs for breast cancer patients grouped by distance from cancer centers. They show the percentage of the population that is alive at a given month after initial diagnosis.

Large variations also exist in the survival rate for breast cancer patients when I group patients by their distances from the nearest comprehensive cancer centers (Fig. 7). Comparing the groups shows that the survival rate is lower for populations who live farther from cancer centers. Patients who are within 10 miles of a cancer center have a 74% survival rate. Patients who are between 10-20 miles have a 64% survival rate. Patients who are between 20-30 miles have a 64% survival rate. Patient who are between 20-30 miles have a 60% survival rate. This initial graphic analysis provides intuition for the impact of the patient's location from cancer centers and shows the apparent disparities in outcomes. Next, I use regressions to clarify how much of the disparities in survival outcomes is attributable to the patient's distance from cancer centers by controlling for other covariates as well as the probability of non-cancer related death.

Table I. Estimations of the effect of distance to a comprehensive cancer center and stage of diagnosis on survival probability for breast cancer patients diagnosed from 2000-2011 in Kentucky using Cox Proportional Hazards Model, Probit Models, and Logit Model

	Cox b/se	Cox2 b/se	Probit 1 b/se	Probit 3 b/se	Logit b/se
main					
stage== 1.0000	0.3564***	0.3545***	-0.2476***	-0.1650***	-0.4815***
	(0.041)	(0.041)	(0.036)	(0.029)	(0.071)
stage== 2.0000	0.8967***	0.8969***	-0.7329***	-0.5431***	-1.3399***
	(0.042)	(0.042)	(0.038)	(0.032)	(0.073)
stage== 3.0000	1.5591***	1.5606***	-2.1237***	-1.9403***	-3.6585***
	(0.046)	(0.046)	(0.056)	(0.052)	(0.102)
Distance_cancer	0.0042*	0.0039*	0.0008	0.0013	0.0036
	(0.002)	(0.002)	(0.000)	(0.002)	(0.003)
Distance_mammography	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
AGE_DX	-0.0650***	-0.1223***	-0.0272	-0.0087	0.0105
	(0.007)	(0.011)	(0.030)	(0.057)	(0.112)
age2	0.0008***	0.0013***	0.0010*	0.0004	0.0007
	(0.000)	(0.000)	(0.000)	(0.001)	(0.002)
agedist	-0.0001*	-0.0001*		-0.0000	-0.0000
	(0.000)	(0.000)		(0.000)	(0.000)
constant			1.0772	-2.7593*	0.7039
			(0.912)	(1.128)	(2.243)
chi-sqr	3732.006	3768.202	3969.536	31543.525	3968.700
dfres					
BIC	162236.9	162211.2	16464.5	19847.4	16485.1

* p<0.05, ** p<0.01, *** p<0.001

Table I shows that living a farther distance away from a comprehensive cancer center (ccmiles) leads to a lower probability of breast cancer survival. There is a negative relationship between travel distance and breast cancer survival across various methods of estimation. Column 1 and 2 show the estimations from the Cox Proportional Hazards model, where the coefficients reflect changes in death probability. Under this model, living 10 miles farther away from a comprehensive cancer center leads to a mortality probability that is 4 percentage points higher, which is a large effect. The effect of distance depends on the age at diagnosis as well. A patient who lives 10 miles farther away, but is diagnosed at a greater age, has a smaller increase in mortality hazard. Columns 3 and 4 show the results from probit model estimations of travel distance and the probability of breast cancer survival. There is a positive, although statistically insignificant, relationship between travel distance and the probability of breast cancer survival. Because the probit model does not effectively adjust for the incomplete observation of patient survival, the results are less meaningful. Column 5 shows the results of a logit regression, where distance is positively related

with survival. Based on the larger standard errors, the estimates from the logit regression are less precise and suffer from incomplete observation of patient survival.

Table I also shows that stage of diagnosis is a main driver of breast cancer mortality. Columns 1 and 2 show that, compared to the earliest stage of diagnosis, the increased mortality risk for a stage 2 diagnosis is 0.36. For a diagnosis at stage 3, the associated marginal mortality probability is more than twice as high as that of a stage 2 diagnosis, at 0.89. For the most advanced stage of diagnosis, the marginal mortality probability is four times as high as that of a stage 3 diagnosis, at 1.56. In Columns 3, 4, and 5, the probit and logit regressions show that each later stage of diagnosis is associated with a higher probability of mortality from breast cancer. Compared to a diagnosis at the earliest stage, a patient who is diagnosed at stage 2 increases the probability of death from breast cancer by 25%. A stage 3 diagnosis increases the chance of death by 73%, and the most advanced stage increases it by 150%.

Table II. Estimations of the effect of demographic factors on survival probability for breast cancer patients diagnosed from 2000-2011 in Kentucky using Cox Proportional Hazards Model, Probit Models, and Logit Model

	Cox b/se	Cox2 b/se	Probit 1 b/se	Probit 3 b/se	Logit b/se
AGE_DX	-0.0650***	-0.1223***	-0.0272	-0.0087	0.0105
	(0.007)	(0.011)	(0.030)	(0.057)	(0.112)
age2	0.0008***	0.0013***	0.0010*	0.0004	0.0007
	(0.000)	(0.000)	(0.000)	(0.001)	(0.002)
agedist	-0.0001*	-0.0001*		-0.0000	-0.0000
	(0.000)	(0.000)		(0.000)	(0.000)
agerace	-0.0041	-0.0034		0.0009	
	(0.002)	(0.002)		(0.003)	
sex== 1.0000	0.2380*	0.2355*	-0.2816*	-0.1718	-0.4923*
	(0.116)	(0.116)	(0.141)	(0.131)	(0.239)
African American	0.4344**	0.3836*	-0.1787***	-0.2326	-0.3205***
	(0.165)	(0.164)	(0.045)	(0.160)	(0.079)
Non-cancer death		-5.5962***		-4.6687	-1.6233
		(0.945)		(3.719)	(7.049)
constant			1.0772	-2.7593*	0.7039
			(0.912)	(1.128)	(2.243)
chi-sqr	3732.006	3768.202	3969.536	31543.525	3968.700
dfres					
BIC	162236.9	162211.2	16464.5	19847.4	16485.1

* p<0.05, ** p<0.01, *** p<0.001

Table II shows that the patient's age at diagnosis (age_dx) has an ambiguous effect on breast cancer mortality. Column 1 shows the estimations from the Cox Proportional Hazards model.

Under this specification, a greater age of diagnosis leads to a lower hazard of cancer mortality initially, then contributes to a higher mortality hazard from cancer. The negative linear effect of age dominates over the squared term. As age increases, the probabililty of non-cancer mortality also increases. The probability of non-cancer related death at each age is obatined from the Census Life Tables and included as an explanatory variable in the regression in Column 2. When the probability of non-cancer related deaths is controlled for, older age has a smaller effect on increased cancer mortality probability. Column 4 shows the results of a probit model estimating the relationship between age and breast cancer survival. It shows that if the age at diagnosis is 10 years greater, the probability of dying is 0.07 higher. Again, because the probit model does not control for the year of diagnosis as effectively as the Cox proportional hazard model, the results are less accurate.

Finally, it shows that being African American increases the marginal mortality probability by 0.38 compared to being Caucasian. The effect of being African American is equivalent to being diagnosed at an age that is 3 years older. Identification as a racial minority other than African American is associated with even higher marginal mortality probability, though the sample size is small.

Table III. Estimations of the effect of socioeconomic factors on survival probability for breast cancer patients diagnosed from 2000-2011 in Kentucky using Cox Proportional Hazards Model, Probit Models, and Logit Model

	Cox b/se	Cox2 b/se	Probit 1 b/se	Probit 3 b/se	Logit b/se
% high school educated	-0.0008	-0.0007	0.0050	0.0058	0.0097
	(0.004)	(0.004)	(0.005)	(0.004)	(0.008)
Physician-population ratio	-23.5350**	-24.3344**	10.6518	9.3676	17.6374
	(8.581)	(8.586)	(8.554)	(7.618)	(15.398)
%medicaid	0.0039	0.0049	-0.0054	0.0029	-0.0086
	(0.008)	(0.008)	(0.009)	(0.008)	(0.015)
%privateinsurance	0.0023	0.0030	0.0001	0.0072	0.0008
	(0.009)	(0.009)	(0.009)	(0.008)	(0.017)
Median income	-0.0837**	-0.0825**	0.0206	-0.0160	0.0333
	(0.030)	(0.030)	(0.030)	(0.026)	(0.054)
constant			1.0772	-2.7593*	0.7039
			(0.912)	(1.128)	(2.243)
chi-sqr	3732.006	3768.202	3969.536	31543.525	3968.700
dfres					
BIC	162236.9	162211.2	16464.5	19847.4	16485.1

* p<0.05, ** p<0.01, *** p<0.001

Table III shows how county level physician to population ratio, medicaid coverage, household income, and education influence survival from breast cancer. Column 1 and 2 show that the concentration of physicians has a positive impact on the patient's survival outcome. Given that the average physician to population ratio is 2 per 1,000, an increase in the ratio by 1 per 1,000 lowers the patient's probability of death by 0.023 with all else constant. Median household income and breast cancer survival probability are inversely correlated. If the county median income rises by $10,000, the patient's mortality hazard from breast cancer becomes lower by 0.08.

Table IV. Estimation of socioeconomic factors on the stage of diagnosis for breast cancer patients diagnosed from 2000-2011 in Kentucky using ordered probit models

	Model 1 b/se	Model 2 b/se	Model 3 b/se
Distance cancer	0.000 (0.001)	0.000 (0.001)	-0.000 (0.001)
Distance mammography	-0.000 (0.001)		
Median income	-0.008 (0.014)	-0.007 (0.015)	-0.008 (0.015)
AGE_DX	-0.057*** (0.015)	-0.057*** (0.015)	-0.057*** (0.015)
age2	0.001* (0.000)	0.001* (0.000)	0.001* (0.000)
age3	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
agedist	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
Physician-population ratio	-13.461** (4.346)	-13.779** (4.274)	-11.870** (4.508)
%medicaid	0.010* (0.004)	0.010* (0.004)	0.013** (0.005)
%privateinsurance	0.008 (0.005)	0.008 (0.005)	0.011* (0.005)
mar_stat== ..0000	-0.119*** (0.020)	-0.119*** (0.020)	-0.119*** (0.020)
sex== 1.0000	0.394*** (0.074)	0.394*** (0.074)	0.394*** (0.074)
African American	0.097*** (0.023)	0.097*** (0.023)	0.099*** (0.023)
Other race	0.012 (0.089)	0.012 (0.089)	0.014 (0.089)
% high school educated	-0.007** (0.002)	-0.007** (0.002)	-0.007** (0.002)

* p<0.05, ** p<0.01, *** p<0.001

Table IV shows that neither the distance to the cancer center nor distance to mammography facilities seem to drive the patients' stage of diagnosis. However, the physician to patient ratio may

be capturing the effect of access to cancers because distance from cancer center is correlated with physician density, as shown before. There is disparity in the stage of diagnosis due to demographic factors such as race and sex. African American patients are diagnosed late by $1/10^{th}$ of a stage than their Caucasian counterparts. Women are diagnosed almost half a stage later than men are. A younger age of diagnosis is correlated with an earlier stage of diagnosis. A patient who is 10 years younger is likely to get diagnosed earlier by half of a stage. This can result from the increase in public health education in the younger generation and access to internet and health information, but not quite from improvements in diagnostic technology throughout different time periods.

The county level proportion of high school graduates, a proxy for the patient's level of education, has a large influence on the stage of diagnosis. For a county with the average percentage of high school educated adults （82.9%）, the diagnosis will be made almost a quarter of a stage earlier compared to a county with 50% high school graduates. The physician to population ratio also has a strong effect on the patient's stage of diagnosis. Given that the average physician to population ratio is 1 per thousand for the patients in the study, patients in a county like Jefferson which has a ratio of 7 per thousand experience an earlier diagnosis by a tenth of stage. If we hope to diagnose patients by half a stage earlier, the physician to population ratio will need to increase from 2 per thousand to 34 per thousand. This entails training many more physicians than the system is currently producing.

Table V. Using model to make stage of diagnosis and prognosis predictions for 4 hypothetical patients **1)** 60 y/o African American female, low socioeconomic status, living 30 miles from a cancer center; **2)** 60 y/o white female, low socioeconomic status, living 30 miles away from a cancer center; **3)** 60 y/o African American female, high socioeconomic status, living less than 10 miles from a cancer center; **4)** 60 y/o white female, high socioeconomic status, living less than 10 miles from a cancer center; calculations found in the appendix

	Expected Stage of Diagnosis	Expected Mortality Hazard
1) AA low socioeconomic status (coal miner)	1.998	-2.43
2) White low socioeconomic status (coal miner)	1.890	-2.65
3) AA high socioeconomic status (professor)	1.441	-4.51
4) White high socioeconomic status (professor)	1.334	-4.70

Imagine there are 4 archetypal breast cancer patients of the same gender and age, but different professions and socioeconomic status. Two are professors who are capable of living within 10 miles of a comprehensive cancer center. They also have high socioeconomic status, reflected by their residence in an area where income, education, and insurance levels are all above the population average by 100%. The other two individuals are coal miners who can only afford to live in a rural area that is 30 miles away from a comprehensive cancer center. They have low socioeconomic status, reflected by their residence in an area where income, education, and insurance levels are all below the average by around 50%.

Comparing the stage of diagnosis of the white coal miner and the white professor, we can see that the coalminer is expected to be diagnosed half a stage later than the professor. In the Cox proportional hazards model, a more positive hazard implies poorer prognosis for the patient. The coal miner's expected mortality hazard is -2.56 while the professor's is -4.70, which implies a worse prognosis for the coal miner compared to the professor. A similar trend exists for the African American coal miner and professor. These results are for hypothetical patients, but they reflect the reality of health disparity that actual patients may experience.

VIII. Discussion and Conclusion

Previous studies that address the influence of travel distance to care on cancer survival have found mixed results (Gill, 2002; Huang, 2009; Massarweh, 2014). These studies have varied in controlling for income, comorbidities, and other socioeconomic factors, all of which I controlled for in my models. My results from the SEER 18 national cohort of breast cancer patients show that those who live farther from cancer centers have higher mortality risk from breast cancer by 4% per 10 additional miles. The disparity in cancer outcomes that results from travel distance holds even after controlling for distance-related confounding variables such as socioeconomic status, availability of physicians, and insurance type. Huang (2009) found that longer distance from cancer centers led to more advanced diagnosis, but I found that distance does not delay diagnosis for this cohort, after controlling for physician density. The poorer cancer outcomes are less attributed to failure to detect breast cancer at an early stage than to geographical barrier in accessing treatment, which makes sense given the abundance of mammography facilities in most locations. Therefore, implementing policies that decrease the travel burden for cancer patients in distant or rural areas may reduce the disparity in outcomes.

Demographic factors that affect survival outcome the most are the patient's age at diagnosis, sex, and race. Past studies using multivariate regression found that a greater age of diagnosis is associated with poorer prognosis for cancer, but only observes patients from two years (Gill, 2002). However, my results from the Cox regression on 5 years of observations show that only ages past 80 years old increase the probability of mortality. Those who are over 80 years old at the time of diagnosis comprise 10% of the patients in the cohort. Overall, those diagnosed at very old ages have an increased marginal probability of mortality. Because the average age at diagnosis for patients is around 60 years old, many patients do not experience increased risk due to old age. Increased age

contributes to a later stage of diagnosis, however. An age of diagnosis that is greater than 57 years old is associated with a more advanced stage of diagnosis, which signals that older patients may have less knowledge about prevention and detect cancer too late. Later detection and diagnosis in older patients may explain the higher probability of death because cancer may metastasize if diagnosed to late. Results also show that older patients are less affected by living far away from a comprehensive cancer center, based on the negative age-distance interaction term. Their schedules may better accomodate the drive to farther cancer centers compared to younger patients.

In previous studies, no association had been found between race and cancer outcomes (Bradley, 2002; Kim, 2013). However, my results shows that probability of mortality from breast cancer is higher for African Americans compared to Caucasians even after controlling for distance traveled to cancer center, insurance type, education, income level, and other factors. Through the ordered probit regression, I found that the main cause for increased probability of death in African American patients is the delay in the stage of diagnosis, which agrees with results found by Massarweh (2014). This suggests that outcomes may be improved with more public health programs targeted at African Americans and minorities regarding early breast cancer detection.

In accordance with Ward (2004), my results show that the mortality risk for breast cancer is higher for people living in areas with worse insurance status, lower median household income, and lower physician to population ratio. Higher median household income was a predictor of lower probabilities of death from breast cancer, which may be due to ability to get better drugs, treatment, and nutrition to fight the cancer. As predicted by the theory that a higher density of physicians will influence cancer detection and patient outcomes, a higher physician to population ratio around the patient's residential location predicts an earlier stage of diagnosis and a lower probability of death. In counties with the highest percentage of Medicaid patients, patients are diagnosed as much as half a

stage later when all other socioeconomics factors are controlled for. However, aggregate insurance status is not significantly associated with cancer mortality outcome, which may be due to laws that assist with breast cancer treatment for low-income areas.

This study has a few limitations. First, the geographical granularity of the socioeconomic data is limited to the county level by the SEER dataset. Ideally, it would be better to use data on the block-group or census tract level. Second, county level income, insurance, and education data were projected onto individual patients due the privacy of individual insurance information. Finally, the results of this paper are specific to breast cancer, but the model can be used to study all other cancer types. Despite these limitations, the combined SEER 18 - AHRF dataset provides more comprehensive data for patient information than has been available in the past. The results provide great insight into the source of disparity in breast cancer mortality and potential routes that can be taken to address the disparity. As I extend this work, I will conduct more comparison on the stage of diagnosis and mortality predictions obtained from my model for real patients with the patients' actual experience and adjust covariates in order to improve the accuracy of the model.

Citations

Asadzadeh, Vostakolaei F, et al. (2011) The validity of the mortality to incidence ratio as a proxy for site-specific cancer survival. The European Journal of Public Health 21(5): 573–577. doi: 10.1093/eurpub/ckq120. Retrieved September 26, 2014, from http://ejournals.ebsco.com/Direct.asp?AccessToken=6LVVHL989HOMOI3I9MXXNI39XMHO8FCLV&Show=Object&msid=604015288

Ayanian, J. et al. (1993) The relationship between health insurance and clinical outcomes in women with breast cancer. The New England Journal of Medicine. 329(5) 326-331 Retrieved September 26, 2014, from http://www.nejm.org/doi/pdf/10.1056/NEJM199307293290507

Bailar, J., & Smith, E. (1986). Progress against cancer? The New England Journal of Medicine, 314(19), 1226-1232.

.Bradley CJ, Given CW, Roberts C (2002) Race, socioeconomic status, and breast cancer treatment and survival. J Natl Cancer Inst 94:490–496

Cancer Facts and Figures 2014. (2014, January 1). Retrieved September 7, 2014, from (http://www.cancer.org/acs/groups/content/@research/documents/webcontent/acspc-042151.pdf)

Cutler, D., & Lleras-Muney, A. (2009). Understanding differences in health behaviors by education. Journal of Health Economics, 29(1), 1-28. Retrieved September 5, 2014, from http://www.sciencedirect.com/science/article/pii/S0167629609001143#

Gill, A. J., & Martin, I. G. (2002). Survival from upper gastrointestinal cancer in New Zealand: The effect of distance from a major hospital, socio-economic status, ethnicity, age and gender. ANZ Journal Of Surgery,72(9), 643-646. Retrieved October 30th, 2014, from http://www.ncbi.nlm.nih.gov/pubmed/23765207

Gomez, S. (2014) Racial/ethnic (RE) differences in the occurrence of HER2 and hormone receptor (HR)-defined breast cancer (BC) in California (CA). J Clin Oncol 32:5s Retrieved December 1st, 2014

Gunderson, C.C. et al. (2013). Distance traveled for treatment of cervical cancer: Who travels the farthest, and does it impact outcome? International Journal of Gynecological Cancer, 23 (6), pp. 1099–1103 Retrieved October 30th, 2014, from http://www.ncbi.nlm.nih.gov/pubmed/23765207

Hall, R.E., Jones, C.I., 2007. The value of life and the rise in health spending. Quarterly Journal of Economics 122 (1), 39–72. Retrieved October 6, 2014, from

Huang, Bin (2009) Does distance matter? Distance to mammography facilities and stage at diagnosis of breast cancer in Kentucky. J Rural Health. 2009 Fall; 25(4): 366–371. doi: 10.1111/j.1748-0361.2009.00245.x

Kim, S. et al. (2014) Sociodemographic Characteristics, Distance to the Clinic, and Breast Cancer Screening Results J Health Dispar Res Pract. 2013 Spring; 6(1): 70. Retrieved October 25, 2014, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3898539/

Lakdawalla, Darius, Eric Sun, Anupam Jena, Carolina Reyes, Dana Golman, and Tomas Philipson (2010). "An Economic Evaluation of the War on Cancer," Leaving the Board Journal of Health Economics, vol. 29(3), pp. 333-46. Retrieved September 26, 2014, from http://www.sciencedirect.com/science/article/pii/S0167629610000214

Lange, F. (2010) The role of education in complex health decisions: evidence from cancer screening. Journal of Health Economics, 30(1), 43-54 Retrieved September 5, 2014, from http://www.sciencedirect.com/science/article/pii/S0167629610001177#

Lichtenberg, F. (2010). Has Medical Innovation Reduced Cancer Mortality. NBER Working Paper. Retrieved September 5, 2014, from http://www.nber.org/papers/w15880

Lichtenberg, F. (2010). "Are Increasing 5-Year Survival Rates Evidence of Success Against Cancer? A Reexamination Using Data from the U.S. and Australia," Forum for Health Economics & Policy, Berkeley Electronic Press, vol. 13(2), Retrieved September 5, 2014, from http://www.nber.org/papers/w16051

Massarweh NN, Chiang YJ, Xing Y, et al.: Association between travel distance and metastatic disease at diagnosis among patients with colon cancer. J Clin Oncol 2014; 32:942–948. Retrieved October 30th, 2014, from http://onlinelibrary.wiley.com/doi/10.1002/jso.23664/full

Rauh Hain (2013) Clemmer J, Clark RM, et al. Racial disparities and changes in clinical characteristics and survival for vulvar cancer over time. Am J Obstet Gynecol 2013;209:468.e1-10. Retrieved October 30th, 2014, from *http://www.ajog.org/article/S0002-9378(13)00753-9/abstract*

Rooney, Timothy (2014) The Relationship between and Geographic Distribution of Breast Cancer Statistics: Diagnosis, Survival, and Mortality in Selected Areas in the United States, 1973-2004. Duke Journal of Economics; Retrieved August 30th, 2014 from *http://econ.duke.edu/uploads/media_items/timothyrooneydjepaper.original.pdf*

Ward, E. et al. (2004) "Cancer Disparities by Race, Ethnicity, and Socioeconomic Status." Cancer Journal for Clinicians 54, 78-93. Retrieved September 26, 2014, from http://onlinelibrary.wiley.com/doi/10.3322/canjclin.54.2.78/pdf

Wang, F., McLafferty, S., et al. (2008) "Late-Stage Breast Cancer Diagnosis and Healthcare Access in Illinois." Professional Geographer, 60 (1) (2008), pp. 54–69. Retrieved September 26, 2014, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2367325/

Welch, HG., Schwartz, LM., Woloshin, S. (2000) Are increasing 5-year survival rates evidence of success against cancer? Jama-Journal of the American Medical Association 283(22): 2975–2978. doi: Retrieved September 26, 2014, from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0083100#pone.0083100-Welch1

World Health Organization (2014) World Cancer Report Retrieved October 25th, 2014, from http://www.esmo.org/Oncology-News/World-Cancer-Report-2014

Appendix:

Variable Definitions

Variable	Description
ccmiles	Distance from nearest comprehensive cancer center
Near_dist/Miles	Distance from nearest mammography center
stage	Stage of diagnosis: 0= in situ or stage 1, 1=localized or stage 2, 2=regional or stage 2, 3=distant or stage 4
%25+ w. HS diploma	Proxy for education level by county
median household ~12	Median household income by county
pavgprivate	% of privately insured patients by county
pavgmedicaid	% of Medicaid insured patients by county
mdratio	Physician to population ratio
age_dx	Age at diagnosis
comorbidities	Probability of non-cancer related death at each age
year	Year of diagnosis
sex	Gender
race	Ethnicity
mar_stat	Marriage Status

Summary Statistics for Key Variables

Variable	Obs	Mean	Std. Dev.	Min	Max
eventdate	37645	79.81703	38.80569	1	144
event	37645	.2200292	.4142716	0	1
stagedum2	37645	.5296587	.4991262	0	1
stagedum3	37645	.2450525	.430124	0	1
stagedum4	37645	.0563421	.2305843	0	1
ccmiles	37645	65.18942	30.8176	1.016297	130.2715
mamdist	37645	18504.29	21359.51	378.2893	122767.1
ageofd	37645	61.34028	13.55138	19	102
age2	37645	3946.266	1688.755	361	10404
agedist	37645	4011.741	2154.397	24.39113	12119.89
agerace	37645	65.69954	21.59691	19	246
mdratio	37645	.0024441	.0019161	.0000849	.0068311
pavgmedicaid	37645	18.22193	7.244666	5.723056	49.39404
avginsured	37645	83.70944	2.013681	77.74286	88.4
medincome	37645	4.366636	.9848377	1.9624	8.3164
gender1	37645	.0056316	.0748331	0	1
race2	37645	.0661442	.2485374	0	1
race3	37645	.0040112	.0632074	0	1
hseducation	37645	82.94534	6.887453	57.4	91.8

Principal Components Analysis Results

```
Principal components (eigenvectors)
```

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Comp8	Comp9	Unexplained
ccmiles	-0.2219	-0.2012	0.4036	-0.1912	0.6600	-0.4421	0.0384	0.2344	0.1527	0
mamdist	-0.2581	-0.2189	0.0648	-0.0642	0.3150	0.8614	-0.0754	0.1717	-0.0275	0
pop2010	0.3412	0.4553	0.1294	-0.0519	0.4635	0.1738	0.2952	-0.5697	-0.0112	0
medincome	0.4455	-0.4339	0.0113	0.0202	-0.0882	0.1034	0.2052	-0.0277	0.7426	0
noinsuran~12	-0.4064	0.4708	-0.1380	0.0259	-0.1031	0.0454	0.5915	0.3287	0.3506	0
hseducation	0.4976	-0.1388	0.0518	-0.0106	0.0228	0.0298	0.4893	0.4909	-0.4985	0
stage	-0.0429	0.0143	0.4111	0.9100	0.0106	0.0214	0.0071	0.0151	0.0000	0
mdratio	0.3932	0.5040	-0.0136	0.0097	0.1401	0.0601	-0.5237	0.4908	0.2298	0
ageofd	-0.0022	0.1448	0.7906	-0.3568	-0.4600	0.1210	-0.0062	-0.0143	0.0132	0

```
.
. *Scree plot of the eigenvalues
. screeplot

. screeplot, yline(1)

.
. *Principal component analysis
. pca $xlist, mineigen(1)

Principal components/correlation                  Number of obs    =     37645
                                                  Number of comp.  =         3
                                                  Trace            =         9
    Rotation: (unrotated = principal)             Rho              =    0.6397
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	3.41756	2.09851	0.3797	0.3797
Comp2	1.31905	.298228	0.1466	0.5263
Comp3	1.02082	.0301453	0.1134	0.6397
Comp4	.990676	.063488	0.1101	0.7498
Comp5	.927188	.116525	0.1030	0.8528
Comp6	.810663	.583955	0.0901	0.9429
Comp7	.226709	.0250495	0.0252	0.9681
Comp8	.201659	.115991	0.0224	0.9905
Comp9	.0856685	.	0.0095	1.0000

```
Principal components (eigenvectors)
```

Variable	Comp1	Comp2	Comp3	Unexplained
ccmiles	-0.2219	-0.2012	0.4036	.6119
mamdist	-0.2581	-0.2189	0.0648	.7049
pop2010	0.3412	0.4553	0.1294	.3116
medincome	0.4455	-0.4339	0.0113	.07322
noinsuran~12	-0.4064	0.4708	-0.1380	.1238
hseducation	0.4976	-0.1388	0.0518	.1255
stage	-0.0429	0.0143	0.4111	.8209
mdratio	0.3932	0.5040	-0.0136	.1365
ageofd	-0.0022	0.1448	0.7906	.3342

```
. pca $xlist, comp($ncomp)
```

Exploratory Data Analysis

My exploratory data analysis involves summary statistics given in the data section and the following verification of the correlations between stage of diagnosis and age, ethnicity, education, and income that Massarweh (2014) has found. Massarweh et al. (2014) shows that age, lower

educational backgrounds, and lower income levels are all correlated with a later stage of diagnosis. He also shows that white patients tend to present at an earlier stage than black patients. Finally, I am verifying correlations between my independent variables to confirm that my data is valid.

Age and Stage of Diagnosis

SEER 18 data shows that there is a negative correlation of 0.0029 between age and stage of diagnosis.

Age and Survival

Plotting survival against the age of diagnosis, we see a slightly negative trend in survival as age increases (SEER 18). This could be due to an increase in the instances of comorbidity with age, or simply natural death from old age. Age at diagnosis will need to be controlled for in the regression.

Income and Stage of Diagnosis

There is a weak negative correlation between income and the stage of diagnosis of -0.0193. Higher income is correlated with an earlier stage of diagnosis.

Education and Stage of Diagnosis

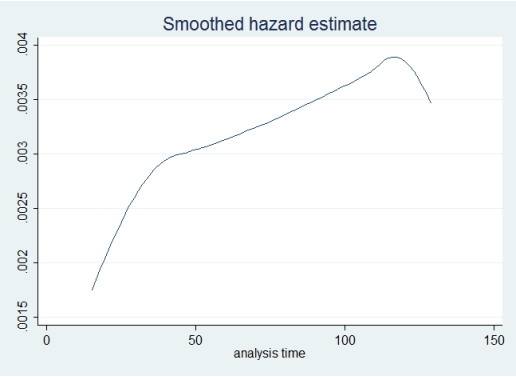
I correlated the stage of diagnosis with the percentage of the population 25+ with no high school diploma (SEER 18 and AHRF). A larger fraction of only high-school educated population in the county is correlated with presenting at a later stage of diagnosis for a patient from that county.

Pairwise Correlations

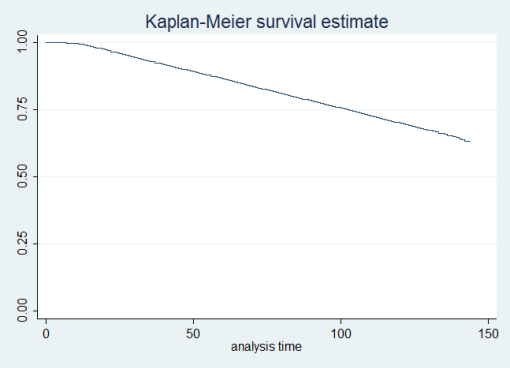
The physician to population ratio was not included in Massarweh (2014)'s model, but is an important measure of resources relevant to survival. The SEER 18 data shows that there is a small

negative correlation of -0.0420 between the ratio and the stage of diagnosis. Therefore, having more doctors available is correlated to having an earlier stage of diagnosis.

Results:



Overall hazard Function for breast cancer patients diagnosed in 2000-2011 in Kentucky. This shows the marginal hazard of dying at each time point after a diagnosis. The hazard rate increases until 110 months, then falls.



Overall cumulative survival function for breast cancer patients diagnosed in 2000-2011. The proportion of patients diagnosed in Kentucky from 2000-2011 who survive until 2011 is 65%.

Table V Calculations

1） 60 y/o African American female, low socioeconomic status, living 30 miles from a cancer center

Expected stage of diagnosis: -0.008*(10)-0.057*60+0.001*(60)^2-13.401*(0.001) +0.010*(60)+0.008*(20)+0.0970-.119*(0)-0.007*(50)+0.394 +1= **1.998**

Expected Mortality Hazard: 0.3545*(1)+0.0039*(30)-0.1223*(60)+0.0013*(60)^2-0.0001*(60)*(10)-0.0034*(60)-24.3344*(0.001)+0.0049*(60)+0.0030*(20)-0.0825*(10)+0.2355+0.3836*(1)-0.0007*(50)-5.5962***(**0.01224**)= - 2.43**

2) 60 y/o white female, low socioeconomic status, living 30 miles away from a cancer center

Expected stage of diagnosis: -0.008*(10)-0.057*60+0.001*(60)^2-13.401*(0.001)+0.010*(60)+0.008*(20)-.119*(0) 0.007*(50)+0.394 +1= **1.890 for 60 y/o**

Expected Mortality Hazard: 0.3545*(0.89)+0.0039*(30)-0.1223*(60)+0.0013*(60)^2-0.0001*(60)*(10) - 24.3344*(0.001)+0.0049*(60)+0.0030*(20)-0.0825*(10)+0.2355-0.0007*(50)-5.5962*(0.01224) = **-2.65**

3) 60 y/o African American female, high socioeconomic status, living less than 10 miles from a cancer center

Expected stage of diagnosis: -0.008*(30)-0.057*60+0.001*(60)^2-13.401*(0.003) +0.010*(20)+0.008*(60)+0.0970-.119*(0)-0.007*(90)+0.394+1 =

= **1.441**

Expected Mortality Hazard: 0.3545*(0.441)+0.0039*(10)-0.1223*(60)+0.0013*(60)^2-0.0001*(60)*(10)-0.0034*(60)-24.3344*(0.003)+0.0049*(20)+0.0030*(60)-0.0825*(30)+0.2355+0.3836*(1)-0.0007*(90)-5.5962*(0.01224) = **-4.51**

4) 60 y/o white female, high socioeconomic status, living less than 10 miles from a cancer center

Expected stage of diagnosis: -0.008*(30)-0.057*60+0.001*(60)^2-13.401*(0.003) +0.010*(20)+0.008*(60)-.119*(0)-0.007*(90)+0.394 +1= 1.334

Expected Mortality Hazard: 0.3545*(0.334)+0.0039*(10)-0.1223*(60)+0.0013*(60)^2-0.0001*(60)*(10) - 24.3344*(0.003)+0.0049*(20)+0.0030*(60)-0.0825*(30)+0.2355-0.0007*(50)-5.5962*(0.01224) = **-4.7**