In presenting this dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I agree that the Library of the University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to copy from, or to publish, this dissertation may be granted by the professor under whose direction it was written when such copying or publication is solely for scholarly purposes and does not involve potential financial gain. In the absence of the professor, the dean of the Graduate School may grant permission. It is understood that any copying from, or publication of, this dissertation which involves potential financial gain will not be allowed without written permission.

_____

Kevin C. Ward

Population-Based Cancer Registries:
The Role of Area-Based Measures of Socioeconomic Status and Cancer Survival

By

Kevin C. Ward
Doctor of Philosophy

Department of Epidemiology

_____
John L. Young, Jr.
Adviser

_____
Michael Goodman
Committee Member

_____
Jonathan Liff
Committee Member

_____
Lance Waller
Committee Member

_____
Joseph Lipscomb
Committee Member

Accepted:

_____
Lisa A. Tedesco, Ph.D.
Dean of the Graduate School

_____
Date

Population-Based Cancer Registries:
The Role of Area-Based Measures of Socioeconomic Status and Cancer Survival

By

Kevin C. Ward
B.I.E., Georgia Institute of Technology, 1993
M.P.H., Emory University, 1998

Adviser: John L. Young, Jr., Dr.P.H., C.T.R.

An Abstract of
A dissertation submitted to the Faculty of the Graduate School
of Emory University in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Department of Epidemiology

2008

ABSTRACT

Area-based measures of socioeconomic status (SES) are widely used in studies of cancer survival. There is consistent evidence that survival varies by SES for many malignancies and that this relationship is resilient to the choice of area-based measure. The specific role of SES in population-based survival statistics and the mechanisms by which SES affects cancer survival are not fully understood.

Three separate studies were conducted utilizing population-based cancer registry data from the Surveillance, Epidemiology, and End Results (SEER) Program of the U.S. National Cancer Institute. The percent of the census tract population living below the poverty level was chosen as the area-based SES measure of interest based on previous research of health disparities using linked census data. The goals of these studies were: 1) to evaluate the validity of one area-based measure of SES; 2) to explore the effect of SES-specific mortality in the calculation of relative survival; and 3) to examine the relationships between SES and survival from non-localized prostate cancer.

Study I identified practical steps to improve geocoding outcomes of registry data and revealed substantial misclassification of SES when geocoding at the level of the ZIP code. Study II demonstrated that the use of SES-specific background mortality in the calculation of relative survival produced a reduction in the survival disparity observed when using national data that do not take SES into consideration. This observation was most pronounced in SES-stratified analyses. Study III, utilizing SEER-Medicare linked data, identified a significant interaction between SES and stage of disease at diagnosis

and showed that while survival by SES does differ among study individuals with equal eligibility for care, much of the SES related disparity is explained by factors for which SES may serve as a surrogate.

Area-based measures of SES play an important role in population-based studies of cancer survival and need to be more fully utilized in the calculation of relative survival. It is important to understand and evaluate the completeness and accuracy of the geocoded data from which area-based measures are obtained. Further research is needed to elucidate the specific, complex pathways by which SES affects cancer survival.

Population-Based Cancer Registries:
The Role of Area-Based Measures of Socioeconomic Status and Cancer Survival

By

Kevin C. Ward
B.I.E., Georgia Institute of Technology, 1993
M.P.H., Emory University, 1998

Adviser: John L. Young, Jr., Dr.P.H., C.T.R.

A dissertation submitted to the Faculty of the Graduate School
of Emory University in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Department of Epidemiology

2008

ACKNOWLEDGEMENTS

I dedicate this dissertation to the memory of my father-in-law, Dr. John W. Youtsey. He is the reason that I entered the field of epidemiology and I am eternally grateful for his direction in my life. He is loved and missed by many, but he lives on in the lives of those who so dearly loved him.

To the members of my dissertation committee, I would like to express my deepest gratitude for your guidance and support during my graduate studies. I have learned so much from each of you and I thank you all for dedicating your time to my professional development. I am especially grateful to Dr. Michael Goodman for his invaluable substantive and editorial advice. It was an honor to learn from someone with such skills in these areas.

To my mentor, boss, and very dear friend, Dr. John Young, I can not begin to express my appreciation for helping to shape my career. I am confident in the skills I have acquired because I have learned from one of the best. I am blessed to have you as a mentor and even more so to have you as a friend.

I would like to thank the National Cancer Institute's SEER Program for their financial support of this work and the Georgia Center for Cancer Statistics staff for their love and encouragement that continues to this day.

Last, but certainly not least, I would like to thank my family. To my wife, Heather, and son, Austin, I am forever grateful for your never-ending love and support. There is nothing in life more precious to me than the two of you. To my father and mother, Charles and Sandra, I thank you for the opportunities you afforded me and the encouragement and love you have always so freely given.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

INTRODUCTION

STUDY MOTIVATION

Area-based measures of socioeconomic status (SES) are readily obtained by linking geocoded data with United States Census Summary Files from the decennial census and are widely used in the cancer literature (1-7). There is consistent evidence that both cancer and non-cancer mortality differ by measures of area-based SES (8-14) and that survival from selected cancers does as well (13, 15-20). Socioeconomic status is a complex measure encompassing both individual and contextual components. Research has shown that the neighborhood in which one lives captures aspects of health beyond those that can be measured by individual SES alone (21-24) and that neighborhood SES might be even more appropriate than individual measures of SES for selected segments of the population, specifically the young and old (23).

To expand research utilizing area-based measures of health in public health surveillance systems, researchers from Harvard University School of Public Health conducted the Public Health Disparities Geocoding Project (4, 25, 26). They set out to determine which area-based census measures of socioeconomic status, at which specific level of geography, were most appropriate for monitoring socioeconomic disparities in health. Using well established a priori criteria to evaluate their results, these researchers suggested that measures of economic deprivation at the level of the census tract, specifically the percent of the census population living below the poverty level, were most effective for evaluating health disparities using linked census data. Use of the

percent of the census population living below the poverty level demonstrated consistent expected gradients across population subgroups, was robust across a range of disease outcomes, allowed for maximal linkage with census data, and was easy to understand and explain.

Despite the impressive research to date in the field of area-based measures of SES and health disparities, there is still much to be learned regarding the use of these measures specifically as they relate to cancer survival. Research is needed to explore the validity of area-based SES assignment from geocoded data with inherent errors in positional accuracy. The role of area-based SES-specific mortality in standard survival calculations from population-based registries has not been fully explored. Finally, the specific relationship between area-based SES and survival needs further examination in the presence of factors known to affect cancer survival in an effort to help elucidate the specific roles by which SES may influence this outcome.

PURPOSE

This dissertation will expand upon current research in the field of area-based measures of SES and cancer survival in an effort to explore the issues described above. Three separate studies were conducted utilizing population-based cancer registry data from the Surveillance, Epidemiology, and End Results (SEER) Program of the U.S. National Cancer Institute using the percent of the census tract population living below the poverty level as the area-based measure of SES. The choice of this specific area-based measure

was based on previous research from the Public Health Disparities Geocoding Project. The primary goals of these studies are described below.

1) The first study evaluated the validity of the area-based measure of SES obtained by linking census information to geocoded address data from the SEER database. Steps to improve the completeness and accuracy of geocoded data were also explored.

2) The second study explored the role of SES-specific background mortality in the calculation of aggregated and SES-stratified relative survival rates. Regression models were used to further evaluate the role of SES in cancer survival while controlling for background mortality and other selected factors known to affect survival. Comparisons with cause-specific survival were also made.

3) The third study examined the extent to which various demographic, clinical, and social factors explained the association between SES and non-localized prostate cancer survival in a population with equal eligibility for care.

This research provides a significant addition to the literature on area-based measures of SES and cancer survival. It will hopefully serve to stimulate further utilization of these measures through a better understanding of their validity, their expanded role in survival calculations, and their relationship with other factors known to affect survival.

REFERENCES

1. Coughlin SS, King J, Richards TB, et al. Cervical cancer screening among women in metropolitan areas of the United States by individual-level and area-based measures of socioeconomic status, 2000 to 2002. Cancer Epidemiology, Biomarkers & Prevention 2006;15:2154-9.
2. Du XL, Fang S, Vernon SW, et al. Racial disparities and socioeconomic status in association with survival in a large population-based cohort of elderly patients with colon cancer. Cancer 2007;110:660-9.
3. Gomez SL, O'Malley CD, Stroup A, et al. Longitudinal, population-based study of racial/ethnic differences in colorectal cancer survival: impact of neighborhood socioeconomic status, treatment and comorbidity. BMC Cancer 2007;7:193.
4. Krieger N, Chen J, Waterman P, et al. Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? The Public Health Disparities Geocoding Project. American Journal of Epidemiology 2002;156:471-82.
5. MacKinnon JA, Duncan RC, Huang Y, et al. Detecting an association between socioeconomic status and late stage breast cancer using spatial analysis and area-based measures. Cancer Epidemiology, Biomarkers & Prevention 2007;16:756-62.
6. Sanderson M, Coker AL, Perez A, et al. A multilevel analysis of socioeconomic status and prostate cancer risk. Annals of Epidemiology 2006;16:901-7.

7. Zell JA, Rhee JM, Ziogas A, et al. Race, socioeconomic status, treatment, and survival time among pancreatic cancer cases in California. Cancer Epidemiology, Biomarkers & Prevention 2007;16:546-52.

8. Faggiano F, Partanen T, Kogevinas M, et al. Socioeconomic differences in cancer incidence and mortality. In: Kogevinas M, Pearce N, Susser M, Boffetta P, eds. Social Inequalities and Cancer: IARC Scientific Publications, 1997:65-176.

9. Geronimus AT. Poverty, time and place: variation in excess mortality across selected US populations, 1980-1990. Journal of Epidemiology and Community Health 1999;53:325-334.

10. Geronimus AT, Colen CG, Shochet T, et al. Urban-rural differences in excess mortality among high-poverty populations: evidence from the Harlem Household Survey and the Pitt County, North Carolina Study of African American Health. Journal of Health Care for the Poor & Underserved 2006;17:532-58.

11. Muller A. Association between income inequality and mortality among US States: considering population at risk.[comment]. American Journal of Public Health 2006;96:590-1.

12. Singh GK, Hiatt RA. Trends and disparities in socioeconomic and behavioral characteristics, life expectancy, and cause-specific mortality of native-born and foreign-born populations in the United States, 1979-2003.[see comment]. International Journal of Epidemiology 2006;35:903-19.

13. Singh GK, Miller BA, Hankey BF, et al. Persistent area socioeconomic disparities in U.S. incidence of cervical cancer, mortality, stage, and survival, 1975-2000. Cancer 2004;101:1051-7.

14. Vinnakota S, Lam NSN. Socioeconomic inequality of cancer mortality in the United States: a spatial data mining approach. International Journal of Health Geographics [Electronic Resource] 2006;5:9.

15. Clark JY, Thompson IM. Military rank as a measure of socioeconomic status and survival from prostate cancer. Southern Medical Journal 1994;87:1141-4.

16. Dickman PW, Auvinen A, Voutilainen ET, et al. Measuring social class differences in cancer patient survival: is it necessary to control for social class differences in general population mortality? A Finnish population-based study. Journal of Epidemiology & Community Health 1998;52:727-34.

17. Du XL, Fang S, Coker AL, et al. Racial disparity and socioeconomic status in association with survival in older men with local/regional stage prostate carcinoma: findings from a large community-based cohort. Cancer 2006;106:1276-85.

18. Kogevinas M, Porta M. Socioeconomic differences in cancer survival: a review of the evidence. IARC Scientific Publications 1997:177-206.

19. McDavid K, Tucker TC, Sloggett A, et al. Cancer survival in Kentucky and health insurance coverage.[see comment]. Archives of Internal Medicine 2003;163:2135-44.

20. Singh GK, Miller BA, Hankey BF, et al. Area Socioeconomic Variations in US Cancer Incidence, Mortality, Stage, Treatment and Survival, 1975-1999: Bethesda, MD., National Cancer Institute, NCI Cancer Monograph Series, Number 4, NIH Pub No. 03-5417, 2003.

21. Winkleby M, Cubbin C, Ahn D. Effect of cross-level interaction between individual and neighborhood socioeconomic status on adult mortality rates. American Journal of Public Health 2006;96:2145-53.

22. Krieger N, Williams DR, Moss NE. Measuring social class in US public health research: concepts, methodologies, and guidelines. Annual Review of Public Health 1997;18:341-78.

23. Rehkopf DH, Haughton LT, Chen JT, et al. Monitoring socioeconomic disparities in death: comparing individual-level education and area-based socioeconomic measures. American Journal of Public Health 2006;96:2135-8.

24. Van Lenthe FJ, Mackenbach JP. Neighborhood and individual socioeconomic inequalities in smoking: the role of physical neighborhood stressors. Journal of Epidemiology & Community Health 2006;60:699-705.

25. Krieger N, Waterman PD, Chen JT, et al. Monitoring socioeconomic inequalities in sexually transmitted infections, tuberculosis, and violence: geocoding and choice of area-based socioeconomic measures--the public health disparities geocoding project (US). Public Health Reports 2003;118:240-60.

26. Krieger N, Chen JT, Waterman PD, et al. Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: The Public Health Disparities Geocoding Project (US). Journal of Epidemiology & Community Health 2003;57:186-99.

Chapter 1

Population-Based Cancer Registries

SURVEILLANCE, EPIDEMIOLOGY, AND END RESULTS PROGRAM

In 1971, the National Cancer Act was signed into law to expand the authority of the National Cancer Institute (NCI) in an effort to advance progress in the "war" against cancer (1). The Act mandated collection, analysis, and dissemination of cancer data and led to the creation of the NCI's Surveillance, Epidemiology and End Results (SEER) Program (2). In 1973, the SEER Program began collecting population-based cancer data in the states of Connecticut, Iowa, New Mexico, Utah, and Hawaii and in the metropolitan areas of Detroit and San Francisco-Oakland. In 1974-1975, SEER expanded its coverage to include the metropolitan areas of Atlanta and Seattle-Puget Sound and then again in 1978 to include 10 rural counties in Georgia. At present, the SEER Program includes 18 registries covering approximately 26 percent of the population of the United States (U.S.) (2).

The SEER Program is the authoritative source of information on cancer incidence and survival in the U.S. and, along with state registries, forms the foundation of cancer surveillance in the United States. High quality surveillance systems allow researchers to monitor changes in cancer incidence and survival, while national death certificate data serve the same purpose in evaluating cancer mortality. It is only through monitoring variations in these measures that one can effectively evaluate programs aimed to improve preventive, diagnostic, screening, and treatment services at the population level.

METROPOLITAN ATLANTA AND RURAL GA SEER REGISTRY

Background

The Metropolitan Atlanta and Rural Georgia (MARGA) SEER Registry is operated by the Georgia Center for Cancer Statistics at Emory University in Atlanta, GA. This registry is responsible for collecting detailed demographic and clinical information on all incident cases of cancer in a five-county area of metropolitan Atlanta (Clayton, Cobb, DeKalb, Fulton, and Gwinnett) and a ten-county area located to the southeast of Atlanta (Glasscock, Greene, Hancock, Jasper, Jefferson, Morgan, Putnam, Taliaferro, Warren, and Washington). The MARGA registry is the oldest SEER registry in the southeastern U.S. State legislation designates cancer as a reportable disease in Georgia, thereby mandating systematic collection of cancer data (3).

Definition of Reportable Diagnoses

The MARGA registry collects information on all individuals diagnosed with a reportable cancer while residing in one of the 15 MARGA counties at the time of diagnosis. Briefly, reportable diagnoses include all *in situ* and invasive neoplasms defined by the International Classification of Diseases for Oncology, Third Edition (4). The non-reportable exceptions to the above definition are basal and squamous cell carcinomas of the non-genital skin, *in situ* carcinoma of the cervix, cervical intraepithelial neoplasia, and prostatic intraepithelial neoplasia (5). Beginning in 2004, the benign and borderline tumors of the brain and central nervous system became reportable as well. While over 95% of the cancers in the MARGA registry are confirmed by pathology, cases diagnosed clinically are also reportable.

The Data Collection Procedures

The diagnosis, treatment, and evaluation of a cancer patient can take place in a variety of settings. These include acute care hospitals and clinics, free-standing surgical and oncology centers, radiation therapy facilities, independent pathology laboratories, physician practices, and hospices. All sources of data relating to the patient's diagnosis or treatment should be reviewed in order to ensure complete ascertainment of necessary information. Tumor registrars are the foundation of cancer data collection as they are specifically trained in this process (6).

Reporting facilities in the MARGA coverage areas are required to identify cancer cases within their individual facilities and report them to the registry on a monthly basis. Facilities that do not have staff registrars to perform these tasks generally have their data abstracted and submitted by contract tumor registrars or field abstractors employed by the registry. Information about MARGA cancer cases diagnosed or treated outside of the 15 SEER counties is captured through a data-sharing agreement with the statewide Georgia Comprehensive Cancer Registry (GCCR) (7). In addition, the GCCR has data exchange agreements with all of the states bordering Georgia and with several other states with large cancer treatment facilities.

Data Items Collected

As with all other cancer registries in the U.S., the MARGA registry collects and reports all data in the electronic format of the North American Association of Central Cancer Registries (NAACCR) (8). NAACCR is the data standards organization responsible for

developing and promoting uniform data standards for cancer registration in the United States and Canada. The SEER Program collects data for the following types of variables:

Demographic

Name, age at diagnosis, race, ethnicity, sex, marital status, date of birth

Geographic

Residence at diagnosis (including address, county, and census tract), birthplace

Cancer Identification

Primary site, laterality, date of diagnosis, histology, behavior, grade, sequence, site-specific biomarkers

Hospital-Specific

Reporting facility, facility-specific treatment

Extent of Disease

Tumor size, tumor extension, lymph node involvement, regional nodes positive, regional nodes examined, summary stage

First Course Treatment

Date of first course treatment, surgery, radiation therapy, chemotherapy, hormone therapy, other therapy

Follow-up

Vital status, date of death or date last known to be alive, cause of death, autopsy

Text

Physical exam, x-rays, microscopic exams, laboratory tests, pathology, surgical procedures, other treatment

Although the above list shows most key variables, it is by no means exhaustive. Detailed information regarding all data items collected by SEER and the appropriate values for each variable can be found in the NAACCR manual (8), the SEER Program Code Manual (5), or the SEER Extent of Disease Manual (9).

Editing and Consolidating Cancer Information

The primary responsibilities of MARGA staff are editing and consolidating incoming cancer abstracts. Computerized edits are used to evaluate completeness of the incoming data and the validity of each individual data item. Visual editing is used to compare the code selected for each data item against the text documentation that is included with each abstract. Visual editing is necessary to ensure accuracy of the coded registry data and to provide feedback to tumor registrars.

Cancer patients often receive care at multiple facilities. For example, a patient diagnosed at one facility may choose to undergo surgery at a different hospital and then receive weekly radiation treatment at a free-standing radiation center. As a result, one patient may have several separate abstracts reported to the registry. The process of record consolidation involves combining multiple sources of cancer information into a single record that most accurately and completely describes the diagnosis, staging, and treatment of each cancer patient.

Patient Follow-up

To examine survival after a cancer diagnosis, registries participating in the SEER Program follow all patients on an annual basis to ascertain vital status, date of last contact and, for those known to be deceased, cause of death. While in the past patient follow-up was largely achieved through letters mailed to patients, physicians and next-of-kin, current practices primarily involve record linkages. The MARGA registry matches its data on a monthly basis against vital records from the state of Georgia. This process allows the registry to identify cancer patients who have recently died and to obtain the date of death and cause of death for each deceased patient. Every year, the registry also matches its database against the National Death Index (10) to identify patients who died outside the state of Georgia. In addition, patient vital status is confirmed by matching registry data against a variety of independent sources such as birth records, voter records, hospital discharge records, Medicare files, and Social Security files. Hospitals participating in the American College of Surgeons Commission on Cancer (CoC) Program also provide updated follow-up information to the MARGA registry since they are required to follow their own cancer patients to meet CoC Program standards (11). The minimum acceptable annual standard for follow-up of living registrants is 95 percent, according to the Registry's contract with the NCI.

Geocoding

An important part of the cancer registry abstract is the patient's residential address at the time of diagnosis. The MARGA registry sends all patient addresses for geocoding to a commercial vendor. Geocoding is the process of assigning geographic coordinates,

typically latitude and longitude, or census tract, to a street address of interest (12, 13). The process of geocoding utilizes a street reference database containing a complete list of street segments in the U.S. Each individual street segment in the reference database contains geographic coordinates and street numbers for the beginning and ending points of the segment on each side of the street. The cancer patient's residential street is first located in the reference database, using other ancillary variables such as city, state, and ZIP code. Next, the location of the cancer patient's house number along the reference database street segment is interpolated using the beginning and ending street numbers for the segment. This process allows the assignment of interpolated latitude and longitude coordinates to the street address of interest. Census tracts are then assigned to the address using census reference IDs associated with each individual segment. When an exact street address cannot be located in the underlying reference database, the centroid of the residence ZIP code is often used instead (14). Thus, it is important to keep in mind that geocoding results for different addresses may have different levels of certainty, which are reflected in the following NAACCR codes (8).

Code	Meaning
1	Census tract based on a complete and valid street address of residence
2	Census tract based on residence ZIP + 4
3	Census tract based on residence ZIP + 2
4	Census tract based on residence ZIP code only
5	Census tract based on ZIP code of PO Box
9	Unable to assign census tract

One of the reasons U.S. registries use geocoded addresses is a lack of individual measures of socioeconomic status (SES). Most cancer registry data are derived from the patient's medical record and these records typically do not contain SES information. Geocoded registry data allow the assignment of area-based measures of SES to the individual cancer record by linking census tracts from the registry to census summary files from the U.S. Census (15-17).

REFERENCES

1. Rettig R. Cancer Crusade: The Story of the National Cancer Act of 1971. Princeton, N.J.: Princeton University Press, 1977.
2. National Cancer Institute. SEER: Surveillance, Epidemiology and End Results Program: National Institute of Health (Publication No. 05-4772), 2005.
3. Division of Public Health. Reporting Cancer: Notifiable Disease Law: Georgia Department of Human Resources, 2007.
4. World Health Organization. International Classification of Diseases for Oncology, Third Edition. Geneva: World Health Organization, 2000.
5. Johnson C, Adamo M, (eds.). SEER Program Coding and Staging Manual 2007. National Cancer Institute, NIH Publication number 07-5581, Bethesda, MD 2007.
6. Fritz A. The coder and the cancer registrar: defining the differences in roles and coding. Journal of Ahima 2005;76:66-7.
7. Division of Public Health. Georgia Comprehensive Cancer Registry: Georgia Department of Human Resources, 2007.
8. Hofferkamp J, Havener L, editors. Standards for Cancer Registries Volume II: Data Standards and Data Dictionary, Twelfth Edition, Version 11.2. Springfield, IL: North American Association of Central Cancer Registries, April 2007.
9. Fritz A, Ries L, editors. SEER Extent of Disease - 1988; Codes and Coding Instructions: National Cancer Institute, Bethesda, MD, 1998.
10. National Center for Health Statistics. National Death Index, 2007.
11. Commission on Cancer. Cancer Program Standards, 2004, Revised Edition. Chicago, IL: American College of Surgeons, 2006.

12. Cromley E, McLafferty S. GIS and public health. New York: The Guilford Press, 2002.

13. Waller L, Gotway C. Applied Spatial Statistics for Public Health Data. Hoboken: John Wiley & Sons, Inc, 2004.

14. Tele Atlas. Tele Atlas Geocoding Service - Reference Documentation, 2006.

15. Chen F, Breiman R, Farley M, et al. Geocoding and linking data from population-based surveillance and the US Census to evaluate the impact of median household income on the epidemiology of invasive Streptococcus pneumonia infections. American Journal of Epidemiology 1998;148:1212-8.

16. Krieger N, Chen J, Waterman P, et al. Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? The Public Health Disparities Geocoding Project. American Journal of Epidemiology 2002;156:471-82.

17. Krieger N, Chen JT, Waterman PD, et al. Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: The Public Health Disparities Geocoding Project (US). Journal of Epidemiology & Community Health 2003;57:186-99.

Chapter 2

Relative Survival

NET SURVIVAL

The probability of death for a cancer patient is subject to competing forces of mortality: death resulting from the addition of the newly diagnosed cancer and death from all other causes (1). The two leading measures of population-based survival in the presence of competing cause mortality are cause-specific survival (2, 3) and relative survival (1, 4-6). Both measures attempt to estimate "net survival", average patient survival from a specific cancer diagnosis corrected for other causes of death.

Cause-specific survival uses standard life table methodology (4, 7, 8) but treats deaths from causes other than the cancer under investigation as censored observations. This approach requires patient-specific cause of death, which in turn depends on the accuracy of information recorded in the death certificate. Difficulty can arise in deciding whether a death was due to cancer or due to other causes.

Relative survival, on the other hand, uses a somewhat more complicated three-step procedure. First, all-cause survival in a cohort of interest is estimated using life table methodology. Next survival of the general population, with demographic characteristics (such as age, sex, race, and calendar year) that match those of the cohort under investigation, is calculated using expected rate tables. Expected rates tables are typically derived from life tables of the national population (6). Finally, estimates of relative

survival are determined by dividing the observed survival of the cohort under investigation by the expected survival of the demographically matched population. This approach avoids limitations reported to be associated with the use of death certificates (9-11). On the other hand, special attention should be given to the choice of the population used to calculate background mortality and the variables upon which it is matched to the cancer cohort. If expected survival is underestimated, relative survival will be overestimated and vice versa.

Of the two methods for estimating net survival, relative survival is the one most commonly used by population-based cancer registries (12). It should be noted that the survival rate resulting from the calculation of relative survival is not really a rate but is instead a proportion. The term rate, however, will be used in this chapter as it is widely accepted. A five-year relative survival rate represents the proportion of patients not succumbing to the cancer under investigation during the five-year period following diagnosis. A relative survival rate less than 100 percent indicates that mortality in the cancer cohort was greater than mortality in the general population during the specified time interval. A peculiarity of this methodology is that at any point in time the survival of the patient population may exceed that of the general population, resulting in a relative survival proportion greater than 100 percent. Typically, the relative survival rate is truncated at 100 percent when this happens.

As with all statistical methods, relative survival calculations are based on a number of assumptions. It is assumed that the background mortality of the matched general

population cohort does not include mortality from the specific cancer of interest, or that such mortality is negligible. In most cases this assumption has been shown to be justified (1). If, however, the contribution of the cancer of interest to background mortality is not negligible, relative survival will be overestimated due to underestimated expected survival. It also is assumed that censored events are independent of the outcome and that the competing causes of death are independent of each other and thus additive. This additive hazards model has been suggested to be more biologically plausible for cancer data (13, 14).

CALCULATION OF RELATIVE SURVIVAL

The methodology used in the calculation of relative survival can be demonstrated using the following simple example. Suppose a small cohort of 10 prostate cancer patients was followed for five years. Table 2.1 presents demographic and survival characteristics of this 10-patient cohort. As shown in the table, all patients were white males between the ages of 60 and 84. Survival times ranged from 10 months to 60 months following diagnosis. During the 5-year follow-up, six patients died of prostate cancer, two patients died of other causes, one patient was alive at the end of the 5-year period, and one patient was lost to follow-up. The numerator of the relative survival rate (observed survival) is calculated using standard life table methodology for all-cause survival. This can be achieved using either actuarial or Kaplan-Meier estimates of patient survival as described elsewhere (15). One of the key attributes of these methods is their ability to address censored observations. In this cohort, the censored observations would include the patient still alive after five years of observation and the patient lost to follow-up. The

two deaths from causes other than prostate cancer, which are censored observations in cause-specific survival calculations, are considered uncensored events in all-cause survival analysis.

Table 2.2 shows observed survival for our hypothetical prostate cancer cohort calculated using actuarial estimates. After dividing the 5-year follow-up into one-year intervals (i = 1, 2, 3, 4, 5), interval-specific probabilities of death (q_i), interval-specific probabilities of survival (p_i) and cumulative probabilities of observed survival ($_1po_i$) are then calculated as:

$$q_i = d_i / [n_i - (w_i/2)]$$

$$p_i = 1 - q_i$$

and

$$_1po_i = \prod p_j \text{, from } j=1 \text{ to } i$$

where d_i represents the number of deaths during the i^{th} interval, n_i represents the number of patients alive at the beginning of the i^{th} interval, and w_i represents the number of patients censored during the i^{th} interval.

The denominator for the relative survival rate (expected survival) is generally calculated using life tables from the national population. The U.S. population census is conducted every 10 years. Decennial life tables are generated using census population counts and national mortality (16). Life tables present the probability of dying between the ages of x

and x+1 for a person of a given race and sex in a specific census year. The core component of the life table is the death rate Q_x. Defined as the proportion of the population at age x expected to die before attaining age x+1, it is calculated using the formula:

$$Q_x = D_x / (3P_x + 1/2D_x)$$

where D_x is the adjusted number of deaths occurring in a population of age x over the three year period surrounding the population census, P_x is the census population of age x at the mid-year of the census period, and deaths are assumed to occur uniformly over the one-year period during which the age advances from x to x+1 (16). Deaths for which age was not stated are allocated proportionately among the different age groups through the use of an adjustment factor. Beers interpolation coefficients are used to smooth single age population death rates (17). This technique is applied to both death (D_x) and population (P_x) values by aggregating the single year data into 5-year age groups and then interpolating back to single age values. The inverse of the life table death rate ($1- Q_x$) represents the proportion of the population surviving the age interval x to x+1 and is the value used in the expected rate table to calculate expected survival. While population life tables in the U.S. are generated by age, race, sex, and census year, they can also be generated to include other variables such as socioeconomic status.

Table 2.3 presents an example of life table death rates (Q_x) from a decennial census for white males between the ages of 60 and 89. From this table, the probability of dying

between the age of 70 and 71 for a white male in this census year is estimated as 0.03237. The corresponding value for a white male between the age of 71 and 72 is estimated as 0.03531. Expected survival for a cancer cohort can be generated using these life table values. The Ederer I method is the simplest method of estimating the expected survival (1). Using this method, interval-specific expected survival probabilities ($pe_i(s)$) are first generated for each individual subject in the cancer cohort based on life table values for a matched individual from the general population. As shown in Table 2.4, the first year survival probability for a 60 year old white male (subject 1) is obtained by taking the inverse of the life table death rate for this man ($pe_1(1)$= 1- 0.0134 = 0.9866). The second year survival probability for subject 1 is obtained by taking the inverse of the life table death rate for a white male of age 61 ($pe_2(1)$= 1- 0.01486 = 0.9851). The results for the remainder of the cohort are generated in the same fashion (Table 2.4).

The next step in the calculation of expected survival is to determine the cumulative expected survival probability ($_1pe_i(s)$) for each individual in the cancer cohort. This is calculated as:

$$_1pe_i(s) = \prod p_j(s) \text{ , from } j=1 \text{ to } i$$

Again using subject 1 as an example, the 5-year cumulative expected survival probability for this individual is calculated as $_1pe_5(1)$ = 0.9866 x 0.9851 x 0.9836 x 0.9821 x 0.9805 = 0.9205 (Table 2.5). The final step in the calculation of expected survival is to generate

cumulative expected survival rates ($_1pe_i$) for the entire cohort of cancer patients from the date of diagnosis to the end of the i^{th} interval. The Ederer I method uses the formula

$$_1pe_i = \sum {}_1pe_i(s) / n_i, \text{ summed from } s=1 \text{ to } n_i$$

with n_i representing the number of patients alive at the beginning of the i^{th} interval. These cumulative expected survival probabilities are presented at the bottom of Table 2.5. As shown in Table 2.6, relative survival rates ($_1RS_i$) can now be calculated for each interval by dividing the observed survival in the cancer cohort ($_1po_i$) by the expected survival obtained from the matched general population cohort ($_1pe_i$).

Table 2.1 Characteristics of 10 prostate cancer patients followed for 5 years to assess survival

Subject	Sex	Race	Age	Survival time (months)	Status
1	m	w	60	10	Dead - PC
2	m	w	66	23	Dead - PC
3	m	w	84	27	Dead - PC
4	m	w	77	42	Dead - OC
5	m	w	73	47	Alive
6	m	w	61	60	Alive
7	m	w	80	14	Dead - OC
8	m	w	69	30	Dead - PC
9	m	w	61	37	Dead - PC
10	m	w	71	55	Dead - PC

Status: Dead-PC (died of prostate cancer)
Dead-OC (died of other cause)
Alive (alive at last contact or close of the study)

Table 2.2 Observed 5-year survival for study cases generated using actuarial methods

i	n_i	d_i	w_i	q_i	p_i	${}_1po_i$
1	10	1	0	0.1000	0.9000	0.9000
2	9	2	0	0.2222	0.7778	0.7000
3	7	2	0	0.2857	0.7143	0.5000
4	5	2	1	0.4444	0.5556	0.2778
5	2	1	0	0.5000	0.5000	0.1389

Table 2.3 Section of the annual life table showing the probability of death between the ages of x and x+1 for white men between the ages of 60 and 89

X=	X0	X1	X2	X3	X4	X5	X6	X7	X8	X9
6	0.01345	0.01486	0.01636	0.01793	0.01953	0.02109	0.02271	0.02458	0.02685	0.02950
7	0.03237	0.03531	0.03839	0.04159	0.04496	0.04856	0.05253	0.05703	0.06228	0.06837
8	0.07552	0.08360	0.09225	0.10109	0.11027	0.12042	0.13100	0.14238	0.15473	0.16810

Table 2.4 Interval-specific expected survival probabilities of each study case for the first 5 years. Generated from population life tables (Table 2.3)

				Interval expected survival probability ($pe_i(s)$)				
Subject	Sex	Race	Age	1	2	3	4	5
1	m	w	60	0.9866	0.9851	0.9836	0.9821	0.9805
2	m	w	66	0.9773	0.9754	0.9732	0.9705	0.9676
3	m	w	84	0.8897	0.8796	0.8690	0.8576	0.8453
4	m	w	77	0.9430	0.9377	0.9316	0.9245	0.9164
5	m	w	73	0.9584	0.9550	0.9514	0.9475	0.9430
6	m	w	61	0.9851	0.9836	0.9821	0.9805	0.9789
7	m	w	80	0.9245	0.9164	0.9077	0.8989	0.8897
8	m	w	69	0.9705	0.9676	0.9647	0.9616	0.9584
9	m	w	61	0.9851	0.9836	0.9821	0.9805	0.9789
10	m	w	71	0.9647	0.9616	0.9584	0.9550	0.9514

Table 2.5 Cumulative expected survival probabilities of each study case for the first 5 years. Generated from interval-specific probabilities (Table 2.2)

				Cumulative expected survival probability ($_1pe_i$ (s))				
Subject	Sex	Race	Age	1	2	3	4	5
1	m	w	60	0.9866	0.9719	0.9560	0.9389	0.9205
2	m	w	66	0.9773	0.9533	0.9277	0.9003	0.8712
3	m	w	84	0.8897	0.7826	0.6801	0.5832	0.4930
4	m	w	77	0.9430	0.8842	0.8238	0.7616	0.6979
5	m	w	73	0.9584	0.9153	0.8709	0.8251	0.7781
6	m	w	61	0.9851	0.9690	0.9517	0.9331	0.9134
7	m	w	80	0.9245	0.8472	0.7690	0.6913	0.6151
8	m	w	69	0.9705	0.9391	0.9059	0.8712	0.8349
9	m	w	61	0.9851	0.9690	0.9517	0.9331	0.9134
10	m	w	71	0.9647	0.9277	0.8891	0.8491	0.8079
			$_1pe_i$ =	0.95849	0.915931	0.872575	0.828679	0.784531

Table 2.6 Relative survival rates for study cases

i	n_i	d_i	w_i	q_i	p_i	$_1po_i$	$_1pe_i$	$_1RS_i$
1	10	1	0	0.1000	0.9000	0.9000	0.9585	0.9390
2	9	2	0	0.2222	0.7778	0.7000	0.9159	0.7642
3	7	2	0	0.2857	0.7143	0.5000	0.8726	0.5730
4	5	2	1	0.4444	0.5556	0.2778	0.8287	0.3352
5	2	1	0	0.5000	0.5000	0.1389	0.7845	0.1770

REFERENCES

1. Ederer F, Axtell LM, Cutler SJ. The relative survival rate: a statistical methodology. National Cancer Institute Monograph 1961;6:101-21.
2. Bull K, Spiegelhalter DJ. Survival analysis in observational studies. Statistics in Medicine 1997;16:1041-74.
3. Cox DR. Regression models and life tables (with discussion). Journal of the Royal Statistical Society 1972;Series B:187-220.
4. Berkson I, Gage RP. Calculation of Survival Rates for Cancer. Proceedings Staff Meet. Mayo Clinic 1950;25:270-86.
5. Esteve J, Benhamou E, Croasdale M, et al. Relative survival and the estimation of net survival: elements for further discussion. Stat Med 1990;9:529-38.
6. Henson DE, Ries LA. The relative survival rate. Cancer 1995;76:1687-8.
7. Cutler SJ, Ederer F. Maximum utilization of the life table method in analyzing survival. Journal of Chronic Diseases 1958;8:699-712.
8. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 1958;53:457-481.
9. Engel LW, Strauchen JA, Chiazze L, Jr., et al. Accuracy of death certification in an autopsied population with specific attention to malignant neoplasms and vascular diseases. American Journal of Epidemiology 1980;111:99-112.
10. Percy C, Stanek E, Gloeckler L. Accuracy of cancer death certificates and its effect on cancer mortality statistics. American Journal of Public Health 1981;71:242-50.

11. Percy CL, Miller BA, Gloeckler Ries LA. Effect of changes in cancer classification and the accuracy of cancer death certificates on trends in cancer mortality. Annals of the New York Academy of Sciences 1990;609:87-97; discussion 97-9.

12. Ries LAG, Melbert D, Krapcho M, et al. SEER Cancer Statistics Review, 1975-2004. National Cancer Institute. Bethesda, MD., http://seer.cancer.gov/csr/1975_2004/, based on November 2006 SEER data submission, posted to the SEER web site, 2007.

13. Buckley JD. Additive and multiplicative models for relative survival rates. Biometrics 1984;40:51-62.

14. Hakulinen T, Tenkanen L. Regression analysis of relative survival rates. Applied Statistics 1987;36:309-371.

15. Kleinbaum D, Klein M. Survival Analysis, A Self-Learning Text, Second Edition. New York: Springer, 2005.

16. Anderson R. Method for constructing complete annual U.S. life tables. National Center for Health Statistics. Vital Health Stat 2(129). 1999.

17. Shryock H, Siegel J. The methods and materials of demography, vol 2, U.S. Bureau of the Census. Washington, D.C.: U.S. Government Printing Office.

Chapter 3

Geocoding and Area-Based Measures of Socioeconomic Status

GEOCODING

The Process

Geocoding refers to the process of assigning geographic coordinates, typically latitude and longitude, or census tract, to a street address of interest and has been described in numerous texts in relation to geographic information systems (1-3). Four primary methods exist for geocoding data to point coordinates (latitude and longitude). Once point coordinates are obtained, other levels of geography can be assigned to the address (e.g. census tracts, census blocks, or ZIP codes). The first method assigns the address of interest to the centroid of the geographic unit, typically the ZIP code, in which it is contained. A centroid is defined as the center point of a polygon. Centroids are often weighted to account for the location of the population within the polygon (4). The second method uses a street reference database (e.g. Topologically Integrated Geographic Encoding and Referencing (TIGER)/Line files from the United States Census Bureau (5)) and an interpolation algorithm to assign individual street addresses along a referenced street segment. If the actual point coordinate of a specific address is located in the reference database, the interpolation technique is not used. The third method uses a parcel database to assign to each address the point coordinates of the matching address parcel. While increasing in popularity and more accurate than the other methods (6), this method requires complete parcel maps that are not yet available nationwide. The forth and final method involves the use of a hand-held GPS device to measure the address from

the actual location of interest. While clearly the most accurate method available, it is also the most expensive, time consuming, and least feasible for large datasets. Discussion from this point forward will focus on the first two methods described above.

The process of geocoding usually begins by "scrubbing" the address file to be geocoded. Complete and accurate address level data are key components of successful geocoding. Address standardization and Coding Accuracy Support System (CASS) verification are two commonly used scrubbing techniques. Address standardization takes each individual street address and parses the address into individual components (i.e. house number, street prefix, street name, street suffix, street type, city, state, ZIP code). The individual address components are then reviewed and recoded to a standard format. For example, the directional street prefixes Nth and North would be standardized to "N". Variations in street names, such as Peachtree St, Pchtree Street, and P'tree St, would all be standardized to "PEACHTREE ST". Address standardization also creates soundex codes (a type of phonetic algorithm) for selected components of the address to be used in future address matching routines. CASS verification attempts to validate each address for mail delivery (7). This process compares the street and city against the ZIP code and makes corrections to the ZIP code or city as necessary. It also evaluates and makes corrections to individual street components as needed. Addresses not likely to geocode at the street level are highlighted during this process.

After the address file has been standardized and verified, each address can be matched against the reference street database. Each individual street segment in the reference

database contains geographic coordinates and street numbers for the beginning and ending points of the segment on each side of the street. The street for geocoding is first located in the reference database, using other ancillary variables such as city, state and ZIP code. Next, the location of the house number along the reference database street segment is interpolated using the beginning and ending street numbers for the segment. In certain instances, no interpolation is necessary as the address is located at the end of the street segment. This process allows the assignment of interpolated latitude and longitude coordinates to the street address being geocoded in a manner such that the assigned coordinate is made proportional to its location on the street segment of a defined length. Census tracts are then assigned to the address using census reference identifiers associated with each individual street segment. Some vendors use a point-in-polygon approach to assign census tracts from point coordinates, although this method has been identified as less accurate (8). When an exact street address cannot be located in the underlying reference database, the centroid of the residence ZIP code is often used to geocode the address instead (4). Centroids of polygons are contained within the reference database for use when necessary.

Geocoding Error

There are a number of potential problems that can arise during the geocoding process described above. Errors in the address file are a major problem (9, 10). Misspelled address components, missing address components (e.g. street numbers, directional components) and incorrect street numbers are some of the most common errors. Post Office (PO) Box addresses do not represent where the patient is living at the time of

diagnosis and can only be geocoded to the ZIP Code centroid of the PO Box. Errors in the reference file also exist (11-15). Research investigating the accuracy of TIGER/Line files estimated positional errors of 30 to 121 meters between a sample of GPS measured positions and TIGER file positions (14). One source of these errors is interpolation. Interpolation techniques make certain assumptions that can introduce error; in certain areas the size of the error may be substantial. The techniques generally assume that all addresses in a range actually exist and that addresses are distributed homogeneously across the street segment range. These assumptions can introduce larger positional inaccuracies in areas with longer street segments or fewer addresses (12, 13). Deficiencies in the references files are also a problem. Rural route addresses are often not found in reference files (16, 17) and newly added streets may be missing as well. If vendors do not update their references files on a regular basis, they can become quickly outdated for geocoding current information.

As a result of the types of errors presented above, numerous researchers have investigated the implications of using geocoded data in epidemiologic investigations. The majority of the research has focused on geocoding match rates (6, 18-22) and positional accuracy of geocoded data (6, 11-13, 16, 20, 21, 23, 24). Geocoding match rates are reported to range between 20 and 100 percent across vendors and datasets while positional accuracy has also been found to vary greatly. As one example, Whitsel et al. compared four commercial vendors for geocoding match rates and positional accuracy, using a large nationwide dataset with established coordinates. Their research found substantial differences in both measures, with vendors matching the smallest proportion of addresses

(match rates for vendors A-D: 98%, 82%, 81%, 30%) having the highest degree of spatial accuracy as measured in meters (positional error in meters for vendors A-D: 1809, 748, 704, 228) (20).

The leading cause of positional errors in most research has been attributed to rurality or population density. Positional errors in geocoded locations are strongly correlated with population density. As the population density increases, positional error has been shown to decrease (6, 11, 12, 16). Cayo et al., for example, showed that while 95 percent of urban addresses geocoded to within 152 meters, the same statistic for rural areas was 2,872 meters. This is largely the result of the underlying street reference databases. Shorter street segments are more commonly found in densely populated urban areas, as there are more intersections. Across these shorter street segments, the interpolation process involved in geocoding has less room for error. Addresses of people living in these more densely populated areas contain more complete address information and fewer rural routes. This results in more street matched addresses, fewer ZIP code centroid matches and, consequently, much less positional error in urban areas.

Positional inaccuracy in the assignment of an address to a point coordinate can lead to biased study results. Hurley et al. and Krieger et al. both demonstrated this issue in separate studies with the use of addresses geocoded to the level of the ZIP code (16, 25). In their study, containing a substantial number of participants with PO Box addresses, Hurley et al. showed that box holders are often not representative of the general

population and that excluding this population from their study could result in selection bias.

Differential match rates by geographic region also can lead to biased results if data are missing in a non-random fashion and confounding is present between the risk factor under examination and geographic region. Oliver et al. identified spatially non-random differences in geocoding completeness in their study of prostate cancer incidence in Virginia (19). While they showed that area-based measures of income and urban status were associated with increased incidence of prostate cancer in their data, they also showed that rural counties had a higher percentage of unsuccessfully geocoded data. This, in conjunction with a statistically significant association of both older age and lower education with unsuccessfully geocoded addresses among males, led them to suggest that geographic confounding could explain some of the findings of their study.

Studies have shown that approximately 75 percent of addresses standardized to the U.S. Postal Service address format, excluding PO Boxes and Rural Routes, can be successfully geocoded to a street level address by geocoding only (18, 26, 27). Methods to improve geocoding match rates have been identified to enhance addresses that were not successfully geocoded to a street address on first attempt. Researchers utilizing geocoded data must realize the error that is inherent in the geocoding process and take steps to minimize its effects on their research. The degree of error that is acceptable to any research project primarily depends on the exposure of interest and the level of specificity that is needed in exposure assignment.

AREA-BASED MEASURES OF SOCIOECONOMIC STATUS

Socioeconomic data are largely missing from United States surveillance systems (28). To address this problem, researchers from the Harvard University School of Public Health conducted the Public Health Disparities Geocoding Project (29-31). Acknowledging 1) that geocoded data exist in most surveillance systems, 2) that U.S. Census data contain many different measures of socioeconomic status, and 3) that these data can be readily linked together, the researchers set out to determine which area-based census measures of socioeconomic status, at which specific level of geography, were most appropriate for monitoring socioeconomic disparities in health.

The researchers utilized public health surveillance systems from Rhode Island and Massachusetts to monitor seven different health outcomes: mortality (all-cause and the five most common causes of death in each state), cancer incidence (all-cause and the 5 most common sites), low birth weight, childhood lead poisoning, sexually transmitted infections, tuberculosis, and nonfatal weapons-related injuries. Files 3A and 3B of the 1990 U.S. Census Summary Tape (32) were used to obtain census tract, census block group, and ZIP code data for 11 single-variable, area-based measures of SES and 8 composite measures the researchers created for this study. These area-based measures were constructed around the categories of occupational class, income, poverty, wealth, educational level, and crowding. The final decision regarding the optimal area-based measure of SES and corresponding level of geography was based upon an *a priori* decision to evaluate each measure with regard to external validity, robustness, completeness, and ease of use at each level of geography.

The results of this study provided evidence that the choice of both area-based measure of SES and level of geography were important considerations. The researchers suggested that measures of economic deprivation at the level of the census tract, specifically the percent of the census population living below the poverty level, are most effective for evaluating health disparities using linked census data. This specific measure demonstrated consistent gradients of association across population subgroups, was robust across a range of disease outcomes, allowed for maximal linkage with census data, and was easy to understand and explain.

While area-based measures of socioeconomic status are clearly useful for monitoring disparities across a wide range of health outcomes, some researchers have questioned their role in relation to individual measures of SES. However, research has shown these area-based measures to be useful for a number of reasons. First, area-based measures provide similar or complementary estimates of socioeconomic disparities in health to many individual measures (29, 33-37). Second, area-based measures of SES may be even more appropriate than individual measures of education and income for individuals under the age of 25 and over the age of 65 (35). Finally, area-based measures provide information regarding neighborhood effects on health above what can be measured from individual SES (38, 39). Area-based measures of SES as described in this chapter are not intended to serve as proxies for individual measures of SES. Rather, they are meant to represent all of the contextual factors affiliated with the neighborhood environment in which one lives. As such, they are not subject to the ecologic fallacy. Instead, the real concern with these measures regards the accuracy with which these neighborhood factors

can be meaningfully captured through census tract data. One portion of this problem is at the census end – can the U.S. Census, using their current sampling procedures, accurately measure SES indicators within census tracts and blocks? The other portion is at the geocoding end – to what degree of accuracy can one geocode data to the appropriate census tract and assign the appropriate measure of area-based SES?

REFERENCES

1. Cromley E, McLafferty S. GIS and public health. New York: The Guilford Press, 2002.
2. Huxhold W. An Introduction to Geographic Information Systems. Oxford: Oxford University Press, 1991.
3. Waller L, Gotway C. Applied Spatial Statistics for Public Health Data. Hoboken: John Wiley & Sons, Inc, 2004.
4. Tele Atlas. Tele Atlas Geocoding Service - Reference Documentation, 2006.
5. U.S. Bureau of the Census. Census 2000 TIGER/Line File Technical Documentation. Washington, D.C.: U.S. Bureau of the Census, 2000.
6. Cayo M, Talbot T. Positional error in automated geocoding of residential addresses. International Journal of Health Geographics 2003;2:10.
7. United States Postal Service. CASS Technical Guide. Available at: (http://www.ribbs.usps.gov/files/cass/technical_guides/casstech.pdf). Accessed December 2006.
8. U.S. Bureau of the Census. Cartographic boundary files. Washington, DC: U.S. Bureau of the Census, 2007.
9. Rushton G, Armstrong M, Gittler J, et al. Geocoding in cancer research: a review. American Journal of Preventive Medicine 2006;30:S16-24.
10. Wiggins L. Using Geographic Information Systems Technology in the Collection, Analysis, and Presentation of Cancer Registry Data: A Handbook of Basic Practices. Springfield, IL: North American Association of Central Cancer Registries, 2002.

11. Bonner MR, Han D, Nie J, et al. Positional accuracy of geocoded addresses in epidemiologic research.[erratum appears in Epidemiology. 2003 Nov;14(6):736]. Epidemiology 2003;14:408-12.

12. Dearwent S, Jacobs R, Halbert J. Locational uncertainty in georeferencing public health datasets. Journal of Exposure Analysis and Environmental Epidemiology 2001;11:329-34.

13. Karimi H. Evaluation of Uncertainties Associated with Geocoding Techniques. Computer-Aided Civil and Infrastructure Engineering 2004;19:170-185.

14. Liadis J. GPS TIGER Accuracy Analysis Tools (GTAAT), Evaluation and Results. In: TIGER Operation Branch Geography Division TUCB, ed, 2000.

15. Nie J, Vito D, Willett N, et al. Validation of TIGER (Topologically Integrated Geographic Encoding and Referencing System) to geocode addresses for epidemiologic research. American Journal of Epidemiology 2001;179:647.

16. Hurley S, Saunders T, Nivas R, et al. Post office box addresses: a challenge for geographic information system-based studies. Epidemiology 2003;14:386-91.

17. Vine M. Geographic information systems: their use in environmental epidemiologic research. Environmental Health Perspectives 1997;105:598-605.

18. McElroy JA, Remington PL, Trentham-Dietz A, et al. Geocoding addresses from a large population-based study: lessons learned. Epidemiology 2003;14:399-407.

19. Oliver M, Matthews K, Siadaty M, et al. Geographic bias related to geocoding in epidemiologic studies. International Journal of Health Geographics 2005;4:29.

20. Whitsel E, Quibrera P, Smith R, et al. Accuracy of commercial geocoding: assessment and implications. Epidemiol Perspect Innov 2006;3:8.

21. Whitsel EA, Rose KM, Wood JL, et al. Accuracy and repeatability of commercial geocoding. American Journal of Epidemiology 2004;160:1023-9.

22. Yang DH, Bilaver LM, Hayes O, et al. Improving geocoding practices: evaluation of geocoding tools. Journal of Medical Systems 2004;28:361-70.

23. Krieger N, Waterman P, Lemieux K, et al. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. American Journal of Public Health 2001;91:1114-6.

24. Ward M, Nuckols J, Giglierano J, et al. Positional accuracy of two methods of geocoding. Epidemiology 2005;16:542-7.

25. Krieger N, Waterman P, Chen J, et al. Zip code caveat: bias due to spatiotemporal mismatches between zip codes and US census-defined geographic areas--the Public Health Disparities Geocoding Project.[see comment]. American Journal of Public Health 2002;92:1100-2.

26. Boscoe F, Kielb C, Schymura M. Assessing and improving census tract completeness. Journal of Registry Management 2002;29:17-20.

27. McLafferty S, Cromley E. Your first mapping project on your own: From A to Z. Journal of Public Health Management Practice 1999;5:76-82.

28. Krieger N, Chen J, Ebel G. Can we monitor socioeconomic inequalities in health? A health survey of US health departments' data collection and reporting practices. Public Heath Reports 1997;112:481-491.

29. Krieger N, Chen J, Waterman P, et al. Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of

area-based measure and geographic level matter? The Public Health Disparities Geocoding Project. American Journal of Epidemiology 2002;156:471-82.

30. Krieger N, Chen JT, Waterman PD, et al. Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: The Public Health Disparities Geocoding Project (US). Journal of Epidemiology & Community Health 2003;57:186-99.

31. Krieger N, Waterman PD, Chen JT, et al. Monitoring socioeconomic inequalities in sexually transmitted infections, tuberculosis, and violence: geocoding and choice of area-based socioeconomic measures--the public health disparities geocoding project (US). Public Health Reports 2003;118:240-60.

32. U.S. Bureau of the Census. Census of Population and Housing, 1990: Summary Tape File 3 Technical Documentation. Washington, DC: U.S. Bureau of the Census, 1991.

33. Datta GD, Colditz GA, Kawachi I, et al. Individual-, neighborhood-, and state-level socioeconomic predictors of cervical carcinoma screening among U.S. black women: a multilevel analysis. Cancer 2006;106:664-9.

34. Diez-Roux AV, Kiefe CI, Jacobs DR, Jr., et al. Area characteristics and individual-level socioeconomic position indicators in three population-based epidemiologic studies.[erratum appears in Ann Epidemiol. 2001 Aug;30(4):924 Note: Roux AV [corrected to Diez-Roux]]. Annals of Epidemiology 2001;11:395-405.

35. Rehkopf D, Haughton L, Chen J, et al. Monitoring socioeconomic disparities in death: comparing individual-level education and area-based socioeconomic measures. American Journal of Public Health 2006;96:2135-8.

36. Robert SA, Strombom I, Trentham-Dietz A, et al. Socioeconomic risk factors for breast cancer: distinguishing individual- and community-level effects. Epidemiology 2004;15:442-50.

37. Subramanian S, Chen J, Rehkopf D, et al. Comparing individual- and area-based socioeconomic measures for the surveillance of health disparities: A multilevel analysis of Massachusetts births, 1989-1991.[see comment]. American Journal of Epidemiology 2006;164:823-34.

38. Diez-Roux AV. Bringing context back into epidemiology: variables and fallacies in multilevel analysis. American Journal of Public Health 1998;88:216-22.

39. MacIntyre S, Ellaway A. Ecological approaches: Rediscovering the role of the physical and social environment. In: Berkman L, Kawachi I, eds. Social Epidemiology. New York: Oxford University Press, 2000:332-348.

Chapter 4

Geocoded Registry Data and the Use of Area-Based Measures of Socioeconomic Status: Misclassification and Steps to Minimize its Effects

Kevin C. Ward[1,2], Michael Goodman[1,2], Lance Waller[3], Jonathan Liff[1,2], Joseph Lipscomb[4], John L. Young, Jr. [1,2]

[1]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA.

[2] Metropolitan Atlanta and Rural Georgia SEER Registry, Georgia Center for Cancer Statistics, Atlanta, GA.

[3]Department of Biostatistics, Rollins School of Public Health, Emory University, Atlanta, GA.

[4]Department of Health Policy and Management, Rollins School of Public Health, Emory University, Atlanta, GA.

*Correspondence to: Kevin C. Ward, MPH, Georgia Center for Cancer Statistics, Emory University, 1462 Clifton Road, NE, 5th Floor, Atlanta, GA. 30322 (email: kward@sph.emory.edu)

Key Words: geocoding, socioeconomic status, area-based measures, misclassification

ABSTRACT

Background: Population-based disease registries routinely geocode patient addresses and link the results with census files to obtain area-based measures of socioeconomic status (SES); however the validity of information obtained by this process remains unclear. This study examines the degree of misclassification that may affect area-based measures of SES in studies utilizing registry data. Steps to improve the completeness and accuracy of geocoding are also explored.

Methods: Residential addresses of all incident cancer cases between 1996 and 2000 from the Metropolitan Atlanta and Rural Georgia SEER Registry were reviewed. Addresses that were not successfully geocoded to the exact street address were systematically processed using a variety of techniques to improve geocoding outcome. The extent of misclassification was assessed by comparing the area-based poverty measures assigned to the unsuccessfully geocoded addresses pre- and post-cleanup. A random sample of successfully geocoded addresses was also evaluated for positional error using data from a hand-held GPS device as the gold standard.

Results: Steps to improve geocoding outcomes were successful for 85% of all addresses. Results were superior for metropolitan (91%) relative to rural (55%) addresses. Misclassification by area-based measure of SES varied depending on the location of the address (metropolitan or rural), by the choice of geographic unit (census tract or census

block) and by the address component upon which geocodes were assigned (street address, residential ZIP code, PO Box ZIP code).

Conclusions: Effective procedures exist for improving geocoding outcomes, especially in metropolitan areas. Researchers utilizing geocoded registry data must have information regarding the completeness and accuracy of the assigned geocodes. In the absence of this information, a proper interpretation of epidemiological findings may be problematic.

INTRODUCTION

Population-based disease registries collect a wealth of information regarding demographic and clinical patient characteristics (1, 2). They are generally limited, however, with respect to other data. Cancer registries in the United States, for example, have for years tried to collect data on birthplace, religion, occupation, tobacco history, alcohol use, and family history of cancer (3) but with limited success as this information is typically absent from the medical record. Individual measures of socioeconomic status (SES) are also missing from cancer data (3). In the absence of individual measures of SES, one alternative is to collect area-based SES information through the process of geocoding.

Geographic information systems (GIS) and the process of geocoding serve as valuable resources for researchers utilizing registry data (4-7) and national agencies in the United States have recommended their increased utilization in existing surveillance systems to enhance surveillance (8, 9). As described in various texts (10-12), current GIS technologies facilitate the assignment of geographic coordinates to the street level, support the linkage of geocoded data to census summary files containing area-based measures of SES, and provide a means for spatial analysis. While area-based measures may serve as proxies for individual SES, they have been shown to be meaningful indicators in their own right (6, 7, 13, 14). The neighborhood in which one lives captures aspects of living conditions not necessarily defined by individual measures; it is a relevant characteristic that applies to all of its residents regardless of age and gender, and it is a moderately stable measure of socioeconomic conditions (13).

Useful data informing studies on socioeconomic factors and health come from the Public Health Disparities Geocoding Project, which was designed to identify the optimal area-based measures of SES (e.g., occupation, income, education, crowding, poverty) and corresponding levels of geography (e.g., zip code, census tract, census block) (15, 16). This research suggests that measures of economic deprivation at the level of the census tract, such as the percent of the census population living below the poverty level, are most effective for evaluating health disparities using linked census data (17). These measures demonstrate consistent gradients across population subgroups, are robust across a range of disease outcomes, allow for maximal linkage, and are easy to understand and explain. While conceptually intuitive, the process of linking these data sources has limitations due to inherent difficulties in assigning geocoded information to the appropriate geographic level (18-29).

Geocoding is the process of assigning geographic coordinates, typically latitude and longitude, or census tract, to a street address of interest. Geocoding match rates are reported to range between 20 and 100 percent (24, 28-33) in different populations and using different definitions of success. When an exact street address cannot be geocoded, the centroid of the residence ZIP code is often used instead (34). Variation in geocoding success has been attributed to problems with the input address data (26, 35, 36), limitations of the underlying reference database (18, 19, 21, 23, 25), methodological issues related to the interpolation process for locating the physical address along a reference street segment (19, 21, 26), and the use of Post Office (PO) Box and rural route addresses (20, 24, 37, 38). Even a 'successfully' geocoded address may be subject to

positional inaccuracy (18-20, 27, 28, 30). All of these process limitations may lead to misclassification in studies utilizing geocoded data.

The primary aims of the present study are to investigate possible methods to increase the frequency of successful geocoding and to quantify the degree of misclassification of one measure of area-based SES that may arise from errors in geocoding. Misclassification may result from assigning a particular address to a wrong census tract or a wrong block group, which may lead to inferring a wrong area-based SES category. The area-based socioeconomic measure of interest in this study is the "percentage of persons living below the U.S. poverty line" as suggested by Krieger et al. (15).

MATERIALS AND METHODS

Population

The population for this study included all incident cases of cancer diagnosed between January 1, 1996 and December 31, 2000, and reported to the Metropolitan Atlanta and Rural Georgia (MARGA) Surveillance, Epidemiology, and End Results (SEER) Registry. This population-based registry was established in 1976 as part of the National Cancer Institute's SEER Program (39) to collect all incident cancers on an annual basis in a 15 county area of Georgia: five counties in metropolitan Atlanta, and ten rural counties located in the central part of the state. Addresses at the time of diagnosis for all eligible cancer records were geocoded by a single commercial vendor using data from the 2000 U.S. decennial census. The certainty of each geocoded record, as provided by the commercial vendor, was captured in the registry database using the codes established by

the North American Association of Central Cancer Registries (3). The census tract certainty codes were assigned as follows. A certainty code of 1 was assigned to records where the geographic coordinates from the geocoding process were based on an exact street address that was located in the underlying street reference database of the commercial vendor. For the purpose of this study, these were the only records considered successfully geocoded. Codes 2 through 4 were used when the exact street address was not locatable in the reference database but a residential ZIP code (ZIP, ZIP+2 or ZIP+4) was present. In this situation, the geographic coordinates assigned were based on the use of a ZIP code centroid (34). Code 5 was used when a PO Box, rather than a residential address, was provided as the address at diagnosis. In this case, the geographic coordinates assigned were based on the ZIP code centroid of the PO Box. Records with a certainty code of 9 could not be geocoded at any level.

Review of unsuccessfully geocoded cases

During the study period, 10 percent of the MARGA SEER Registry data were not geocoded to an exact street address, resulting in a certainty code other than 1. For those records, an attempt was made to identify alternative methods of obtaining more complete and accurate addresses. Registry data were divided into two groups: one containing addresses with presumed street level errors and the other containing PO Box and rural route addresses. Due to the inherent differences between these groups, separate methods were developed for improving their respective geocoding outcomes.

As shown in Figure 4.1, all records in the group with presumed street level errors were batch processed using six different services: address standardization to the official United States Postal Service address format (40); Coding Accuracy Support System (CASS) certified address matching (41); a fee-based reverse directory service; a fee-based identification tracking service; linkage to mortality records; and linkage to voter records for the state of Georgia. The first two services attempted to modify records to correct errors in the existing address while the other services searched for a more complete and/or more accurate address from an alternate source. Probabilistic record linkage software was used to ensure an alternate source address corresponded with the actual address of interest.

After the batch procedures were completed, records were resent to the Registry's commercial vendor for a second geocoding attempt. Records that still remained unsuccessfully geocoded were then manually matched against the United States Census Bureau Topologically Integrated Geographic Encoding and Reference (TIGER) files (42) for multiple years (1995, 1998, and 2000). A 20 percent random sample of records not geocoded following this manual TIGER file review were sent back to the original reporting facility in an effort to obtain better address information from the hospital medical record or billing system. In the event new address information was identified, records were again resent to the commercial vendor for a final geocoding attempt.

Address standardization, CASS certified address matching, and manual review of TIGER files are of no value when there is only a PO Box or rural route address. All records in

this category were first processed through the identification tracking service, which provided an address history and a corresponding date range for each location. The goal of this process was to identify the residential address at the time of diagnosis. Following the recommendation by Hurley et al. (20), PO Box addresses were also sent to the Postmaster of the corresponding city with a request for the physical street address of the box owner. All PO Box and rural route records were also matched against state voter files (1994-2002). Voter registration in the state of Georgia requires the listing of a residential address. If an individual with a PO Box address in the Registry database voted both before and after the date of diagnosis and the addresses on the voter files were the same, the voter file address was presumed to be the street address at diagnosis. The same applied to rural routes although this was a less common occurrence. Finally, all records were matched against both statewide mortality files and the reverse directory service. Since it was not possible to know if the address at the time of death or the current address from the reverse directory service truly corresponded to the address of the individual at the time of diagnosis, information gained from these sources was used for validation only.

Geocoding results both pre- and post-address cleanup at the level of the census tract and census block group were linked to the Census 2000 Summary File 3 (43) to obtain the percent of the population at each level of geography living below the poverty level. Misclassification was assessed by comparing pre- and post-cleanup results.

Evaluation of misclassification among successfully geocoded cases

A stratified random sample of 150 metropolitan and 150 urban/rural successfully geocoded cancer cases (certainty code 1) from the study population were selected for field geocoding. Field staff trained in GPS data collection traveled to all 300 address locations and obtained latitude and longitude coordinates using a hand-held GPS device (Garmin, GPSMAP 76S) from the location of the residence mailbox. In the event an address could not be located (i.e. the address no longer existed or could not be found using all available map sources), the address closest to the original location of interest was selected and geocoded. All GPS devices were regularly calibrated and repeated measures from the exact same location were obtained prior to each trip. To examine inter-observer agreement, a 10 percent random sample of locations was geocoded independently by two separate field investigators.

Latitude and longitude coordinates for all 300 addresses from both the GPS device and the Registry's commercial vendor were imported into the GIS and mapping software, ArcGIS 9, developed by ESRI of Redlands, California. Block group shape files from ESRI for all counties in Georgia were loaded into ArcGIS. A point-in-polygon technique was used to place each set of geographic coordinates for an address in its corresponding block group and census tract. Geocoding results from both sources were linked to census data to obtain the percent of the population in that geographic unit living below the poverty level.

Data analysis

The percent of the population living below the poverty level was classified first as a dichotomous variable with the cutoff point of ≥20 percent; and then as a 3-category variable further dividing the <20 percent group into subcategories of <10 percent and 10-19.9 percent. Areas with ≥20 percent of the population living below the poverty threshold meet the federal definition of a "poverty area" (44). For the year 2000, the poverty threshold was defined as $17,463 for a family of 4 with 2 children under the age of 18 (45). The characteristics of successfully geocoded data were compared to those not geocoded based on an exact street address at the point of study initiation using frequency analyses and corresponding chi-square tests. The extent of misclassification was assessed by comparing the geocoded data before and after the clean-up process. The results of this comparison were expressed as the percent of cases assigned to a wrong census tract, wrong census block group, or wrong area-based poverty level. Each percentage estimate was accompanied by a 95% confidence interval (CI). Data from the field study were also assessed for misclassification using the hand-held GPS measurements as the gold standard.

To further examine the potential for misclassification due to geocoding positional inaccuracy in the presence of a successfully geocoded address, we used Vincenty's formula (46) to calculate the geodesic distances between the pairs of latitude/longitude coordinates obtained from the field study and from the commercial vendor. This method used an ellipsoidal model of the earth. Positional errors were then quantified in meters

and divided into categories of 0-50 m; 51-100 m; 101-200 m; 201-1,000 m; and 1,001+ m once again using the hand-held GPS measurements as the gold standard.

RESULTS

A total of 53,563 cancer cases were reported to the MARGA SEER Registry between 1996 and 2000. Of those, 48,209 cases (90%) were geocoded based on a complete, valid street address. Selected case characteristics and their relation to geocoding success are summarized in Table 4.1. A comparison of successfully geocoded records (certainty code 1) to those that were unsuccessful (certainty codes 2-9) demonstrated small, albeit significant at the 0.05 level, differences with respect to gender, age, frequency of unknown stage of disease at diagnosis (9% vs. 11%), and percent late stage disease (44% vs. 48%). Other differences were more pronounced. Black cancer patients constituted a higher proportion of the unsuccessfully geocoded addresses relative to the addresses successfully geocoded to the street level (34% vs. 26%) as did cases living in non-metropolitan areas (16% vs. 4%) and poverty areas (25% vs. 16%).

Nearly 9 percent of the records from Metropolitan SEER and 32 percent of the records from Rural SEER were not originally geocoded with certainty code 1. Rural SEER had a much larger percentage of PO Box (9.5% vs. 1.1%) and rural route addresses (6.4% vs. <0.1%) relative to Metropolitan SEER. Table 4.2 summarizes success of the clean-up procedures by address type. As expected, clean-up was most effective for street addresses, followed by PO Box addresses and then rural routes. Regardless of the address type, clean-up for metropolitan records yielded better results. At the conclusion

of all steps taken to clean and enhance the address data, only 1.5 percent of the total records remained unsuccessfully geocoded to a street level address. The percentage was substantially lower in Metropolitan (0.8%) than in Rural SEER (14.2%), although the actual numbers were slightly higher. This was largely attributed to difficulty in geocoding rural route addresses, an observation that has been reported in prior studies (38).

Table 4.3 presents a summary of the batch clean-up processes used alone and in combination with other methods. The results of each individual process are shown on the diagonal, with the denominator of each percent representing records that were cleaned to allow successful geocoding. The tracking service alone, for example, provided 75 percent of the successful results. In each column, cell values represent the additional percent gain from adding a second batch process to the one reported on the diagonal. As an example, adding CASS as a second source to the tracking service provided an additional 17.6 percent gain for a combined 93 percent.

A comparison of vendor provided coordinates to those obtained by GPS measurement among 300 addresses successfully geocoded to the street level demonstrated a median error of 34.7 meters (range: 1-1,322) for Metropolitan SEER and 195.1 meters (range: 8-6,216) for Rural SEER. The distributions were skewed due to a small proportion of extreme values. While 91 percent of Metropolitan SEER addresses were geocoded to within 100 meters, only 33 percent of Rural SEER addresses achieved the same level of concordance with 17 percent of errors in excess of 1,000 meters.

Misclassification of the 300 field addresses with respect to geographic units (census tract and block group) and an area-based measure of poverty is presented in the first two data columns of Table 4.4. In Metropolitan SEER, 3.3 percent (95% CI 1.5-7.6) of the cases successfully geocoded at the street level were misclassified into a wrong census tract or block group by the commercial vendor. The misclassification of the area-based measure of poverty due to positional errors ranged from 1.3 to 2.0 percent at the geographic level of the census tract. In Rural SEER, census tract and block group misclassification of successfully geocoded street level addresses occurred in 5.3 percent (95% CI 2.8-10.2) and 12 percent (95% CI 7.7-18.2) of cases respectively, while the misclassification by the area-based measure of poverty was comparable to that found in Metropolitan SEER, at least at the level of the census tract.

Misclassification of addresses originally geocoded to the ZIP code centroid was far more pronounced (Table 4.4). Nearly 60 percent of Metropolitan SEER cases and 27 percent of Rural SEER cases geocoded to the ZIP code were placed into a wrong census tract. The percentages misclassified were even greater at the block group level (69% vs. 60%, respectively). Using the census tract poverty variable divided into three levels, 21 percent of the Metropolitan SEER data were misclassified compared to 15 percent in Rural SEER. PO Box addresses presented a larger problem, particularly in the Metropolitan SEER area. Using the PO Box address instead of the actual residential address resulted in 82 percent misclassification of cases in Metropolitan SEER and 22 percent in Rural SEER. Misclassification by block group was again more extensive in

both areas. When linked to census data, the census tract poverty variable (3 groups) was misclassified in 44 percent of Metropolitan and 8 percent of Rural SEER cases.

DISCUSSION

Our results indicate that some degree of positional error in geocoding based research is inevitable. The need for concern about this error is largely dictated by the research question of interest. If, for example, the goal is to produce rates of disease at a small level of geography such as the census tract, misclassification to a wrong geographic unit is of primary concern. If the goal is to assign exposure status based on the distance from a specific location, positional accuracy in meters matters greatly. If the objective is to link geocoded data to area-based measures of SES, misclassification by the SES measure is what matters the most. Varying degrees of geocoding match rates, historical data geocoded by multiple vendors, census definitions that change over time (i.e. a poverty area in 1990 may no longer be a poverty area in 2000), and the common use of geocoding ZIP code centroids when street level addresses can not be located, can all produce high levels of misclassification that can bias study results.

At the initiation of this study, 9 percent of the Metropolitan SEER data and 32 percent of the Rural SEER data (10% overall) could not be geocoded based on a complete, valid street address. At the end of the clean-up process, only 1 percent of Metropolitan data and 14 percent of Rural data (1.5% overall) remained in this category. Rural routes posed the greatest geocoding challenge because they are generally not contained within most geocoding reference databases. During the course of this study we were able to

evaluate and compare different processes for improving historical address information. Ensuring that records meet Postal Service address standards (40) was a simple first step that should be applied to all incoming registry data. The fee-based tracking service was the most effective single approach while voter records were especially useful for obtaining the residential address for a PO Box mailing address. Researchers must clearly consider the availability and costs associated with different batch processing methods; however, overall costs were generally minimal.

A second objective of this study was to quantify the degree of information bias that might arise from errors in geocoded data. Smaller levels of geography produced greater potential for misclassification. As census tracts and block groups are based on population density, they are smaller and more numerous in metropolitan areas. These features account for the more common assignment to erroneous tracts and block groups of addresses originally geocoded to ZIP code centroids in Metropolitan relative to Rural SEER. The degree of positional error in meters was much greater in rural areas relative to metropolitan areas, an observation reported in several previous studies (18, 27, 30). This finding can be attributed at least in part to the fact that reference database files used for geocoding are more complete in densely populated areas and are more accurate in areas with short street segments. Our finding that in rural areas successfully geocoded addresses were more frequently assigned to the wrong tract and block group is supported by this information.

With respect to the area-based poverty measure, misclassification depended on the specific geographic location of interest. This was largely explained by the relative heterogeneity or homogeneity of socioeconomic conditions within a given county (Figure 4.2). All of the census tracts in two counties in Rural SEER were classified in the highest poverty category, while three counties in Metropolitan SEER had fewer than 4 percent of their tracts classified as such. Future studies may benefit from the use of mapping software that could help identify counties more susceptible to misclassification. Researchers could then expend additional effort to validate geocoded addresses within these areas.

The results of our analyses should be viewed in the broader context of total misclassification affecting registry data. Misclassification exists at multiple levels (street, residence, ZIP centroid, PO ZIP centroid) and to varying degrees in geocoded data. If an entire dataset from a metropolitan area were geocoded using residence ZIP code centroids, an estimated 58-61 percent of the data would be misclassified by census tract, 67-70 percent by census block group, and 7-22 percent by census tract derived poverty measure (Table 4.4). This level of misclassification would not be acceptable to most researchers. If all of the data were instead geocoded to a street level address, an estimated 1-8 percent of the data would be misclassified by census tract and block group, and only 0.4-6.0 percent by the same poverty measure. The actual misclassification in most U.S. datasets probably lies somewhere between these two extremes with a large percentage of the data geocoded to the street level and a smaller portion based on zip code centroids.

Despite the potentially useful findings in this study, our analyses have several notable limitations. First, it is unclear whether the results obtained in Georgia are applicable to similar registries elsewhere in the country. The street level and PO Box misclassification at selected levels of geography (census tract and block group) were similar to those found in prior studies in other areas (20, 47), but this does not ensure consistency at the ZIP code centroid level. The relative homogeneity of socioeconomic conditions (i.e. poverty areas vs. non-poverty areas) found within most counties in the SEER areas of GA may not hold in other states. Second, despite the extensive data cleanup efforts we were unable to successfully geocode 1.5 percent of addresses. This percentage was much higher (14%) in Rural SEER, leaving a greater degree of uncertainty with rural data. Third, our area-based poverty measure was based on the address at diagnosis between 1996 and 2000 but was linked to the 2000 Census data. This could introduce additional misclassification as the socioeconomic conditions of an area can change over time. In the near future, the U.S. Census Bureau will be releasing 5-year rolling estimates that will help address this concern. Finally, we used the GPS measurements as our gold standard for calculating positional error. The GPS measurements themselves and the locations from which measurements were taken are also subject to error. Nevertheless, our results were in line with the positional misclassification reported in previous research (47).

Taking steps to understand and minimize bias related to area-based exposure assignment is a critical part of any study utilizing GIS. A comprehensive practice guideline document for GIS technology is available from the North American Association of

Central Cancer Registries (36). Other reference materials are also available to help maximize geocoding match results (24, 33). These processes should be followed by the use of an accurate, complete, and reliable geocoding system. Recent studies have suggested parcel matching as a more accurate alternative to interpolation protocols that are carried out by most geocoding applications (19, 26, 30).

CONCLUSION

Based on our findings, it is imperative that investigators both understand and report the percent of data geocoded based on ZIP centroids as the level of potential misclassification within the study results will increase as this percentage increases. We also suggest caution in the use of PO Box addresses for the assignment of area-based SES measures as they are subject to the highest degree of misclassification in metropolitan areas. In addition, we recommend a close examination of rural addresses as they are more likely to contain a large percent of unsuccessfully geocoded records and possess greater positional errors in the successfully geocoded records. In conclusion, researchers utilizing geocoded registry data must have information regarding the completeness and accuracy of the assigned geocodes. In the absence of this information, a proper interpretation of epidemiological findings is problematic.

FIGURE 4.1 Process diagram for systematic re-evaluation of unsuccessfully geocoded addresses with presumed street level errors

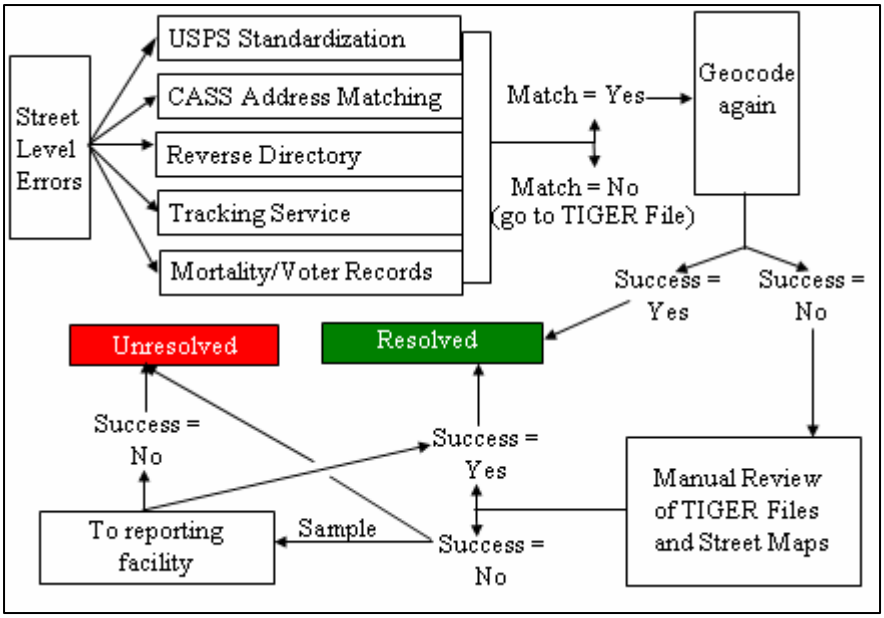

FIGURE 4.2 Metropolitan Atlanta and Rural Georgia SEER census tract poverty areas (>=20% of the population living below the poverty level)

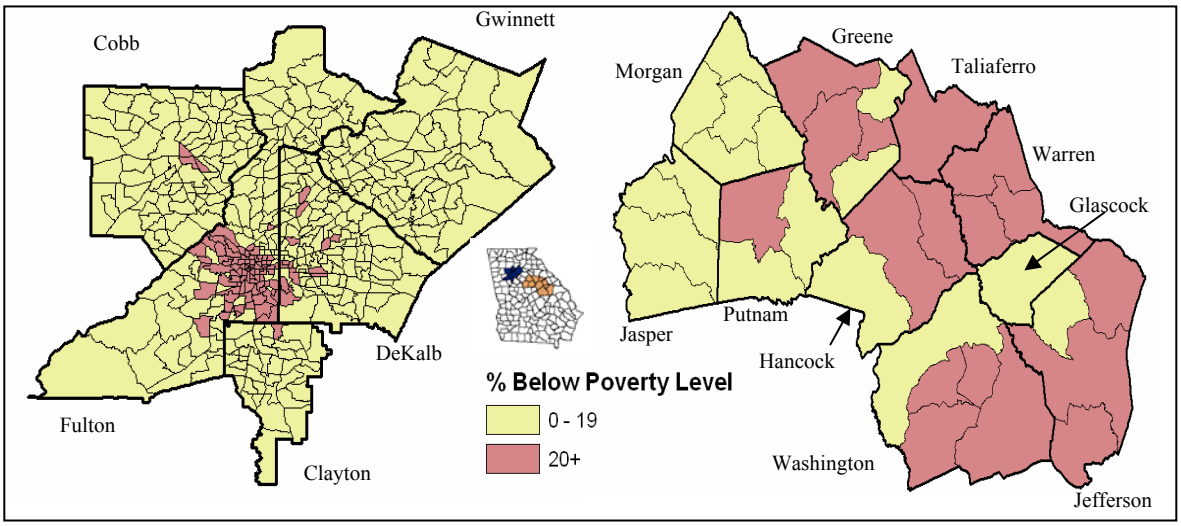

TABLE 4.1 Baseline characteristics of Metropolitan Atlanta and Rural Georgia SEER Registry cases according to initial geocoding status, 1996-2000

Characteristic		Geocoded to Street Address (n = 48,209)		Not Geocoded to Street Address (n = 5,354)		p value
		No.	%	No.	%	
Sex						0.0355
	Male	23,298	48.3%	2,669	49.9%	
	Female	24,911	51.7%	2,685	50.1%	
Race*						<0.0001
	White	34,718	72.3%	3,406	64.0%	
	Black	12,506	26.0%	1,815	34.1%	
	Other	790	1.6%	98	1.8%	
Age						0.0046
	<45	6,936	14.4%	789	14.7%	
	45-64	19,368	40.2%	2,028	37.9%	
	65+	21,905	45.4%	2,537	47.4%	
Stage						<0.0001
	Known	43,798	90.9%	4,759	88.9%	
	Unknown	4,411	9.1%	595	11.1%	
Stage*						<0.0001
	Early	24,345	55.6%	2,488	52.3%	
	Late	19,453	44.4%	2,271	47.7%	
Geographic Area						<0.0001
	Metro	46,354	96.2%	4,486	83.8%	
	Urban	1,220	2.5%	347	6.5%	
	Rural	635	1.3%	521	9.7%	
% Pop below Poverty*~						<0.0001
	0-9.9	29,143	60.5%	2,598	49.4%	
	10.19.9	11,445	23.7%	1,317	25.1%	
	20+	7,621	15.8%	1,342	25.5%	

* Counts, percentages and p values exclude records with unknown values
~Based on Zip Code Tabulation Area

TABLE 4.2 Distribution of initially unsuccessfully geocoded records ultimately successfully cleaned and geocoded, by initial address type, 1996-2000

	Metro SEER (N=4,486)		Rural SEER (N=868)		All GA SEER (N=5,354)	
Records Cleaned & Geocoded	No.	Percent Successful	No.	Percent Successful	No.	Percent Successful
PO Box	481	83.1%	166	64.3%	647	77.3%
Rural Route	5	71.4%	31	17.7%	36	19.8%
Street Address	3,590	92.1%	283	65.1%	3,873	89.3%

TABLE 4.3 Batch processed records - percent gain in successfully geocoded addresses resulting from the addition of a subsequent source of information to an existing source of information*

Additional Source	**Existing Source**					
	Tracking	Reverse Dir.	CASS	USPS	Voter	Mortality
Tracking	75.10%	58.70%	38.10%	53.70%	39.80%	62.60%
Reverse Dir.	3.20%	19.60%	10.10%	14.40%	7.20%	17.50%
CASS	17.60%	45.10%	54.60%	22.20%	33.10%	46.60%
USPS	11.10%	27.30%	0.00%	32.50%	19.50%	27.00%
Voter	6.60%	29.50%	20.40%	29.00%	41.90%	35.30%
Mortality	5.60%	16.00%	10.10%	12.70%	11.50%	18.10%

* Numbers on the diagonal represent the percent of successfully geocoded records (N=4556) obtained by any single batch source. Row percents off the diagonal display the gain received by adding the row level variable as another source of information to the existing column level source.

TABLE 4.4 Geocoding misclassification by geocode type (street, residence ZIP centroid, PO ZIP centroid), census geography, and poverty group

	Street Level		Residence ZIP Centroid		PO ZIP Centroid	
Misclassification by:	Metro SEER	Rural SEER	Metro SEER	Rural SEER	Metro SEER	Rural SEER
Tract						
Percent	3.3%	5.3%	59.5%	27.0%	81.8%	21.7%
Confidence Interval	(1.5, 7.6)	(2.8,10.2)	(57.9, 61.2)	(22.4, 32.2)	(78.0, 85.0)	(16.1, 28.6)
Block Group						
Percent	3.3%	12.0%	68.9%	59.3%	91.2%	66.3%
Confidence Interval	(1.5, 7.6)	(7.7,18.2)	(67.3, 70.4)	(53.7, 64.7)	(88.3, 93.5)	(58.7, 73.1)
Tract Poverty 2-groups*						
Percent	1.3%	1.3%	8.0%	11.9%	18.9%	5.4%
Confidence Interval	(0.4, 4.7)	(0.4, 4.7)	(7.2, 9.0)	(8.8, 16.0)	(15.6, 22.6)	(2.9, 10.0)
Tract Poverty 3-groups#						
Percent	2.0%	1.3%	20.9%	15.1%	43.8%	7.8%
Confidence Interval	(0.7, 5.7)	(0.4, 4.7)	(19.6, 22.3)	(11.6, 19.5)	(39.4, 48.3)	(4.7, 12.9)
Blk Grp Poverty 2-groups*						
Percent	0.7%	4.0%	11.5%	35.7%	20.5%	32.5%
Confidence Interval	(0.2, 3.6)	(1.9, 8.4)	(10.4, 12.6)	(30.6, 41.3)	(17.1, 24.4)	(25.8, 40.1)
Blk Grp Poverty 3-groups#						
Percent	1.3%	5.3%	26.4%	38.7%	48.1%	32.5%
Confidence Interval	(0.4, 4.7)	(2.8,10.2)	(25.0, 27.9)	(33.4, 44.3)	(43.6, 52.6)	(25.8, 40.1)

* Census assigned poverty [% living below poverty line]: (0-19.9, 20+)

Census assigned poverty [% living below poverty line]: (0-9.9, 10-19.9, 20+)

REFERENCES

1. Johnson C. SEER Program Coding and Staging Manual 2004, Rev 1. Bethesda, MD: National Cancer Institute, NIH Publication Number 04-5581, 2004.
2. Krieger N, Chen J, Ebel G. Can we monitor socioeconomic inequalities in health? A health survey of US health departments' data collection and reporting practices. Public Heath Reports 1997;112:481-91.
3. Havener L, Hultstrom D, editors. Standards for Cancer RegistriesVolume II: Data Standards and Data Dictionary. Springfield, IL: North American Association of Central Cancer Registries, 2006.
4. Chen F, Breiman R, Farley M, et al. Geocoding and linking data from population-based surveillance and the US Census to evaluate the impact of median household income on the epidemiology of invasive Streptococcus pneumonia infections American Journal of Epidemiology 1998;148:1212-8.
5. Krieger N, Chen J, Waterman P, et al. Race/ethnicity and changing US socioeconomic gradients in breast cancer incidence: California and Massachusetts, 1978-2002 (United States). Cancer Causes & Control 2006;17:217-26.
6. Rehkopf D, Haughton L, Chen J, et al. Monitoring socioeconomic disparities in death: comparing individual-level education and area-based socioeconomic measures. American Journal of Public Health 2006;96:2135-8.
7. Subramanian S, Chen J, Rehkopf D, et al. Comparing individual- and area-based socioeconomic measures for the surveillance of health disparities: A multilevel

analysis of Massachusetts births, 1989-1991.[see comment] American Journal of Epidemiology 2006;164:823-34.

8. National Cancer Institute Cancer Surveillance Research Implementation Plan (1999). Available at: http://cancercontrol.cancer.gov/sig/index.htm. Accessed February, 2007.

9. U.S. Department of Health and Human Services. Healthy People 2010. 2nd ed. With Understanding and Improving Health and Objectives for Improving Health. Washington, DC: US Government Printing Office, 2000.

10. Cromley E, McLafferty S. GIS and public health. New York: The Guilford Press, 2002.

11. Huxhold W. An Introduction to Geographic Information Systems Oxford: Oxford University Press, 1991.

12. Waller L, Gotway C. Applied Spatial Statistics for Public Health Data. Hoboken: John Wiley & Sons, Inc, 2004.

13. Krieger N. Measuring Social Class in US Public Health Research. Annual Review of Public Health 1997;18:341-78.

14. Krieger N, Chen JT, Waterman PD, et al. Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: The Public Health Disparities Geocoding Project (US). Journal of Epidemiology & Community Health 2003;57:186-99.

15. Krieger N, Chen J, Waterman P, et al. Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of

area-based measure and geographic level matter?: the Public Health Disparities Geocoding Project. American Journal of Epidemiology 2002;156:471-82.

16. Krieger N, Chen JT, Waterman PD, et al. Painting a truer picture of US socioeconomic and racial/ethnic health inequalities: the Public Health Disparities Geocoding Project. American Journal of Public Health 2005;95:312-23.

17. Krieger N, Chen JT, Waterman PD, et al. Race/ethnicity, gender, and monitoring socioeconomic gradients in health: a comparison of area-based socioeconomic measures--the public health disparities geocoding project. American Journal of Public Health 2003;93:1655-71.

18. Bonner MR, Han D, Nie J, et al. Positional accuracy of geocoded addresses in epidemiologic research.[erratum appears in Epidemiology. 2003 Nov;14(6):736]. Epidemiology 2003;14:408-12.

19. Dearwent S, Jacobs R, Halbert J. Locational uncertainty in georeferencing public health datasets. Journal of Exposure Analysis and Environmental Epidemiology 2001;11:329-34.

20. Hurley S, Saunders T, Nivas R, et al. Post office box addresses: a challenge for geographic information system-based studies. Epidemiology 2003;14:386-91.

21. Karimi H. Evaluation of Uncertainties Associated with Geocoding Techniques. Computer-Aided Civil and Infrastructure Engineering 2004;19:170-85.

22. Krieger N, Waterman P, Lemieux K, et al. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. American Journal of Public Health 2001;91:1114-6.

23. Liadis J. GPS TIGER Accuracy Analysis Tools (GTAAT), Evaluation and Results. In: TIGER Operation Branch Geography Division TUCB, ed. 2000.

24. McElroy JA, Remington PL, Trentham-Dietz A, et al. Geocoding addresses from a large population-based study: lessons learned. Epidemiology 2003;14:399-407.

25. Nie J, Vito D, Willett N, et al. Validation of TIGER (Topologically Integrated Geographic Encoding and Referencing System) to geocode addresses for epidemiologic research. American Journal of Epidemiology 2001;179:647.

26. Rushton G, Armstrong M, Gittler J, et al. Geocoding in cancer research: a review. American Journal of Preventive Medicine 2006;30:S16-24.

27. Ward M, Nuckols J, Giglierano J, et al. Positional accuracy of two methods of geocoding. Epidemiology 2005;16:542-7.

28. Whitsel E, Quibrera P, Smith R, et al. Accuracy of commercial geocoding: assessment and implications. Epidemiol Perspect Innov 2006;3:8.

29. Whitsel EA, Rose KM, Wood JL, et al. Accuracy and repeatability of commercial geocoding. American Journal of Epidemiology 2004;160:1023-9.

30. Cayo M, Talbot T. Positional error in automated geocoding of residential addresses. International Journal of Health Geographics 2003;2:10.

31. Geronimus A. Poverty, time and place: variation in excess mortality across selected US populations, 1980-1990. Journal of Epidemiology and Community Health 1999;53:325-34.

32. Oliver M, Matthews K, Siadaty M, et al. Geographic bias related to geocoding in epidemiologic studies. International Journal of Health Geographics 2005;4:29.

33. Yang DH, Bilaver LM, Hayes O, et al. Improving geocoding practices: evaluation of geocoding tools. Journal of Medical Systems 2004;28:361-70.

34. Tele Atlas. Tele Atlas Geocoding Service - Reference Documentation, 2006.

35. Goldberg D. From text to geographic coordinates: The current state of geocoding. URISA (www.urisa.org/goldberg) 2006.

36. Wiggins L, . Using Geographic Information Systems Technology in the Collection, Analysis, and Presentation of Cancer Registry Data: A Handbook of Basic Practices. Springfield, IL: North American Association of Central Cancer Registries, 2002.

37. Krieger N, Waterman P, Chen J, et al. Zip code caveat: bias due to spatiotemporal mismatches between zip codes and US census-defined geographic areas--the Public Health Disparities Geocoding Project.[see comment]. American Journal of Public Health 2002;92:1100-2.

38. Vine M. Geographic information systems: their use in environmental epidemiologic research. Environmental Health Perspectives 1997;105:598-605.

39. Ries L, Harkins D, Krapcho M, et al. SEER Cancer Statistics Review, 1975-2003. Bethesda, MD: National Cancer Institute, 2006.

40. United States Postal Service. Postal Addressing Standards, Publication 28. Available at: (http://pe.usps.gov/cpim/ftp/pubs/Pub28/pub28.pdf). Accessed December 2006.

41. United States Postal Service. CASS Technical Guide. Available at: (http://www.ribbs.usps.gov/files/cass/technical_guides/casstech.pdf). Accessed December 2006.

42. U.S. Bureau of the Census. Census 2000 TIGER/Line File Technical Documentation. Washington, D.C.: U.S. Bureau of the Census, 2000.

43. U.S. Bureau of the Census. 2000 Census Population and Housing, Summary File 3: Technical Documentation. 2002.

44. U.S. Bureau of the Census. Poverty Areas. Available at: (http://www.census.gov/hhes/www/poverty/definitions.html). Accessed December 2006.

45. U.S. Bureau of the Census. Poverty Thresholds 2000. Available at: (http://www.census.gov/hhes/www/poverty/threshld/thresh00.html). Accessed December 2006.

46. Vincenty T. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. Survey Review 1975;176:88-93.

47. Ratcliff J. On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units. International Journal of Geographical Information Science 2001;15:473-85.

Chapter 5

Determining Population-Based Relative Survival: The Importance of Using Local Area and SES-Specific Measures of Background Mortality

Kevin C. Ward[1,2], Michael Goodman[1,2], Jonathan Liff[1,2], Lance Waller[3], Joseph Lipscomb[4], John L. Young, Jr. [1,2]

[1]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA.

[2] Metropolitan Atlanta and Rural Georgia SEER Registry, Georgia Center for Cancer Statistics, Atlanta, GA.

[3]Department of Biostatistics, Rollins School of Public Health, Emory University, Atlanta, GA.

[4]Department of Health Policy and Management, Rollins School of Public Health, Emory University, Atlanta, GA.

*Correspondence to: Kevin C. Ward, MPH, Georgia Center for Cancer Statistics, Emory University, 1462 Clifton Road, NE, 5th Floor, Atlanta, GA. 30322 (email: kward@sph.emory.edu)

Key Words: relative survival, SES, additive hazard, expected rates

ABSTRACT

Background: Relative survival describes the probability of surviving a diagnosis of cancer in the presence of competing causes of death. In the U.S., this is calculated by controlling for age-, race- and sex-matched background mortality experienced by the entire population in a given calendar period. As mortality is known to also vary by socioeconomic status (SES), this study evaluated the effect of using local, area-based SES-specific background mortality on relative survival estimates.

Methods: Age-, race-, sex- and SES-specific life tables were developed for Metropolitan Atlanta around the 2000 U.S. Census. For the two most common malignancies in men and women (breast cancer and prostate cancer), aggregated and SES stratified age-, race- and cancer-specific 5-year relative survival rates were calculated and compared using 1) traditional national life tables and 2) SES-specific life tables limited to Metropolitan Atlanta. Results using regression models for relative survival were compared to traditional Cox proportional hazards models.

Results: A consistent pattern of decreased survival with increased area-based poverty was observed in this study. The use of national life tables to estimate background mortality generally overestimated relative survival in low poverty areas and underestimated relative survival in high poverty areas. The use of SES-specific life tables somewhat diminished the observed SES disparities in survival. When modeling the data, the excess risk of death among persons living in poorer areas was lower when

appropriately controlling for SES-specific background mortality. Cause-specific regression models underestimated the role of poverty in breast cancer analyses but produced almost identical results for prostate cancer.

Conclusions: Area-based measures of socioeconomic status need to be considered when controlling for background mortality using population-based registry data. This is especially important in SES stratified analyses and regression models that include SES. Further research and discussion is needed to address this important cancer surveillance issue at the national level.

INTRODUCTION

When a person is diagnosed with cancer the overall probability of death can be viewed as an additive effect of two forces of mortality: death from non-cancer causes experienced by individuals from the general population, similar with respect to age, race, sex, calendar year, and other factors (background mortality), and excess death resulting from the addition of the newly diagnosed cancer (excess mortality) (1). From the patient's perspective, surviving the diagnosis becomes the focal point of concern and a primary goal of treatment is to decrease excess mortality. Measures of average survival within population subgroups serve as guides for both clinicians and patients in planning treatment strategies.

Relative survival is the most commonly reported outcome measure used by population-based cancer registries to describe the probability of surviving a diagnosis of cancer in the presence of competing causes of death (2). Defined as the observed survival in a cohort of cancer patients divided by the expected survival in a comparable age-, race-, sex-, and calendar year-matched cohort from the general population, it provides a mechanism to control for non-cancer related mortality (3-7). Mortality has been shown, however, to differ by factors other than age, race, sex, and calendar year. Numerous studies have clearly demonstrated that mortality varies by measures of socioeconomic status (SES) for both cancer and non-cancer causes of death (8-14). A summary of the literature on socioeconomic disparities in total mortality showed a consistent negative gradient with individuals living in areas of lower SES experiencing higher levels of mortality (8). This gradient was observed for both males and females, although it was

more pronounced in males. Special attention should be given to the choice of the population used to calculate background mortality in relative survival analyses and the variables used to match this population to the cancer cohort. If expected survival is underestimated in the comparison population, relative survival in the cancer cohort will be overestimated and vice versa.

Prior studies have examined the need to control for SES-specific background mortality when measuring relative survival. Two recent studies, by Dickman et. al. (15) and Kravdal (16), demonstrated that the use of a single national life table to measure background mortality for all SES groups can bias relative survival estimates, especially when comparing different SES strata. These studies were conducted in countries where national identification numbers allow linking registry data with individual level census measures. In the U.S., census data at an individual level are not available publicly and current National Cancer Institute (NCI) publications presenting relative survival rates do not take SES into consideration (2). Instead the NCI has chosen to use cause-specific survival to control for background mortality in SES-specific analyses using area-based measures of SES (17). Cause-specific survival, like relative survival, uses standard life table methodology (18-20) but considers deaths from causes other than the cancer under investigation as censored observations (21, 22). This approach requires patient-specific cause of death and the validity of results is impacted by the accuracy and completeness of information on the death certificate. Relative survival avoids the reliance on death certificate coded causes which have been shown to have limitations due to various types of coding errors (23-26). As one example, SEER data from 1996-2000 showed that 5%

of all deaths were coded with an unknown cause and another 5% were coded as unspecified cancer (27).

The calculation of SES-specific relative survival in U.S. registries must presently use area-based rather than individual measures. United States registries consistently obtain address at diagnosis for each cancer case and geocode each address to the level of the census tract (28). Census data can be linked to area-based measures of SES to assign neighborhood level characteristics to the underlying records, thus allowing SES-specific analysis (29-32). Little work has been done to evaluate the effect of using area-based SES-specific background mortality in the calculation of relative survival rates within the U.S.

The primary aim of this study is to evaluate the role of SES-specific background mortality in the calculation of aggregated and SES-stratified relative survival rates. Regression modeling will further evaluate the role of SES in cancer survival while controlling for background mortality and other selected factors known to affect survival. Comparisons with cause-specific survival will also be made.

MATERIALS AND METHODS

Source data

The data for this study came from four sources: the Metropolitan Atlanta Surveillance, Epidemiology, and End Results (SEER) Registry (MASR), the Vital Records Department of the Georgia Department of Human Resources (DHR), the National Center for Health

Statistics (NCHS), and the 2000 U.S. Decennial Census. The MASR data provided overall and SES-specific observed survival in the study cohort. Mortality data from the NCHS were used in conjunction with population data from the 2000 U.S. Census to create national life tables. Mortality data from the DHR and population data from the 2000 U.S. Census were used to create local SES-specific life tables.

The MASR is a population-based cancer registry collecting information on all incident cancers in a 5-county area surrounding the city of Atlanta. All primary invasive cancers in individuals between 15 and 84 years of age diagnosed between January 1, 1996 and December 31, 2000, were eligible for inclusion. Cases were excluded if they were out of the appropriate age range (7.6%), alive with no survival time available (0.1%), reported as being from a racial group other than white or black (1.8%), or missing the necessary area-based SES measure (0.3%). In addition, people were excluded if their cancer was a second or subsequent primary cancer (12.5%), or was diagnosed at death or autopsy only (1.2%). Follow-up through 2005 was successful for 98% of all cases.

Mortality data from the NCHS and DHR were restricted to the years 1999-2001 to follow the standard methodology for generating life tables around the decennial census (33). While the NCHS data were available for the entire U.S., data from DHR were limited to the 5 counties included in the MASR.

Choice of area-based SES measure

In the absence of individual SES information, census tract poverty has been shown to serve as a meaningful alternative (34-36). In the U.S., census tracts are small relatively permanent statistical subdivisions of a county that were designed to be homogeneous with respect to socioeconomic status and living conditions (37). For the purpose of this study, addresses from the MASR and DHR were geocoded to the level of the census tract by a single commercial vendor and linked with poverty data from the 2000 U.S. Census Summary File 3 (38). Federal standards define census tracts with twenty percent or more of the population living below the poverty level as "poverty areas" (39). The designation of poverty level varies with family size, income, and year (e.g., $17,463 for a family of 4 in calendar year 2000). Each census tract was classified into one of three groups according to the percentage of the tract population living below the poverty level: 0-9.9% (low poverty), 10-19.9% (middle poverty), and 20-100% (high poverty).

Creation of expected rate tables incorporating a measure of SES

The U.S. population census is conducted every 10 years. Decennial life tables are generated using census population counts and national mortality. Life tables present the probability of dying between the ages of x and x+1 for a person of a given race and sex. Life tables are then used to develop expected rate tables. Expected rate tables present the probability of survival from age x to x+1 for a person of a given race and sex and are used to generate the expected survival (denominator) for relative survival calculations.

For this study, two groups of expected rate tables were created and their impacts on relative survival were compared: standard U.S. national tables (US) that do not consider SES and SES-specific MASR tables (MASR-SES). Separate age-specific tables were created for white males, white females, black males and black females. All tables excluded data for ages less than 15 due to the believed underenumeration of the population by the census at younger ages (33). Data for ages above 89 were also excluded due to known problems with the accuracy of conventional death rates in this population (33). Race groups were based on unbridged race categories from the 2000 Census and were limited to white and black. Race bridging refers to the process of assigning an individual of multiple races to a single race category (40). Unbridged data were used because race bridged population information was not available at the level of the census tract. Research has demonstrated that race bridging has a small-to-negligible impact on white and black populations (40).

For the MASR-SES tables, geocoded mortality data were linked to census files by tract to assign the area-based poverty measure to each record. Records were divided into the three poverty categories previously defined (0-9.9%, 10-19.9%, and 20-100%) and counts were obtained by age, race, and sex. All census tracts in the MASR were then divided into the same three poverty categories and population counts by age, race, and sex were obtained from the census file. For each poverty stratum, expected rate tables were created by age, race, and sex using SES-specific mortality rates. A brief summary of the methodology used to generate the expected rate tables is presented in Appendix 5.A.

Analysis

Primary analyses in this study were restricted to the two leading incident cancers in the MASR population: female breast cancer and prostate cancer. Analyses of baseline characteristics across SES categories were conducted in SAS version 9.1 (Cary, NC). Relative survival calculations were performed using SEER*Stat software (41). Five-year relative survival rates by age, race, and SES were generated using both the standard U.S. national expected rate tables and the MASR SES-specific expected rate tables. Results using each method were compared.

There are other methods to estimate relative survival, in addition to the stratified method described above. One of these is the use of a generalized linear regression model with a Poisson error structure, which allows simultaneous control of multiple factors in a single analysis (1). This additive hazards model allows the incorporation of the MASR SES-specific expected rate tables to model the baseline hazard of background mortality in addition to the excess hazard caused by cancer. The general formula for this model is:

$$h(t,x) = h^{*}(t,x_1) + \exp(x\beta)$$

where h(t,x) represents the overall hazard, t represents the time since diagnosis, x represents the covariate vector for each cancer patient, $h^{*}(t,x_1)$ represents the background hazard estimated in the general population with covariate vector x_1, and $\exp(x\beta)$ represents the excess hazard due to cancer. Since stage is a strong predictor of survival and because poverty could affect the timing of diagnosis, analyses were stratified by early

and late stage as defined by the national SEER Program (42). Analyses for prostate cancer were limited to regional or distant stage disease since 5-year relative survival for localized disease was virtually 100% (2).

Since cause-specific survival is an analogous methodology used to control for background mortality in many studies, results from the above model using MASR SES-specific expected rate tables were also compared against various Cox regression models. Comparisons were restricted to the area-based poverty variable since this was our primary interest. As discussed previously, cause-specific analyses rely on the use of coded cause of death to classify individuals into the appropriate outcome category and this is considered the key limitation of Cox regression using population-based registry data. Three Cox models were compared against the results from the relative survival generalized linear model. The first was a truly cause-specific analysis where only the cancer of interest was considered as the cause of death; the second model included other cancer deaths; and the final model included all causes of death.

RESULTS

Baseline characteristics of the MASR invasive female breast cancer and prostate cancer cases are presented in Table 5.1 by area-based SES. While overall there were more breast cancer cases during the study period, prostate cancers were more common in high poverty areas. Study subjects in these high poverty areas were more likely to be black, older, deceased within 5 years following their diagnosis of cancer, and Fulton County residents. Fulton County is where the state capital city, Atlanta, Georgia, is located. The

cancers among high poverty area residents were characterized as later stage and higher grade disease than those of the lower poverty area residents.

An evaluation of age-, sex-, and race-specific mortality rates for the MASR counties stratified by poverty levels showed that national rates typically overestimate local background mortality rates in low poverty areas of the MASR and underestimate them in the high poverty areas. Differences are consistently larger in males relative to females. The largest absolute differences between the national rates and those for age-, sex-, race- and poverty-specific strata will cause the greatest change in relative survival rates when controlling for background mortality by SES (see Appendix 5.B).

Regardless of the approach used to calculate relative survival, a fairly consistent pattern of decreasing survival with increasing area-based poverty was observed (Table 5.2). Overall differences between relative survival rates computed using national and MASR SES-specific expected survival were generally small. For analyses with all poverty groups combined, age- and race-specific relative survival rates never differed by more than 1 percent between the two methods. Somewhat larger effects were observed when comparing differences in relative survival stratified by SES. In general, relative survival rates calculated using national expected survival overestimated survival in low poverty areas and underestimated survival in high poverty areas, relative to rates calculated using MASR-SES expected survival. Differences were generally more pronounced for prostate cancer and in the areas of high poverty. Figure 5.1 plots the relative survival rates for prostate cancer for both whites and blacks comparing the two methods. At 5 years,

relative survival for white males using national expected survival was 84.9 percent. When MASR-SES expected survival was used instead, 5-year relative survival of 92.7 percent was observed. These differences increased over follow-up time and were also observed in black males, although to a lesser degree. The same general pattern was seen in breast cancer (Figure 5.2) but with smaller differences observed.

Table 5.3 presents the results of stage-stratified regression models. In all categories, as poverty levels increased so did the excess mortality. For cancer patients residing in high poverty areas, the excess risk of death from their cancer diagnosis was consistently lower when modeling background mortality using MASR SES-specific expected survival. As an example, using national expected survival, excess mortality among prostate cancer patients with late stage disease residing in the highest poverty areas was 2.56 times higher (95% CI 1.49, 4.40) than that among similar individuals in the lowest poverty areas. In the same model using MASR SES-specific expected survival, the relative excess mortality was 2.22 (95% CI 1.29, 3.80).

Table 5.4 compares results from the additive hazards model presented in Table 5.3 to the traditional proportional hazards model used in Cox regression models. For breast cancer, cause-specific models underestimated the role of area-based poverty in the risk of death from breast cancer and did so to a greater extent for higher poverty areas. Models incorporating other cancers and other causes provided results most consistent with those of the additive hazards model. For prostate cancer, cause-specific models produced results almost identical to the additive hazards model.

DISCUSSION

Our study demonstrated decreased cancer survival and increased all cause mortality among individuals living in lower SES areas. These observations are in agreement with previous reports. A recent analysis of SEER data showed that, after controlling for stage, individuals living in high poverty census tracts had worse survival from each of the cancers investigated relative to those living in lower poverty tracts (17). There is also consistent evidence that overall mortality and each of its components, cancer and non-cancer mortality, vary by SES. For these reasons, calculations of relative survival that do not control for SES may be subject to error.

Take for example a white male, age 66, living in the highest poverty area. Using national expected rate tables that currently ignore SES would underestimate his background mortality of this individual (Appendix 5.B: 2,299 per 100,000 compared to 3,744 per 100,000) and thus overestimate his expected survival; consequently his relative survival would be underestimated. In comparison, the relative survival of an otherwise similar man living in one of the lowest poverty areas would be overestimated due to the overestimation of his background mortality (Appendix 5.B: 2,299 per 100,000 compared to 2,009 per 100,000). The net effect of these types of errors would lead, in most circumstances, to an apparently widened survival gap between the poverty strata than the use of SES-specific estimates would indicate.

A review of SES-specific all cause mortality compared to a national reference illustrates the degree to which the use of national expected survival may affect study results. Since the numerator of the relative survival rate, observed survival, remains constant, all of the variation in these calculations is driven by the denominator. Age, sex, race, and poverty level-specific strata with the largest absolute differences in all cause mortality from the equivalent age-, sex-, and race-specific national rates will yield the most biased relative survival rates when background mortality by SES is ignored. In our population, the magnitude of error is largest in the low poverty strata for blacks and the high poverty strata for whites. It appears to be more pronounced in men relative to women.

Two other important findings arise from the results of this study. In this population, when a SES-stratified analysis is not being conducted, there is minimal benefit gained to using SES-specific expected survival as the errors cancel each other out. This study also highlights the need for caution in the use of cancer registry data for calculating cause-specific survival. Cause-specific survival relies heavily on the quality of coding of cause of death and some evidence indicated that this may vary by SES (43). In our study, the cause-specific results when modeling prostate cancer were similar to those from the relative survival models, whereas the cause-specific results when modeling breast cancer underestimated the relative survival results. In women diagnosed with localized breast cancer, one possible explanation for the observed underestimation is misclassification of the cause of death. It appears that in high poverty areas, breast cancer mortality is most comparable to the Cox model that includes breast cancer deaths along with deaths from common metastatic sites, perhaps because metastatic sites may be erroneously listed as

the underlying cause of death. Among women with late stage disease, however, breast cancer mortality is closest to all-cause mortality presumably because cancer contributed to mortality from other causes (e.g. via treatment related toxicity) that were not captured in the cancer-specific results.

Our study is conceptually similar to the earlier publication by Dickman et al., which was based on Finnish data (15). Although the findings and conclusions from their research are comparable to ours, their analysis used individual measures of social class that are not available in population-based registries in the U.S. It is also important to point out that SES-specific mortality is different in different populations, and for this reason, data from Finland may not apply to the United States as the data from Georgia may not apply to Utah.

The strengths of our study include greater than 98% follow-up and a high proportion (98.5%) of registry data geocoded to a street level address. For the purpose of this study we developed a multi-step procedure to clean and improve address data that could not be successfully geocoded through standard registry practices. Consistent methods were used to create both the national and MASR SES-specific expected rate tables. The limitations that warrant consideration include the assumption that SES is homogeneous within a census tract and the exclusion of certain age groups at both extremes when developing expected rate tables for this study. Additional methods to develop SES-specific expected rate tables to include these other age groups are needed. It would also be beneficial to re-examine the results of this study in other cancers, in other populations, such as those with

a larger percentage of residents living in high poverty areas, and for longer periods of follow-up. A comparison of results using local SES-specific expected rate tables to those using U.S. national SES-specific tables could lead the discussion regarding optimal methods for widespread use of SES-specific expected rates in relative survival analysis.

CONCLUSION

Area-based measures of socioeconomic status need to be considered when controlling for background mortality using population-based registry data. This is especially important in SES stratified analyses. The development of SES-specific expected rate tables at the local level, however, requires substantial effort and time. Further research and discussion is needed to address this important cancer surveillance issue at the national level.

TABLE 5.1 Baseline characteristics of the MASR invasive female breast and prostate cancer cases* according to census tract poverty measure, 1996-2000

Characteristic		Low Poverty (0-9.9%) N=8,835		Middle Poverty (10-19.9%) N=2,792		High Poverty (20+%) N=1,602	
		No.	%	No.	%	No.	%
Sex							
	Male	4,255	48.2	1,344	48.1	887	55.4
	Female	4,580	51.8	1,448	51.9	715	44.6
	p value ~				0.999		<0.001
Race							
	White	7,247	82.0	1651	59.1	271	16.9
	Black	1,588	18.0	1,141	40.9	1,331	83.1
	p value ~				<0.001		<0.001
Age							
	15-54	2,948	33.4	851	30.5	379	23.7
	55-69	3,768	42.6	1,137	40.7	712	44.4
	70-84	2,119	24.0	804	28.8	511	31.9
	p value ~				<0.001		<0.001
Stage							
	Local	6,360	72.0	1,900	68.1	1,051	65.6
	Regional/Distant	2,132	24.1	774	27.7	475	29.7
	Unknown	343	3.9	118	4.2	76	4.7
	p value ~				<0.001		<0.001
Grade							
	Well\Mod Diff	5,782	65.5	1,681	60.2	930	58.0
	Poor\Undiff	2,210	25.0	803	28.8	461	28.8
	Unknown	843	9.5	308	11.0	211	13.2
	p value ~				<0.001		<0.001
Vital Status at study cutoff							
	Alive	7,070	80.0	1,950	69.8	949	59.2
	Deceased	1,765	20.0	842	30.2	653	40.8
	p value ~				<0.001		<0.001
Geographic Area							
	Clayton	416	4.7	409	14.6	29	1.8
	Cobb	2,326	26.3	407	14.6	53	3.3
	DeKalb	2,033	23.0	999	35.8	236	14.7
	Fulton	2,183	24.7	734	26.3	1,284	80.2
	Gwinnett	1,877	21.3	243	8.7	0	0.0
	p value ~				<0.001		<0.001
Diagnosis year							
	1996	1,554	17.6	584	20.9	312	19.5
	1997	1,846	20.9	548	19.6	330	20.6
	1998	1,719	19.5	545	19.5	334	20.8
	1999	1,829	20.7	540	19.4	338	21.1
	2000	1,887	21.3	575	20.6	288	18.0
	p value ~				0.002		0.020

~ p value based on chi-square test (compared to low poverty group)
*Study population limited to cases eligible for survival analysis

TABLE 5.2 Comparison of age- and race-specific 5-year relative survival rates for two leading incident cancers in GA by percent of the census tract population living below the federal poverty level using expected survival from national life tables versus MASR-SES life tables, 1996 to 2000

				National Expected Survival							
				Low Poverty (0-9.9%)		Middle Poverty (10-19.9%)		High Poverty (20+%)		ALL	
Cancer	Sex	Race	Age*	No.	5-yr RS	No.	5-yr RS	No.	5-yr RS	No.	5-yr RS
Breast	Female	White	All	3,841	92.4	883	88.6	142	80.4	4,866	91.4
			15-54	1,770	91.6	329	88.7	33	76.8	2,132	90.9
			55-69	1,288	93.3	288	88.5	54	82.3	1,630	92.1
			70-84	783	93.0	266	88.5	55	76.7	1,104	91.2
		Black	All	739	83.6	565	75.2	573	73.3	1877	78.1
			15-54	504	83.4	350	73.4	262	68.0	1,116	76.7
			55-69	167	82.3	142	80.5	183	76.2	492	79.7
			70-84	68	88.4	73	69.7	128	81.9	269	81.9
Prostate	Male	White	All	3,406	100.0	768	94.7	129	84.9	4,303	100.0
			15-54	446	98.1	86	95.6	14	73.6	546	97.5
			55-69	1,839	99.9	375	96.1	56	87.1	2,270	99.8
			70-84	1,121	100.0	307	92.3	59	83.9	1,487	100.0
		Black	All	849	99.9	576	95.9	758	90.8	2,183	96.8
			15-54	228	99.4	86	92.3	70	92.2	384	97.2
			55-69	474	100.0	332	98.4	419	90.5	1,225	97.7
			70-84	147	97.4	158	89.7	269	90.4	574	93.3

* Restricted to ages 15-84

Relative survival rates over 1.0 have been adjusted to 1.0.

TABLE 5.2 (continued) Comparison of age- and race-specific 5-year relative survival rates for two leading incident cancers in GA by percent of the census tract population living below the federal poverty level using expected survival from national life tables versus MASR-SES life tables, 1996-2000

				MASR SES-specific Expected Survival							
				Low Poverty (0-9.9%)		Middle Poverty (10-19.9%)		High Poverty (20+%)		ALL	
Cancer	Sex	Race	Age*	No.	5-yr RS	No.	5-yr RS	No.	5-yr RS	No.	5-yr RS
Breast	Female	White	All	3,841	92.1	883	89.2	142	82.9	4,866	91.3
			15-54	1,770	91.2	329	88.9	33	77.2	2,132	90.7
			55-69	1,288	92.5	288	89.4	54	83.8	1,630	91.8
			70-84	783	93.6	266	89.0	55	79.6	1,104	92.0
		Black	All	739	82.8	565	74.9	573	74.6	1,877	78.0
			15-54	504	82.6	350	73.1	262	68.6	1,116	76.4
			55-69	167	80.9	142	80.2	183	77.7	492	79.7
			70-84	68	89.7	73	69.7	128	82.5	269	82.5
Prostate	Male	White	All	3,406	100.0	768	96.9	129	92.7	4,303	99.8
			15-54	446	97.5	86	96.4	14	76.3	546	96.9
			55-69	1,839	99.8	375	97.8	56	92.2	2,270	99.6
			70-84	1,121	100.0	307	95.5	59	93.6	1,487	100.0
		Black	All	849	99.8	576	95.4	758	94.2	2,183	97.0
			15-54	228	97.9	86	91.9	70	94.5	384	96.3
			55-69	474	100.0	332	98.1	419	94.5	1,225	98.0
			70-84	147	97.5	158	88.5	269	92.9	574	94.2

* Restricted to ages 15-84

Relative survival rates over 1.0 have been adjusted to 1.0.

FIGURE 5.1 Comparison of expected survival effects in annual prostate cancer relative survival rates by race, (White N=129 and Black N=758), 1996-2000 cohort followed through 2005, MASR, high poverty category (20+%)

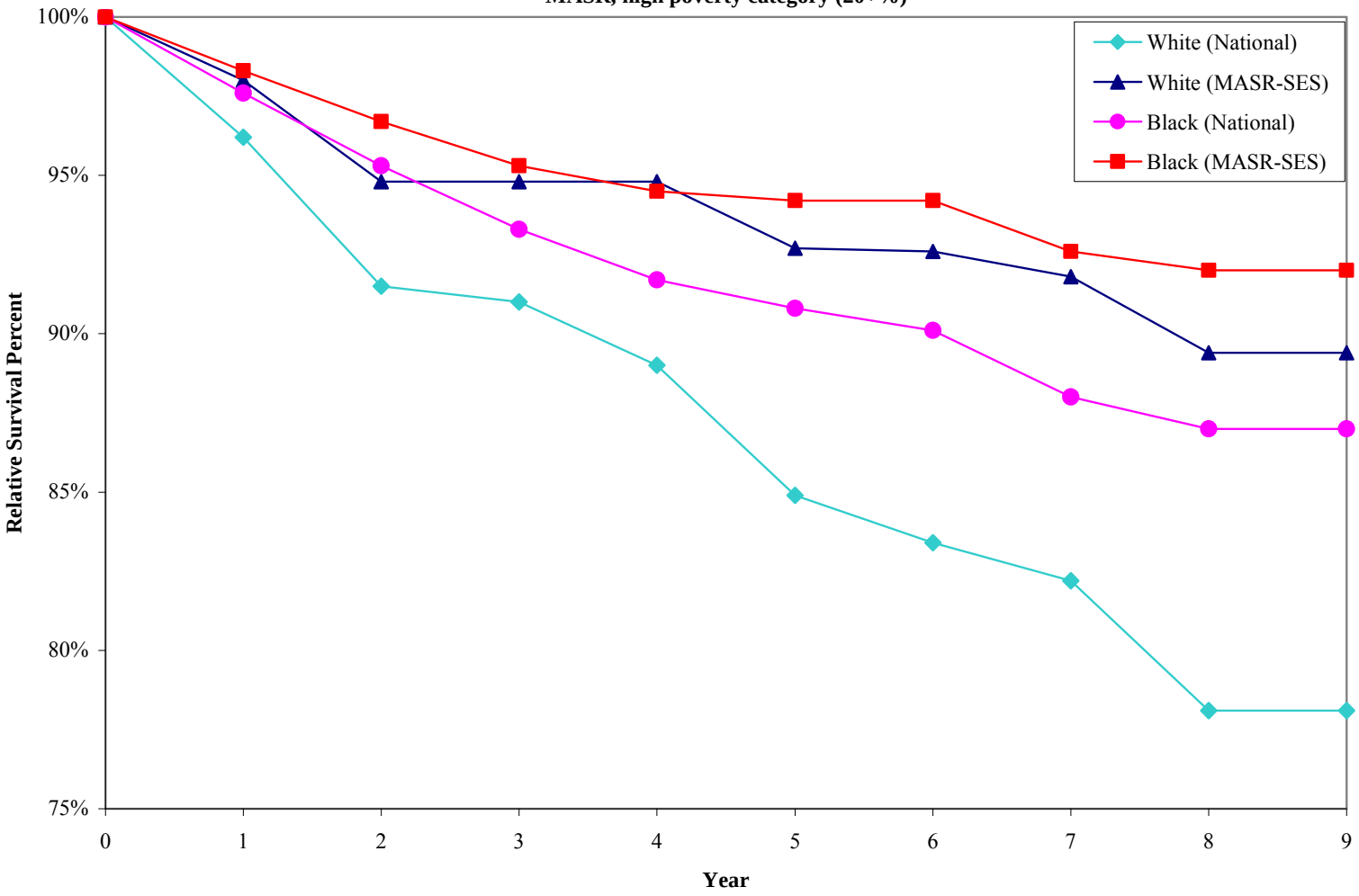

FIGURE 5.2 Comparison of expected survival effects in annual female breast cancer relative survival rates by race (White N=142 and Black N=573), 1996-2000 cohort followed through 2005, MASR, high poverty category (20+%)

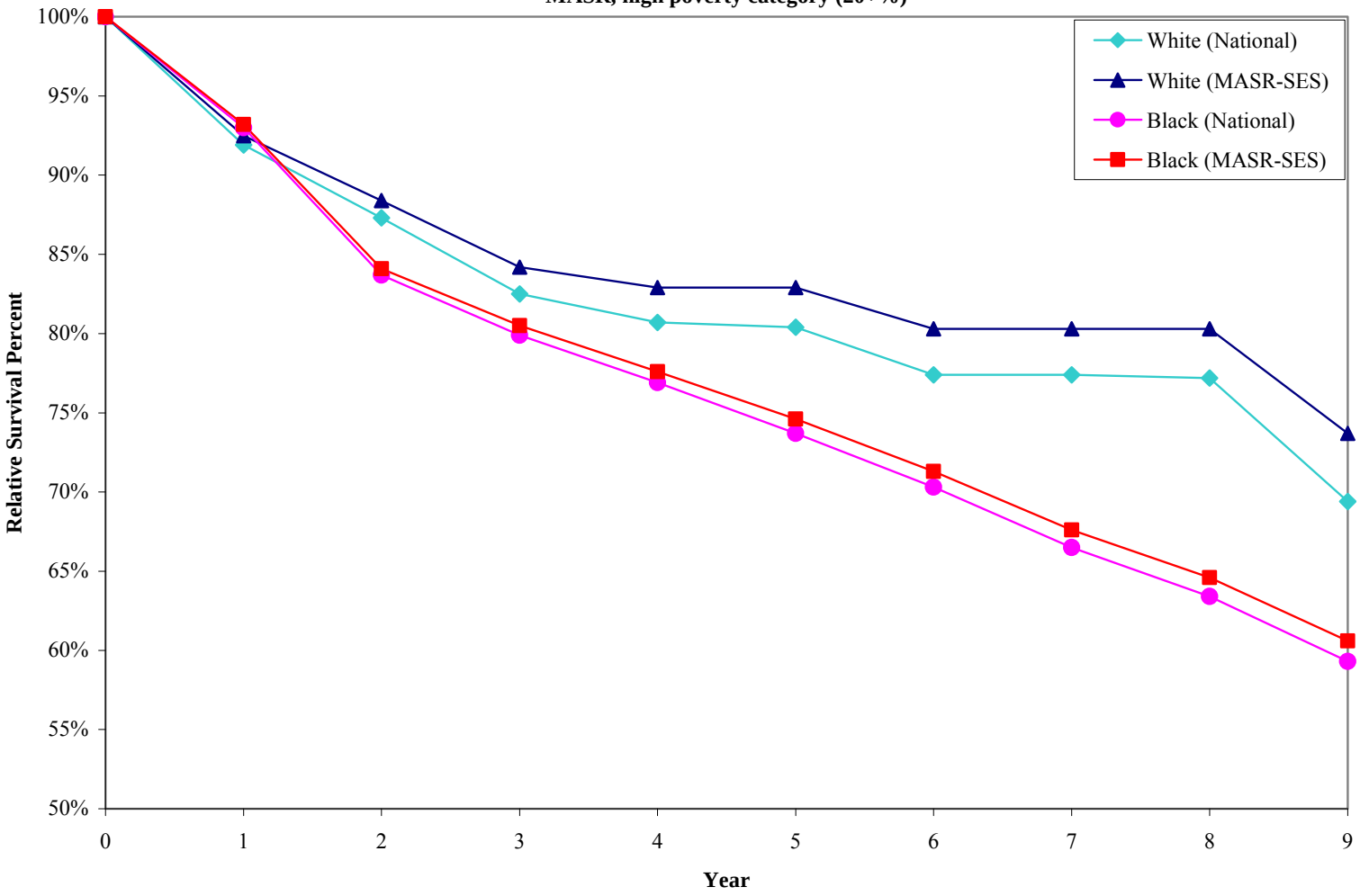

Table 5.3 Regression models for relative survival comparing the use of national expected survival versus MASR SES-specific survival

	National Expected Survival		MASR SES-Specific Expected Survival	
Parameter	Relative Excess Risk	95% CI	Relative Excess Risk	95% CI
Breast Cancer Local Stage (N=4,139)				
Race				
Black	1.55	0.95, 2.52	1.62	1.02, 2.57
White	1.00	Referent	1.00	Referent
Age				
70-84	0.64	0.20, 2.07	0.53	0.13, 2.07
55-69	0.54	0.27, 1.09	0.53	0.27, 1.05
15-54	1.00	Referent	1.00	Referent
Grade				
Unknown	3.77	1.48, 9.69	3.45	1.49, 8.02
Poorly Diff \ Undiff	6.37	2.80, 14.48	5.77	2.83, 11.78
Well \ Mod Diff	1.00	Referent	1.00	Referent
Pct Below Poverty				
20+	1.83	0.94, 3.59	1.45	0.72, 2.89
10-19.9	1.31	0.76, 2.26	1.16	0.68, 1.97
0-9.9	1.00	Referent	1.00	Referent
Breast Cancer Reg \ Distant Stage (N=2,517)				
Race				
Black	1.64	1.36, 1.98	1.68	1.39, 2.02
White	1.00	Referent	1.00	Referent
Age				
70-84	not displayed due to		not displayed due to	
55-69	interaction with follow-up time		interaction with follow-up time	
15-54				
Grade				
Unknown	4.26	3.29, 5.52	4.21	3.26, 5.45
Poorly Diff \ Undiff	2.66	2.12, 3.34	2.63	2.10, 3.30
Well \ Mod Diff	1.00	Referent	1.00	Referent
Pct Below Poverty				
20+	1.79	1.40, 2.28	1.69	1.32, 2.15
10-19.9	1.43	1.17, 1.74	1.38	1.13, 1.68
0-9.9	1.00	Referent	1.00	Referent

Table 5.3 (continued). Regression models for relative survival comparing the use of national expected survival versus MASR SES-specific survival

	National Expected Survival		MASR SES-Specific Expected Survival	
Parameter	Relative Excess Risk	95% CI	Relative Excess Risk	95% CI
Prostate Cancer Reg \ Distant Stage (N=864)				
Race				
Black	1.11	0.69, 1.77	1.17	0.73, 1.86
White	1.00	Referent	1.00	Referent
Age				
70-84	1.80	1.25, 2.66	1.77	1.03, 3.02
55-69	0.68	0.40, 1.17	0.67	0.40, 1.14
15-54	1.00	Referent	1.00	Referent
Grade				
Unknown	31.43	8.38, 117.89	30.28	8.70, 105.43
Poorly Diff \ Undiff	14.26	3.96, 51.40	13.61	4.07, 45.49
Well \ Mod Diff	1.00	Referent	1.00	Referent
Pct Below Poverty				
20+	2.56	1.49, 4.40	2.22	1.29, 3.80
10-19.9	1.42	0.87, 2.33	1.31	0.80, 2.13
0-9.9	1.00	Referent	1.00	Referent

Table 5.4 Comparison of relative excess risk of death by census tract poverty level using different regression techniques

	MASR SES-Specific Expected Surv		Cox Regression (Specific Cancer)		Cox Regression (Cancer)		Cox Regression (All Cause)	
Parameter	Relative Excess Risk	95% CI	Hazard Ratio	95% CI	Hazard Ratio	95% CI	Hazard Ratio	95% CI
Breast Cancer Local Stage (N=4,139)								
Pct Below Poverty								
20+	1.45	0.72, 2.89	1.16	0.70, 1.92	1.43	0.92, 2.23	1.34	0.98, 1.84
10-19.9	1.16	0.68, 1.97	1.11	0.77, 1.61	1.19	0.85, 1.66	1.27	1.01, 1.59
0-9.9	1.00	Referent	1.00	Referent	1.00	Referent	1.00	Referent
Breast Cancer Reg \ Distant Stage (N=2,517)								
Pct Below Poverty								
20+	1.69	1.32, 2.15	1.50	1.19, 1.91	1.56	1.24, 1.95	1.65	1.34, 2.03
10-19.9	1.38	1.13, 1.68	1.32	1.09, 1.59	1.31	1.09, 1.57	1.34	1.13, 1.58
0-9.9	1.00	Referent	1.00	Referent	1.00	Referent	1.00	Referent
Prostate Cancer Reg \ Distant Stage (N=864)								
Pct Below Poverty								
20+	2.22	1.29, 3.80	2.24	1.46, 3.44	2.08	1.39, 3.12	1.93	1.35, 2.74
10-19.9	1.31	0.80, 2.13	1.29	0.86, 1.93	1.23	0.84, 1.80	1.24	0.90, 1.71
0-9.9	1.00	Referent	1.00	Referent	1.00	Referent	1.00	Referent

Appendix 5.A

The core component of the life table is the death rate q_x. Defined as the proportion of the population at age x that is expected to die before attaining age x+1, it is generally calculated using the formula

$$q_x = D_x / (3P_x + D_x/2)$$

where D_x is the adjusted number of deaths occurring in a population age x over the three year period surrounding the population census, P_x is the census population age x at the mid-year of the census period and deaths are assumed to occur uniformly over the 1 year period during which the age advances from x to x+1 (33). Deaths for which age was not stated are allocated proportionately among the different age groups through the use of an adjustment factor. Beers interpolation coefficients are used to smooth single age population death rates (44). This technique is applied to both death (D_x) and population (P_x) values by aggregating the single year data into 5-year age groups and then interpolating back to single age values. The compliment of the life table death rate ($1- q_x$) represents the proportion of the population surviving the age interval x to x+1 and is the value used in the expected rate table to calculate expected survival. Life table death rates are calculated at individual years of age from 15 to 89 for all 12 combinations of race (white/black), sex (male/female) and census tract poverty measure (low, medium, high poverty). Similar methodology is used to create the U.S national tables except without stratification by SES.

Appendix 5.B Age, sex and race-specific mortality rates* for combined MASR counties by % of the census tract population living below the federal poverty level, 1999-2001, with national mortality as a reference

Males

	White (N=12,734)				Black (N=7,969)			
	Poverty Level				Poverty Level			
Age	Low (0-9.9%)	Middle (10-19.9%)	High (20+%)	National	Low (0-9.9%)	Middle (10-19.9%)	High (20+%)	National
15-19 years	95.8	127.0	90.6	88.1	103.6	115.5	231.6	132.5
20-24 years	145.4	142.4	106.7	123.3	193.3	212.6	222.7	237.7
25-29 years	114.3	135.9	203.1	116.8	142.4	146.5	338.4	244.0
30-34 years	113.5	142.0	243.2	135.8	179.8	215.4	462.5	275.1
35-39 years	126.6	230.0	412.6	191.2	236.1	354.4	779.3	366.0
40-44 years	210.7	445.0	512.1	277.7	293.1	557.9	1050.8	548.5
45-49 years	316.9	596.0	860.1	411.5	530.7	751.2	1452.5	855.2
50-54 years	415.1	755.1	1456.4	583.2	731.1	1229.3	1760.7	1206.0
55-59 years	681.0	1177.8	1827.0	922.1	1390.1	1556.2	2521.5	1779.3
60-64 years	1218.2	1909.3	2917.3	1457.1	1981.3	2581.8	3178.7	2475.6
65-69 years	2009.3	2811.1	3744.0	2299.3	3094.2	3619.8	4790.2	3528.5
70-74 years	3354.9	4223.5	6138.7	3600.5	4765.7	5112.8	5841.2	5186.6
75-79 years	5218.0	6468.8	7617.1	5619.6	7555.0	7664.9	8451.4	7362.4
80-84 years	9193.0	10295.6	12600.2	8987.6	12918.7	9126.1	10952.4	10331.4

*Rates are per 100,000

Georgia mortality provided by Vital Records Department of the GA Department of Human Resources

National mortality data provided by NCHS (www.cdc.gov/nchs).

Differs from national rate by greater than 500/100,000

Appendix 5.B (continued). Age, sex and race-specific mortality rates* for combined MASR counties by % of the census tract pop. living below the fed. poverty level, 1999-2001, with national mortality as a reference

Females

	White (N=10,676)				Black (N=6,426)			
	Poverty Level				Poverty Level			
Age	Low (0-9.9%)	Middle (10-19.9%)	High (20+%)	National	Low (0-9.9%)	Middle (10-19.9%)	High (20+%)	National
15-19 years	43.1	59.0	12.6	39.5	55.0	42.9	37.8	43.2
20-24 years	48.2	39.8	35.2	42.7	72.8	62.6	74.1	71.9
25-29 years	45.6	42.0	24.3	47.7	68.3	69.7	117.1	97.1
30-34 years	49.8	92.5	128.0	64.8	78.8	99.2	224.3	140.5
35-39 years	73.2	138.2	142.3	101.3	92.5	180.3	420.2	217.8
40-44 years	99.0	192.3	334.4	151.6	189.0	305.3	463.6	337.0
45-49 years	164.1	298.6	340.1	223.3	276.1	368.4	719.5	489.4
50-54 years	258.4	427.7	594.2	345.0	430.9	588.1	903.1	685.9
55-59 years	464.0	737.6	737.3	568.8	727.0	1024.0	1291.3	1022.8
60-64 years	778.9	1168.5	1759.1	918.4	1148.9	1376.8	1928.2	1496.7
65-69 years	1298.5	1754.7	2248.7	1458.4	1985.2	2240.0	2749.2	2206.2
70-74 years	2259.2	2594.6	3406.9	2287.0	3016.4	3132.8	3567.9	3241.3
75-79 years	3791.5	4130.6	5006.4	3692.2	5548.3	4791.4	5062.5	4779.7
80-84 years	6750.1	6302.5	7128.2	6233.5	7871.3	7503.3	7273.3	6984.6

*Rates are per 100,000

Georgia mortality provided by Vital Records Department of the GA Department of Human Resources

National mortality data provided by NCHS (www.cdc.gov/nchs).

Differs from national rate by greater than 500/100,000

REFERENCES

1. Dickman PW, Sloggett A, Hills M, et al. Regression models for relative survival. Stat Med 2004;23:51-64.
2. Ries LAG, Melbert D, Krapcho M, et al. SEER Cancer Statistics Review, 1975-2004. National Cancer Institute Bethesda, MD, http://seercancergov/csr/1975_2004/, based on November 2006 SEER data submission, posted to the SEER web site, 2007.
3. Brown CC. The statistical comparison of relative survival rates. Biometrics 1983;39:941-8.
4. Ederer F, Axtell LM, Cutler SJ. The relative survival rate: a statistical methodology. National Cancer Institute Monograph 1961;6:101-21.
5. Esteve J, Benhamou E, Croasdale M, et al. Relative survival and the estimation of net survival: elements for further discussion. Stat Med 1990;9:529-38.
6. Henson DE, Ries LA. The relative survival rate. Cancer 1995;76:1687-8.
7. Parkin DM, Hakulinen T. Analysis of Survival. In: Jensen OM, Parkin DM, MacLennan R, et al., eds. Cancer Registration Principles and Methods. Lyon: International Agency for Research on Cancer: IARC Scientific Publications, 1991:159-76.
8. Faggiano F, Partanen T, Kogevinas M, et al. Socioeconomic differences in cancer incidence and mortality. In: Kogevinas M, Pearce N, Susser M, et al., eds. Social Inequalities and Cancer: IARC Scientific Publications, 1997:65-176.

9. Geronimus AT. Poverty, time and place: variation in excess mortality across selected US populations, 1980-1990. Journal of Epidemiology and Community Health 1999;53:325-34.

10. Geronimus AT, Colen CG, Shochet T, et al. Urban-rural differences in excess mortality among high-poverty populations: evidence from the Harlem Household Survey and the Pitt County, North Carolina Study of African American Health. J Health Care Poor Underserved 2006;17:532-58.

11. Muller A. Association between income inequality and mortality among US States: considering population at risk.[comment]. Am J Public Health 2006;96:590-1.

12. Singh GK, Hiatt RA. Trends and disparities in socioeconomic and behavioral characteristics, life expectancy, and cause-specific mortality of native-born and foreign-born populations in the United States, 1979-2003.[see comment]. Int J Epidemiol 2006;35:903-19.

13. Singh GK, Miller BA, Hankey BF, et al. Persistent area socioeconomic disparities in U.S. incidence of cervical cancer, mortality, stage, and survival, 1975-2000. Cancer 2004;101:1051-7.

14. Vinnakota S, Lam NSN. Socioeconomic inequality of cancer mortality in the United States: a spatial data mining approach. Int J Health Geographics 2006;5:9.

15. Dickman PW, Auvinen A, Voutilainen ET, et al. Measuring social class differences in cancer patient survival: is it necessary to control for social class differences in general population mortality? A Finnish population-based study. J Epidemiol Community Health 1998;52:727-34.

16. Kravdal O. A cancer survival model that takes sociodemographic variations in "normal" mortality into account: comparison with other models. J Epidemiol Community Health 2002;56:309-18.

17. Singh GK, Miller BA, Hankey BF, et al. Area Socioeconomic Variations in US Cancer Incidence, Mortality, Stage, Treatment and Survival, 1975-1999. Bethesda, MD., National Cancer Institute, NCI Cancer Monograph Series, Number 4, NIH Pub No. 03-5417, 2003.

18. Berkson I, Gage RP. Calculation of Survival Rates for Cancer. Proceedings Staff Meet Mayo Clinic 1950;25:270-86.

19. Cutler SJ, Ederer F. Maximum utilization of the life table method in analyzing survival. Journal of Chronic Diseases 1958;8:699-712.

20. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 1958;53:457-81.

21. Bull K, Spiegelhalter DJ. Survival analysis in observational studies. Statistics in Medicine 1997;16:1041-74.

22. Cox DR. Regression models and life tables (with discussion). Journal of the Royal Statistical Society 1972;Series B:187-220.

23. Begg CB, Schrag D. Attribution of deaths following cancer treatment.[comment]. J Natl Cancer Inst 2002;94:1044-5.

24. Engel LW, Strauchen JA, Chiazze L, Jr., et al. Accuracy of death certification in an autopsied population with specific attention to malignant neoplasms and vascular diseases. Am J Epidemiol 1980;111:99-112.

25. Percy C, Stanek E, 3rd, Gloeckler L. Accuracy of cancer death certificates and its effect on cancer mortality statistics. Am J Public Health 1981;71:242-50.

26. Percy CL, Miller BA, Gloeckler Ries LA. Effect of changes in cancer classification and the accuracy of cancer death certificates on trends in cancer mortality. Ann N Y Acad Sci 1990;609:87-97; discussion -9.

27. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER 17 Regs Limited-Use, Nov 2006 Sub (1973-2004 varying) - Linked To County Attributes - Total U.S., 1969-2004 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2007, based on the November 2006 submission.

28. Hofferkamp J, Havener L, editors. Standards for Cancer Registries Volume II: Data Standards and Data Dictionary, Twelfth Edition, Version 11.2. Springfield, IL: North American Association of Central Cancer Registries, April 2007.

29. Krieger N, Chen JT, Waterman PD, et al. Race/ethnicity, gender, and monitoring socioeconomic gradients in health: a comparison of area-based socioeconomic measures--the public health disparities geocoding project. Am J Public Health 2003;93:1655-71.

30. Krieger N, Chen JT, Waterman PD, et al. Painting a truer picture of US socioeconomic and racial/ethnic health inequalities: the Public Health Disparities Geocoding Project. Am J Public Health 2005;95:312-23.

31. Krieger N, Chen JT, Waterman PD, et al. Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of

area-based measure and geographic level matter?: the Public Health Disparities Geocoding Project. Am J Epidemiol 2002;156:471-82.

32. Subramanian SV, Chen JT, Rehkopf DH, et al. Racial disparities in context: a multilevel analysis of neighborhood variations in poverty and excess mortality among black populations in Massachusetts.[erratum appears in Am J Public Health. 2005 Mar;95(3):375]. Am J Public Health 2005;95:260-5.

33. Anderson R. Method for constructing complete annual U.S. life tables. National Center for Health Statistics. Vital Health Stat 2(129). 1999.

34. Geronimus AT. Use of Census-based Aggregate Variables to Proxy for Socioeconomic Group: Evidence from National Samples. Am J Epidemiol 1998;148:475-86.

35. Krieger N. Overcoming the absence of socioeconomic data in medical records: validation and application of a census-based methodology. Am J Public Health 1992;82:703-10.

36. Subramanian SV, Chen JT, Rehkopf DH, et al. Comparing individual- and area-based socioeconomic measures for the surveillance of health disparities: A multilevel analysis of Massachusetts births, 1989-1991.[see comment]. Am J Epidemiol 2006;164:823-34.

37. U.S. Bureau of the Census. Geographic Areas Reference Manual, U.S. Bureau of the Census, Washington, DC (2004) Available at: http://www.census.gov/geo/www/garm.html. Accessed January, 2007.

38. U.S. Bureau of the Census. 2000 Census Population and Housing, Summary File 3: Technical Documentation. 2002.

39. U.S. Bureau of the Census. Poverty Areas. (http://www.census.gov/hhes/www/poverty/definitions.html). (last accessed December 2006).

40. Ingram D, Parker J, Schenker N, et al. United States Census 2000 population with bridged race categories. National Center for Health Statistics. Vital Health Stat 2(135). 2003.

41. Surveillance Research Program, National Cancer Institute SEER*Stat software (www.seer.cancer.gov/seerstat) version 6.3.6.

42. Johnson C, Adamo M, (eds.). SEER Program Coding and Staging Manual 2007. National Cancer Institute, NIH Publication number 07-5581, Bethesda, MD 2007.

43. Samphier ML, Robertson C, Bloor MJ. A possible artefactual component in specific cause mortality gradients. Social class variations in the clinical accuracy of death certificates. J Epidemiol Community Health 1988;42:138-43.

44. Shryock H, Siegel J. The methods and materials of demography, vol 2, U.S. Bureau of the Census. Washington, D.C.: U.S. Government Printing Office.

Chapter 6

Examining the Role of Area-Based Poverty in Survival from Non-Localized Prostate Cancer in the Medicare Population

Kevin C. Ward[1,2], Michael Goodman[1,2], Joseph Lipscomb[3], Lance Waller[4], Jonathan Liff[1,2], John L. Young, Jr. [1,2]

[1]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA.

[2] Metropolitan Atlanta and Rural Georgia SEER Registry, Georgia Center for Cancer Statistics, Atlanta, GA.

[3]Department of Health Policy and Management, Rollins School of Public Health, Emory University, Atlanta, GA.

[4]Department of Biostatistics, Rollins School of Public Health, Emory University, Atlanta, GA.

*Correspondence to: Kevin C. Ward, MPH, Georgia Center for Cancer Statistics, Emory University, 1462 Clifton Road, NE, 5th Floor, Atlanta, GA. 30322 (email: kward@sph.emory.edu)

Key Words: prostate cancer, survival, socioeconomic status, area-based measures, Cox regression

ABSTRACT

Background: The mechanisms by which socioeconomic status (SES) affect prostate cancer survival are not well understood. Stage, grade, and treatment are all important explanatory factors, but they cannot completely explain survival disparities across different SES strata. We examined non-localized prostate cancer survival among the socioeconomically diverse Medicare population with equal eligibility for care. The purpose of this study is to explore the extent to which various demographic, clinical, and social factors explain the association between SES and prostate cancer survival in this population.

Methods: The SEER-Medicare linked data was used to identify a population-based cohort of 5,368 men, between the ages of 66 and 79 years, diagnosed with non-localized prostate cancer. The cohort was restricted to men enrolled in fee-for-service Medicare Part A and Part B and all men were followed though 2004. Death from prostate cancer was the primary outcome variable in Cox regression models

Results: Relative to low poverty areas, high poverty areas were characterized by a larger percentage of blacks, a smaller percentage of married men, and a larger percentage of men with comorbidities and metastatic tumors. Significant interaction was observed between poverty and stage. Unadjusted hazard ratios (HR) comparing the highest to lowest poverty strata were 1.48 (95% CI 0.86, 2.54) for stage III cancers, 2.09 (95% CI 1.35, 3.22) for stage IV-M0 cancers, and 1.03 (95% CI 0.79, 1.09) for stage IV-M1

cancers. After adjustment for marital status, stage migration, comorbidities, age, grade, treatment, area of residence, and race, the HRs were as follows: 1.21 (95% CI 0.66, 2.21) for stage III cancers, 1.58 (95% CI 0.94 2.67) for state IV-M0 cancers, and 0.90 (95% CI 0.72, 1.13) for stage IV-M1 cancers.

Conclusions: Prostate cancer survival differs by SES even among individuals with equal eligibility for care. Much of the SES related disparities are explained by factors for which SES may serve as a surrogate. These underlying factors include stage at diagnosis, social support as reflected in marital status, and comorbidities.

INTRODUCTION

Prostate cancer, the leading incident cancer in men (1), is a unique malignancy because of its typical indolent nature. When diagnosed at a localized stage, average survival approaches 100 percent after controlling for background sources of mortality (1). Once the disease has extended beyond the prostate, however, the cancer is considered non-localized and survival begins to decline. While disparities in prostate cancer survival have been extensively studied in the literature, most research has specifically focused on the relationship of survival and race (2-14) rather than survival and SES (15-18). In a recent publication from the Surveillance, Epidemiology, and End Results (SEER) Program, men in high poverty census tracts had poorer 5-year prostate cancer-specific survival for both whites and blacks. The largest differences were observed in men with distant stage disease (18).

SES is a complex measure encompassing both individual and contextual components (19, 20). Stage, grade, and treatment are all important explanatory factors in this relationship, but they cannot completely explain survival disparities across different SES strata (13, 15, 16). Observed SES effects on survival may also operate through numerous complex pathways including but not limited to access to care, utilization of care, quality of care, cancer screening, treatment delivery, treatment delays, social support and co-morbid conditions. The mechanisms by which SES affects survival are not well understood.

This paper examines non-localized prostate cancer survival among Medicare recipients. The SEER-Medicare linked data provide an opportunity to study a socioeconomically

diverse population with equal eligibility for care and to enhance the SEER data by adding information on treatment and comorbidities. The aim of these analyses is to explore the extent to which various demographic, clinical, and social factors explain the association between SES and prostate cancer survival in this population.

MATERIALS AND METHODS

Data Sources

The SEER Program is the leading source of population-based data on cancer incidence and survival in the United States (21). The 18 registries that currently comprise this program cover approximately 25% of the U.S. population and collect detailed information on cancer patient demographics, tumor characteristics, extent of disease, and first course therapy.

Medicare is the federally funded program offering health insurance to 97% of U.S. residents over the age of 65 (22). Medicare Part A covers inpatient care at hospitals and skilled nursing facilities. It is generally available without a premium to individuals meeting Medicare eligibility requirements. Medicare Part B requires a premium but adds the additional coverage of physician and outpatient services. In addition to traditional fee-for-service insurance, Medicare also offers HMO services. Individuals must be enrolled in both Medicare Part A and Part B to join a Medicare HMO (22).

Every three years, SEER data are linked with administrative claims and enrollment files for the Medicare Program. Details of the linkage have been previously described (23).

The most current linkage includes SEER data through diagnosis year 2002 and Medicare claims through 2005.

Case Selection

The study population was restricted to invasive prostate cancer cases (International Classification of Diseases for Oncology, 2nd Edition, code: 'C619') diagnosed between 1996 and 2002 in one of the 12 SEER Registries that collected data during this entire period. Cases were limited to those with an American Joint Committee on Cancer (AJCC) 5th edition stage of III or IV (i.e. cancers that have extended beyond the prostate) (24). Individuals under the age of 66 were excluded to allow one full year of Medicare coverage prior to diagnosis. Men over the age of 79 were also excluded due to an age-SES interactive effect that made the interpretation of the results difficult in the presence of other existing sources of interaction. The study population was further restricted by excluding second and later primaries, autopsy and death certificate only cases, cases with a race other than black or white, and cases with an unknown census tract. The resulting dataset contained 8,833 unique prostate cancer cases.

To ensure availability of cancer patient care during the entire study period, cases were also excluded from analyses if not enrolled in both fee-for-service Medicare Part A and Part B throughout follow-up (N=3,216). In addition, to optimize ascertainment of comorbidities, cases were further excluded if not enrolled with the same coverage as above for the entire year prior to diagnosis (N=249). As a result the final study population included 5,368 prostate cancer cases.

Outcome Variable

The primary outcome of interest was death from prostate cancer. SEER Registries follow all cases on an annual basis to ascertain vital status, date of last contact, and cause of death (26). Causes of death were obtained from the underlying cause provided on state mortality files or National Death Index files (27). For this study, follow-up was complete through December 31, 2004. Cases dying of causes other than prostate cancer were treated as censored observations as were cases surviving past the follow-up period. Survival time was calculated in months from diagnosis to death or censoring since the individual day of diagnosis is not reported to SEER.

Area-Based Measures of SES

In the U.S., census tracts are small permanent statistical subdivisions of a county that are designed to be homogeneous with respect to socioeconomic status and living conditions (28). Census tracts can be linked with poverty data from the 2000 U.S. Census Summary File 3 (29). Federal standards define census tracts with twenty percent or more of the population living below the poverty level as "poverty areas" (30). The designation of poverty level varies with family size, income, and year (e.g., $17,463 for a family of 4 in calendar year 2000).

Research suggests that measures of economic deprivation at the level of the census tract, such as the percent of the census population living below the poverty level, are the most effective measures for evaluating health disparities using linked census data (31). These

measures demonstrate consistent gradients across population subgroups, are robust across a range of disease outcomes, allow for maximal linkage, and are easy to understand and explain. For this study, each census tract was classified according to the percent of the tract population living below the poverty level into one of three groups: 0-9.9% (low poverty), 10-19.9% (middle poverty), and 20-100% (high poverty).

Other Covariates

The covariates analyzed for this study were grouped into categories as follows:

- Host Factors
  - race: white or black
  - diagnosis age: 66-69, 70-74, 75-79
  - comorbidities: Charlson score 0, Charlson score 1+
- Geography
  - county at diagnosis: metropolitan or urban/rural
- Family Support
  - marital status: married, single, separated/divorced, widowed
- Tumor Characteristics
  - grade: well or moderately differentiated, poorly differentiated, unknown
  - stage: AJCC stage III, stage IV-M0 or stage IV-M1
  - tumor upstaged from clinical stage I/II: yes or no
- Treatment
  - first course treatment given: yes or no

Grades 1 (well differentiated) and 2 (moderately differentiated) were combined due to the small number of well-differentiated tumors (n=76). Stage of disease was categorized using the definitions from the AJCC 5th edition (24). In addition, stage IV cancers were further classified into those with distant metastasis (stage IV-M1) and those without distant metastasis (stage IV-M0). An additional dichotomous variable was created to identify tumors clinically diagnosed as stage I or II but subsequently upstaged to pathologic stage III or IV.

While SEER collects detailed information on first course therapy, the only date of therapy collected is the date of first course treatment from any modality. SEER only releases treatment data on surgery and radiation therapy, along with date of first course therapy from any modality, due to known limitations with the completeness of the data on chemotherapy and hormone therapy. Medicare claims were used to supplement SEER treatment data and to update the date of first course therapy where the documented administration of hormone therapy from Medicare claims provided a date of first course treatment earlier than what was present in SEER. Medicare physician and outpatient files were searched for hormone therapy administration in the 6 month period after diagnosis using HCPCS codes as described elsewhere (32).

Analyses

All data were analyzed using SAS 9.1 (Cary, NC). Distributions of case characteristics across poverty strata were compared using chi-square tests. Unadjusted prostate-specific and other-cause 5-year survival measures were calculated for each study variable using

life table methodology with monthly intervals (33-35). Differences in survival across poverty categories were tested for significance using log-rank tests. Unadjusted and adjusted hazard ratios (HR) and 95% confidence intervals (CI) were calculated using Cox proportional hazards regression models (36). The proportional hazards assumption was assessed for each variable by examining the log-log survival curves (37). In the multivariable analyses each individual covariate was tested for interaction with the primary exposure variable (poverty). If interaction was present, stratified analyses were conducted. A forward stepwise regression approach was used to examine the effects of individual covariates on the association between poverty and prostate cancer mortality.

RESULTS

Figure 6.1 presents the life table survival curves for the three poverty strata. Individuals in the lowest poverty group (0-9.9%) experienced the best survival while those in the highest poverty group (20+%) experienced the poorest survival. Survival experience of men in the middle poverty stratum was similar to that in the low poverty group. The differences in survival were statistically significant (log-rank $p<.001$) across the groups and increased with time.

The baseline characteristics of the study cases are presented in Table 6.1. Overall, 63% of the cases were in the low poverty group while only 13% were in the high poverty group. Relative to the low poverty census tracts, high poverty tracts were characterized by a larger percentage of blacks (in fact, only 12% of the black population resided in the lowest poverty tracts), a smaller percentage of married men and a larger percentage of

men with comorbidities. High poverty areas also included a larger percentage of poorly differentiated tumors, unknown grade tumors and metastatic tumors. Statistically significant differences across the three poverty groups were observed for all covariates except age of diagnosis.

Unadjusted measures of prostate cancer-specific and other-cause survival are presented in Table 6.2. The 5-year prostate cancer-specific survival was 78% for low poverty areas, 74% for middle poverty areas and 66% for high poverty areas. Stage and grade were much stronger predictors of prostate cancer-specific survival than other-cause survival. By contrast, comorbidities and the absence of documented first course therapy were the stronger predictors of other-cause survival.

Compared to individuals living in the low poverty areas, those in the highest poverty group had a significant 66% increase in prostate cancer-specific mortality and a significant 45% increase in other-cause mortality. Relative to married men, the HRs reflecting prostate cancer-specific mortality were 1.78 (95% CI 1.48, 2.16) for single men, 1.95 (95% CI 1.59, 2.39) for separated or divorced men, and 2.21 (95% CI 1.85, 2.63) for widowed men. Analyses of the association between marital status and other-cause mortality produced similar results. Among all independent variables of interest, only non-metropolitan (urban or rural) county of residence showed significant association with prostate cancer-specific, but not with other-cause survival.

Table 6.3 presents the results of the forward stepwise Cox regression modeling for prostate cancer-specific survival. Due to significant (p=0.02) interaction between poverty and AJCC stage, all Cox regression analyses were stratified by stage. Among patients with stage IV metastatic disease (M1), the analysis of the unadjusted association between poverty and survival demonstrated essentially null results. By contrast, among cases with stage IV disease, but without distant metastases (M0), men residing in high poverty areas experienced a more than two-fold increase in prostate cancer-specific mortality (HR=2.09: 95% CI: 1.35-3.22) compared to men residing in the low poverty areas. The corresponding results for stage III prostate cancer were less pronounced and did not reach statistical significance.

The step-wise addition of potential confounders resulted in gradual attenuation of the observed association between poverty and survival among stage III and stage IV (M0) cases, whereas the results for stage IV (M1) disease continued to show no discernable departure from the null. In the final all-inclusive model, which adjusted for marital status, upstaged clinical disease, the presence of comorbidities, age, grade, treatment, race, and county of diagnosis the results were as follows. Among stage III cases the HRs for high (versus low) and middle (versus low) poverty areas were 1.21 (95% CI: 0.66-2.21) and 0.83 (95% CI: 0.52-1.32), respectively. For stage IV (M0) disease the corresponding results were 1.58 (95% CI: 0.94-2.67) for high poverty and 1.38 (95% CI: 0.90-2.11) for middle poverty; whereas among cases with metastatic disease the HR for high poverty was 0.90 (95% CI: 0.72-1.13) and the HR for middle poverty was 0.88 (95% CI: 0.75-1.05).

The final stratified model confirmed that the independent effects of social, demographic and clinical factors on prostate cancer survival could vary by stage (Table 6.4). Residence in a non-metropolitan county and divorce/separation from a spouse were associated with increased mortality for stage III, but not for stage IV cases. Treatment appeared to improve survival in stage IV (M1) cases, but not in stage III or stage IV (M0) cases. Presence of comorbidities affected stage IV (M1) and particularly stage IV (M0), but not stage III prostate cancer mortality. On the other hand, upstaged clinical disease and tumor grade demonstrated significant associations with survival irrespective of stage. After adjusting for other demographic, social and clinical characteristics there was no evidence of the association between race and survival; the HRs were close to 1.0 for all disease stages.

DISCUSSION

Area-based measures of socioeconomic status were utilized in this study for two primary reasons. First, population-based disease registries do not generally collect individual measures of SES as their data primarily come from medical records. More importantly, research has indicated that among persons over the age of 65, area-based measures of SES may be more relevant than individual measures such as education and annual income (38). The neighborhood in which one lives captures aspects of living conditions not necessarily defined by individual measures; it is a relevant characteristic that applies to all of its residents regardless of age and gender, and it is a moderately stable measure of socioeconomic conditions (19).

In this study, we found that the association between SES and survival is complex as it appears to differ by stage. While the effect was virtually absent among men with metastatic disease, poverty played a more important role in cases with less advanced disease. This is not surprising as stage III, stage IV-M0 and stage IV-M1 prostate cancers are very different conditions with respect to the natural history of disease and recommended treatment approaches (39). Once prostate cancer has metastasized to distant nodes or organs (typically bone), the prognosis is poor irrespective of access to care. In other stages, treatment selection depends primarily on age, comorbidities and current symptoms and typically involves external beam radiation therapy, androgen ablation therapy or a combination of both. Regardless of the modality, repeated visits to the treating oncologist are generally necessary for administration of therapy so access to care is critically important.

Access to care has been shown in several studies to explain away the racial differences in prostate cancer survival (5, 8). In addition, McDavid and colleagues demonstrated that disparities in prostate cancer survival exist by level of individual insurance after controlling for age, sex, race, stage and treatment (17). In their study, 3-year relative survival from prostate cancer was best for patients with private insurance followed by Medicare with supplement, Medicare, other federally funded insurance, not insured and then Medicaid. In our study, restriction of the population to fee-for-service Medicare recipients provided a mechanism to control for eligibility for care. While equal eligibility

does not equate to equal access, it does make certain that all patients have similar health insurance.

In our SEER-Medicare population with equal eligibility for care, we found that controlling for demographic, clinical, and social support factors resulted in an attenuation of the observed associations between poverty and survival. Factors other than SES appear to be more important in the multivariable analysis. For example, marital status appears to be an important determinant of survival although the magnitude of its effect differs by stage. Of interest are the differences observed for separated, divorced or widowed men. Our results indicate the lack of social support, where it once existed, clearly played a detrimental role. Marriage has been shown in other studies to have a favorable effect on cancer survival (40-45) and has been directly associated with the utilization of more aggressive treatment (46). Some have suggested it gives the individual a greater constitution to fight the disease (41).

Stage migration (47), as measured by upstaged clinical disease in this study, also played an important role in explaining SES related differences in survival. While the AJCC recommends collection of both clinical and pathologic stage for most cancers, this level of detail is not generally present in SEER data. Prostate cancer represents the only exception. Since 1995, SEER has collected both clinical and pathologic tumor extension thus allowing the examination of stage migration. Men diagnosed clinically with stage I or II prostate cancer are likely to have a very different and more favorable prognosis than men diagnosed clinically as having stage III or IV disease. Men with upstaged clinical

disease were likely free of cancer specific symptoms and were probably identified via prostate cancer screening. Since the majority of upstaged cases were moved to stage III, larger effects of this variable on the SES-survival association were seen among stage III cases.

The presence of comorbid disease at the time of diagnosis also played an important role in our analysis. Comorbidities have been associated with increasing age and lower education and are predictive of longer hospital stays, increased hospital mortality and frequency of readmission (48-51). Klaubande et al. specifically showed that the Charlson comorbidity measure calculated from Medicare physician claims was predictive of both less aggressive treatment and higher mortality (52). While it is difficult to explain the exact pathway by which comorbidities operated in our study, in the final model the presence of comorbid disease showed a significant inverse association with survival among men with stage IV prostate cancer.

Any inference of our study findings to other populations requires caution. Our results are not necessarily applicable to men under the age of 65 or over the age of 80. They are also not entirely generalizable to men with insurance coverage other than Medicare fee-for-service. Studies have shown that cancer care and cancer survival does vary by insurance type (17). On the other hand, a substantial proportion of prostate cancers occur between the ages of 66 and 79, and a large percentage of men in this age group have fee-for-service Medicare.

The use of area-based measures of SES assumes homogeneity within the geographic units, census tracts in our case. As the size of census tracts are based on population density, tracts are generally larger in non-metropolitan areas. If the tracts get too large, the populations may be less homogeneous and observed survival differences may be biased (53). It is also important to point out that the SEER-Medicare dataset does not provide information regarding the certainty of the census tracts obtained. When an exact street address cannot be geocoded, the centroid of the residence ZIP code is used instead (54). If the centroid does not possess the same SES characteristics as the exact street address, this may lead to misclassification.

Our analysis used cause-specific survival analysis. This approach requires patient-specific cause of death and the validity of results is determined by the accuracy and completeness of the death certificate coding. As comorbidities increase with age, competing causes of death become an important component of the death rate. While relative survival analyses would have been preferable, these are not possible at the present time because SES-specific expected rate tables are not available at the national level in the U.S. This issue is less of a concern, however, for prostate cancer. Several studies that conducted detailed reviews of death certificates have documented that cause-specific analyses are adequate for prostate cancer (55, 56). Our analyses support these observations by demonstrating that tumor specific characteristics such as grade and stage were much stronger predictors of prostate cancer-specific mortality as compared to other-cause mortality.

The results of our study indicate that prostate cancer survival clearly differs by SES even among individuals with equal eligibility for care. We also found that much of the SES related disparities are explained by factors for which SES may serve as a surrogate. These underlying factors include stage at diagnosis, social support as reflected in marital status, and comorbidities. While consideration of these extraneous factors clearly attenuates the observed association, it does not completely eliminate the effect of SES on survival. It is possible that lower SES affects survival by mechanisms that were not captured in these analyses. For example, we did not have the data on the quality of care received by individuals in our study. Although the study was restricted to individuals with equal eligibility for care, we did not have any reliable measures of access to care and health care utilization. It is also possible that even after controlling for extraneous factors, the observed association can be attributed to residual confounding. For example, the Charlson index only includes some of the conditions known to affect mortality. Although we were able to control for first course therapy in this study, it is possible that a more sophisticated analysis taking into consideration specific treatment modalities would have further attenuated the effect of SES on survival.

CONCLUSION

The clear and consistent evidence of SES-related survival disparities for non-localized prostate cancer offers limited insight into potential intervention strategies. Such strategies require a comprehensive deconstruction of the SES-mortality association to identify factors that can be modified. In the presence of equal eligibility for care, our

study points toward lack of social support and late stage or delayed diagnosis as the two most important modifiable determinants of survival.

Figure 6.1 Prostate cancer survival by area-based poverty measure in the Medicare population, stages III-IV, ages 66-79, 12 SEER areas, 1996-2002

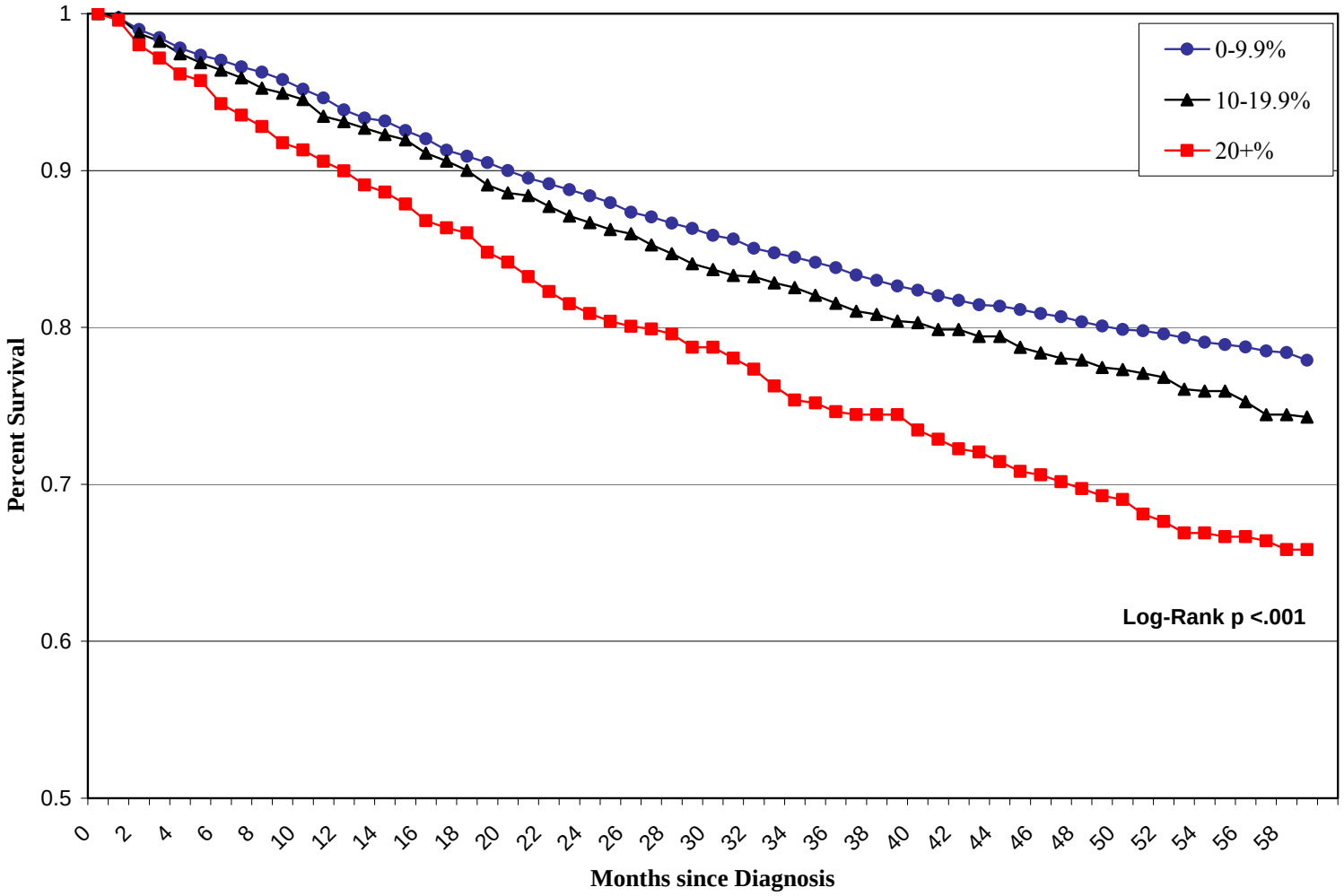

TABLE 6.1 Baseline characteristics of SEER Registry invasive non-localized prostate cancer cases* according to census tract poverty measure

Characteristic	Low Poverty (0-9.9%) N=3,392		Middle Poverty (10-19.9%) N=1,268		High Poverty (20+%) N=708	
	No.	%	No.	%	No.	%
Race**						
White	3,319	97.8	1,084	84.5	359	50.7
Black	73	2.2	184	14.5	349	49.3
Diagnosis Age						
66-69	1,144	33.7	429	33.8	224	31.6
70-74	1,354	39.9	493	38.9	291	41.1
75+	894	26.4	346	27.3	193	27.3
County at diagnosis**						
Metropolitan	2,844	83.8	902	71.1	593	83.8
Urban or Rural	548	16.2	366	28.9	115	16.2
Marital ~**						
Married	2,657	80.7	933	75.8	372	55.0
Single	209	6.4	104	8.5	122	18.1
Separated/Divorced	158	4.8	85	6.9	96	14.2
Widowed	268	8.1	108	8.8	86	12.7
Grade**						
Well\Mod Diff	1,633	48.2	581	45.8	291	41.1
Poorly Diff	1,487	43.8	558	44.0	335	47.3
Unknown	272	8.0	129	10.2	82	11.6
AJCC Stage 5th ed**						
Stage III	1,850	54.5	638	50.3	285	40.3
Stage IV - M0	561	16.6	196	15.5	110	15.5
Stage IV - M1	981	28.9	434	34.2	313	44.2
Upstaged from Clinical Stage I/II**						
Yes	1,301	38.3	408	32.2	179	25.3
No	2,091	61.7	860	67.8	529	74.7
Rcvd Any Treatment**						
Yes	3,301	97.3	1,213	95.7	668	94.4
No	91	2.7	55	4.3	40	5.6
Charlson Comorbidity Score**						
0	2,967	87.5	1,105	87.2	574	81.1
>=1	425	12.5	163	12.8	134	18.9

*Study population limited to cases eligible for survival analysis

**Statistically significant at p<.001

~ Unknown values not included (N=170)

TABLE 6.2 Unadjusted 5-year other-cause and prostate-specific survival with corresponding hazard ratios (HR) and 95% confidence intervals (CI) for primary exposure (poverty) and selected covariates

Characteristic		5-Year Prostate Specific Survival (%)[#]	5-yr Prostate Specific HR	95% CI	5-Yr Other Cause Survival (%)[#]	5-yr Other Cause HR	95% CI
Poverty							
	Low (0-9.9%)	77.9	1.00		84.6	1.00	
	Middle (10-19.9%)	74.3	1.16	(1.01, 1.34)	79.6	1.34	(1.18, 1.78)
	High (20+%)	65.8	1.66	(1.42, 1.94)	78.4	1.45	(1.14, 1.58)
Race							
	White	77.0	1.00		83.4	1.00	
	Black	63.2	1.81	(1.55, 2.12)	76.4	1.57	(1.28, 1.92)
Diagnosis Age							
	66-69	85.2	1.00		88.6	1.00	
	70-74	76.3	1.71	(1.46, 2.01)	83.2	1.47	(1.22, 1.77)
	75+	60.9	3.20	(2.73, 3.74)	72.4	2.70	(2.24, 3.25)
County at diagnosis							
	Metropolitan	76.5	1.00		82.6	1.00	
	Urban or Rural	71.5	1.18	(1.03, 1.36)	82.9	0.94	(0.78, 1.13)
Marital							
	Married	79.2	1.00		84.8	1.00	
	Single	67.1	1.78	(1.48, 2.16)	78.7	1.47	(1.15, 1.89)
	Separated/Divorced	62.4	1.95	(1.59, 2.39)	80.7	1.37	(1.03, 1.83)
	Widowed	60.1	2.21	(1.85, 2.63)	69.6	2.21	(1.79, 2.74)
Grade							
	Well\Mod Diff	88.6	1.00		86.6	1.00	
	Poorly Diff	68.9	3.10	(2.68, 3.59)	81.3	1.48	(1.27, 1.73)
	Unknown	30.9	11.28	(9.48, 13.43)	59.7	4.21	(3.39, 5.24)
AJCC Stage 5th ed							
	Stage III	95.0	1.00		88.6	1.00	
	Stage IV - M0	80.5	4.24	(3.32, 5.42)	83.7	1.55	(1.25 ,1.91)
	Stage IV - M1	36.7	21.50	(17.7, 26.12)	67.5	3.53	(3.01, 4.14)
Upstaged from Clinical Stage I/II							
	Yes	97.7	1.00		92.2	1.00	
	No	62.1	22.41	(16.1, 31,.2)	76.1	3.60	(2.97, 4.35)
Rcvd Any Treatment							
	Yes	76.2	1.00		83.7	1.00	
	No	51.8	3.56	(2.79, 4.54)	41.7	6.04	(4.74, 7.70)
Charlson Comorbidity Score							
	0	78.1	1.00		85.4	1.00	
	>=1	56.6	2.53	(2.20, 2.91)	61.5	3.36	(2.86, 3.94)

#Calculated using life table methodology (all log-rank tests were significant with the exception of County at Diagnosis in the other cause model)

TABLE 6.3 Cox regression modeling to examine the role of effect modification (covariate with poverty) and individual covariates in explaining area-based differences in poverty

	High (20+%) vs. Low (0-9.9%)		Middle (10-19.9%) vs. Low (0-9.9%)	
Covariates in model	HR	95% CI	HR	95% CI
Poverty stratified by AJCC Stage				
Stage III	1.48	(0.86, 2.54)	1.41	(0.74, 2.54)
Stage IV - M0	2.09	(1.35, 3.22)	1.39	(0.94, 2.07)
Stage IV - M1	1.03	(0.79, 1.09)	0.93	(0.79, 1.09)
Poverty, Marital Status, stratified by AJCC Stage				
Stage III	1.42	(0.82, 2.45)	1.04	(0.66, 1.62)
Stage IV - M0	1.74	(1.09, 2.77)	1.34	(0.89, 2.02)
Stage IV - M1	0.94	(0.78, 1.13)	0.93	(0.79, 1.09)
Poverty, Marital Status, Upstaged clinical disease, stratified by AJCC Stage				
Stage III	1.32	(0.76, 2.30)	0.97	(0.62, 1.52)
Stage IV - M0	1.68	(1.06, 2.68)	1.33	(0.88, 2.00)
Stage IV - M1	0.94	(0.78, 1.14)	0.92	(0.79, 1.09)
Poverty, Marital Status, Upstaged clinical disease, Comorbidities stratified by AJCC Stage				
Stage III	1.31	(0.76, 2.28)	0.97	(0.62, 1.52)
Stage IV - M0	1.55	(0.97, 2.48)	1.36	(0.90, 2.05)
Stage IV - M1	0.91	(0.76, 1.10)	0.91	(0.78, 1.07)
Poverty, Marital Status, Upstaged clinical disease, Comorbidities, Diagnosis Age, stratified by AJCC Stage				
Stage III	1.31	(0.76, 2.27)	0.97	(0.61, 1.52)
Stage IV - M0	1.60	(1.00, 2.57)	1.46	(0.96, 2.20)
Stage IV - M1	0.93	(0.77, 1.12)	0.92	(0.78, 1.09)

TABLE 6.3 (continued) Cox regression modeling to examine the role of effect modification (covariate with poverty) & individual covariates in explaining area-based differences in poverty

	High (20+%) vs. Low (0-9.9%)		Middle (10-19.9%) vs. Low (0-9.9%)	
Covariates in model	HR	95% CI	HR	95% CI
Poverty, Marital Status, Upstaged clinical disease, Comorbidities, Diagnosis Age, Grade, stratified by AJCC Stage				
Stage III	1.28	(0.74, 2.21)	0.93	(0.59, 1.45)
Stage IV - M0	1.58	(0.99, 2.53)	1.41	(0.93, 2.14)
Stage IV - M1	0.93	(0.77, 1.12)	0.89	(0.76, 1.05)
Poverty, Marital Status, Upstaged clinical disease, Comorbidities, Diagnosis Age, Grade, Treatment, County Dx, stratified by AJCC Stage				
Stage III	1.24	(0.72, 2.14)	0.84	(0.53, 1.32)
Stage IV - M0	1.54	(0.96, 2.47)	1.36	(0.90, 2.07)
Stage IV - M1	0.90	(0.75, 1.09)	0.88	(0.75, 1.04)
Poverty, Marital Status, Upstaged clinical disease, Comorbidities, Diagnosis Age, Grade, Treatment, County Dx, Race, stratified by AJCC Stage				
Stage III	1.21	(0.66, 2.21)	0.83	(0.52, 1.32)
Stage IV - M0	1.58	(0.94, 2.67)	1.38	(0.90, 2.11)
Stage IV - M1	0.90	(0.72, 1.13)	0.88	(0.75, 1.05)

TABLE 6.4 Hazard ratios (HR) and 95% confidence intervals (CI) for primary exposure and other covariates in the final all inclusive stratified Cox regression model

		Stage III		Stage IV-M0		Stage IV-M1	
		HR	95% CI	HR	95% CI	HR	95% CI
Poverty							
	Low (0-9.9%)	1.00		1.00		1.00	
	Middle (10-19.9%)	0.83	(0.52, 1.32)	1.38	(0.90, 2.11)	0.88	(0.75, 1.05)
	High (20+%)	1.21	(0.66, 2.21)	1.58	(0.94, 2.67)	0.90	(0.72, 1.13)
Race							
	White	1.00		1.00		1.00	
	Black	1.07	(0.51, 2.26)	0.94	(0.52, 1.69)	1.00	(0.80, 1.25)
Age of diagnosis							
	66-69	1.00		1.00		1.00	
	70-74	1.09	(0.70, 1.72)	1.36	(0.85, 2.14)	1.13	(0.94, 1.37)
	75+	1.29	(0.78, 2.13)	2.33	(1.48, 3.68)	1.28	(1.06, 1.55)
County of diagnosis							
	Metropolitan	1.00		1.00		1.00	
	Urban or Rural	2.15	(1.43, 3.23)	1.20	(0.77, 1.86)	0.99	(0.83, 1.18)
Marital status							
	Married	1.00		1.00		1.00	
	Single	0.71	(0.31, 1.64)	1.20	(0.63, 2.27)	1.40	(1.13, 1.73)
	Separated/Divorced	2.34	(1.25, 4.37)	1.67	(0.95, 2.91)	1.10	(0.86, 1.40)
	Widowed	0.84	(0.38, 1.84)	1.82	(1.06, 3.14)	1.33	(1.10, 1.62)
Grade							
	Well\Mod Diff	1.00		1.00		1.00	
	Poorly Diff	3.39	(2.25, 5.11)	2.25	(1.51, 3.36)	1.95	(1.63, 2.33)
	Unknown	2.84	(0.86, 9.42)	4.23	(1.91, 9.38)	2.83	(2.32, 3.45)
Upstaged from Clinical Stage I/II							
	Yes	1.00		1.00		1.00	
	No	4.22	(2.72, 6.55)	5.45	(1.99, 14.89)	1.79	(0.25, 12.82)
Received Any Treatment							
	Yes	1.00		1.00		1.00	
	No	0.89	(0.22, 3.66)	0.44	(0.06, 3.34)	1.70	(1.30, 2.22)
Charlson Comorbidity Score							
	0	1.00		1.00		1.00	
	>=1	1.51	(0.86, 2.64)	2.26	(1.49, 3.43)	1.36	(1.15, 1.60)

REFERENCES

1. Ries LAG, Melbert D, Krapcho M, et al. SEER Cancer Statistics Review, 1975-2004. National Cancer Institute Bethesda, MD, http://seercancergov/csr/1975_2004/, based on November 2006 SEER data submission, posted to the SEER web site, 2007.
2. Bach PB, Schrag D, Brawley OW, et al. Survival of blacks and whites after a cancer diagnosis.[see comment]. JAMA 2002;287:2106-13.
3. Fowler JE, Jr., Bigler SA, Bowman G, et al. Race and cause specific survival with prostate cancer: influence of clinical stage, Gleason score, age and treatment.[see comment]. J Urol 2000;163:137-42.
4. Hart KB, Wood DP, Jr., Tekyi-Mensah S, et al. The impact of race on biochemical disease-free survival in early-stage prostate cancer patients treated with surgery or radiation therapy. Int J Radiat Oncol Biol Phys 1999;45:1235-8.
5. Optenberg SA, Thompson IM, Friedrichs P, et al. Race, treatment, and long-term survival from prostate cancer in an equal-access medical care delivery system. JAMA 1995;274:1599-605.
6. Pienta KJ, Demers R, Hoff M, et al. Effect of age and race on the survival of men with prostate cancer in the Metropolitan Detroit tricounty area, 1973 to 1987. Urology 1995;45:93-101; discussion -2.
7. Polednak AP. Black-white differences in survival from late-stage prostate cancer. Ethn Dis 2003;13:220-5.
8. Powell IJ, Schwartz K, Hussain M. Removal of the financial barrier to health care: does it impact on prostate cancer at presentation and survival? A

comparative study between black and white men in a Veterans Affairs system. Urology 1995;46:825-30.

9. Roach M, 3rd, Krall J, Keller JW, et al. The prognostic significance of race and survival from prostate cancer based on patients irradiated on Radiation Therapy Oncology Group protocols (1976-1985). Int J Radiat Oncol Biol Phys 1992;24:441-9.

10. Roach M, 3rd, Lu J, Pilepich MV, et al. Race and survival of men treated for prostate cancer on radiation therapy oncology group phase III randomized trials. J Urol 2003;169:245-50.

11. Robbins AS, Whittemore AS, Thom DH. Differences in socioeconomic status and survival among white and black men with prostate cancer.[see comment]. Am J Epidemiol 2000;151:409-16.

12. Robbins AS, Whittemore AS, Van Den Eeden SK. Race, prostate cancer survival, and membership in a large health maintenance organization.[see comment]. J Natl Cancer Inst 1998;90:986-90.

13. Robbins AS, Yin D, Parikh-Patel A. Differences in prognostic factors and survival among White men and Black men with prostate cancer, California, 1995-2004. Am J Epidemiol 2007;166:71-8.

14. Thompson I, Tangen C, Tolcher A, et al. Association of African-American ethnic background with survival in men with metastatic prostate cancer.[see comment]. J Natl Cancer Inst 2001;93:219-25.

15. Clark JY, Thompson IM. Military rank as a measure of socioeconomic status and survival from prostate cancer. South Med J 1994;87:1141-4.

16. Du XL, Fang S, Coker AL, et al. Racial disparity and socioeconomic status in association with survival in older men with local/regional stage prostate carcinoma: findings from a large community-based cohort. Cancer 2006;106:1276-85.

17. McDavid K, Tucker TC, Sloggett A, et al. Cancer survival in Kentucky and health insurance coverage.[see comment]. Arch Intern Med 2003;163:2135-44.

18. Singh GK, Miller BA, Hankey BF, et al. Area Socioeconomic Variations in US Cancer Incidence, Mortality, Stage, Treatment and Survival, 1975-1999. Bethesda, MD., National Cancer Institute, NCI Cancer Monograph Series, Number 4, NIH Pub No. 03-5417, 2003.

19. Krieger N, Williams DR, Moss NE. Measuring social class in US public health research: concepts, methodologies, and guidelines. Annu Rev Public Health 1997;18:341-78.

20. Liberatos P, Link BG, Kelsey JL. The measurement of social class in epidemiology. Epidemiol Rev 1988;10:87-121.

21. National Cancer Institute. SEER: Surveillance, Epidemiology and End Results Program. National Institute of Health (Publication No. 05-4772), 2005.

22. U.S. Department of Health and Human Services. Centers for Medicare and Medicaid Services, 2007. (http://www.cms.hhs.gov/home/medicare.asp).

23. Warren JL, Klabunde CN, Schrag D, et al. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. Med Care 2002;40:IV-3-18.

24. American Joint Committee on Cancer. AJCC Cancer Staging Manual, 5th Edition. Philadelphia: Lippincott-Raven, 1997.

25. National Center for Health Statistics. Health, United States, 2006, With Chartbook on Trends in the Health of Americans. Hyattsville, MD:, 2006.

26. Fritz A, Ries L, (eds). The SEER Program Code Manual. National Cancer Institue, Bethesda, MD, 1998.

27. National Center for Health Statistics. National Death Index, 2007. (http://www.cdc.gov/nchs/ndi.htm).

28. U.S. Bureau of the Census. Geographic Areas Reference Manual, 2007. (http://www.census.gov/geo/www/garm.html).

29. U.S. Bureau of the Census. 2000 Census Population and Housing, Summary File 3: Technical Documentation. 2002.

30. U.S. Bureau of the Census. Poverty Areas, 2007. (http://www.census.gov/hhes/www/poverty/definitions.html).

31. Krieger N, Chen JT, Waterman PD, et al. Race/ethnicity, gender, and monitoring socioeconomic gradients in health: a comparison of area-based socioeconomic measures--the public health disparities geocoding project. Am J Public Health 2003;93:1655-71.

32. Shahinian VB, Kuo Y-F, Freeman JL, et al. Risk of fracture after androgen deprivation for prostate cancer. N Engl J Med 2005;352:154-64.

33. Berkson I, Gage RP. Calculation of Survival Rates for Cancer. Proceedings Staff Meet Mayo Clinic 1950;25:270-86.

34. Cutler SJ, Ederer F. Maximum utilization of the life table method in analyzing survival. Journal of Chronic Diseases 1958;8:699-712.

35. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 1958;53:457-81.

36. Cox DR. Regression models and life tables (with discussion). Journal of the Royal Statistical Society 1972;Series B:187-220.

37. Kleinbaum D, Klein M. Survival Analysis, A Self-Learning Text, Second Edition. New York: Springer, Chapter 4.

38. Rehkopf DH, Haughton LT, Chen JT, et al. Monitoring socioeconomic disparities in death: comparing individual-level education and area-based socioeconomic measures. Am J Public Health 2006;96:2135-8.

39. National Comprehensive Cancer Network and the American Cancer Society. Prostate Cancer: Treatment Guidelines for Patients, Version 1. INTOUCH Magazine 1999:74-103.

40. Ell K, Nishimoto R, Mediansky L, et al. Social relations, social support and survival among patients with cancer. J Psychosom Res 1992;36:531-41.

41. Harvei S, Kravdal O. The importance of marital and socioeconomic status in incidence and survival of prostate cancer. An analysis of complete Norwegian birth cohorts. Prev Med 1997;26:623-32.

42. Kravdal O. The impact of marital status on cancer survival. Soc Sci Med 2001;52:357-68.

43. Kvikstad A, Vatten LJ, Tretli S. Widowhood and divorce in relation to overall survival among middle-aged Norwegian women with cancer. Br J Cancer 1995;71:1343-7.

44. Waxler-Morrison N, Hislop TG, Mears B, et al. Effects of social relationships on survival for women with breast cancer: a prospective study. Soc Sci Med 1991;33:177-83.

45. Polednak AP. Survival of breast cancer patients in Connecticut in relation to socioeconomic and health care access indicators. J Urban Health 2002;79:211-8.

46. Denberg TD, Beaty BL, Kim FJ, et al. Marriage and ethnicity predict treatment in localized prostate carcinoma. Cancer 2005;103:1819-25.

47. Feinstein AR, Sosin DA, Wells CK. The Will Rogers phenomenon: improved technologic diagnosis and stage migration as a source of nontherapeutic improvement in cancer prognosis. Trans Assoc Am Physicians 1984;97:19-24.

48. Klabunde CN, Potosky AL, Harlan LC, et al. Trends and black/white differences in treatment for nonmetastatic prostate cancer. Med Care 1998;36:1337-48.

49. van den Akker M, Buntinx F, Metsemakers JF, et al. Multimorbidity in general practice: prevalence, incidence, and determinants of co-occurring chronic and recurrent diseases. J Clin Epidemiol 1998;51:367-75.

50. Iezzoni LI, Heeren T, Foley SM, et al. Chronic conditions and risk of in-hospital death. Health Serv Res 1994;29:435-60.

51. Librero J, Peiro S, Ordinana R. Chronic comorbidity and outcomes of hospital care: length of stay, mortality, and readmission at 30 and 365 days. J Clin Epidemiol 1999;52:171-9.

52. Klabunde CN, Potosky AL, Legler JM, et al. Development of a comorbidity index using physician claims data. J Clin Epidemiol 2000;53:1258-67.

53. Woods LM, Rachet B, Coleman MP. Choice of geographic unit influences socioeconomic inequalities in breast cancer survival. Br J Cancer 2005;92:1279-82.

54. Tele Atlas. Tele Atlas Geocoding Service - Reference Documentation, 2006.

55. Albertsen PC, Walters S, Hanley JA. A comparison of cause of death determination in men previously diagnosed with prostate cancer who died in 1985 or 1995. J Urol 2000;163:519-23.

56. Penson DF, Albertsen PC, Nelson PS, et al. Determining cause of death in prostate cancer: are death certificates valid? J Natl Cancer Inst 2001;93:1822-3.

Chapter 7

CONCLUSIONS

Population-based cancer registries are a rich source of information regarding the clinical aspects of the disease but are in fact quite limited when it comes to providing information about past exposures leading up to the diagnosis of disease. This limitation of registry data is largely the result of the source records from which the registry data are obtained (i.e. the hospital medical record) and the lack of consistent information on exposures that can be gleaned from this source. Innovative research, however, has shown that existing registry data can be comprehensively expanded by utilizing external data sources and record linkages (1). One readily available external source of complete information is the U.S. Census Summary data (2). These files contain a wealth of information on area-based measures of socioeconomic status (SES) and can easily be linked to population-based registry address data that has been geocoded. Through this process, area-based measures of SES can be appended to the cancer record and utilized in research. The three studies presented in this dissertation were conducted to better understand the role of area-based measures of SES in cancer survival research using population-based registry data. This body of work builds on the results of the Public Health Disparities Geocoding Project (3-5), utilizing the specific area-based measure of SES recommended in this comprehensive project: the percent of the census tract population living below the poverty level.

The first study evaluated the validity of the chosen area-based measure of SES obtained by linking census information to geocoded address data from the Metropolitan Atlanta and Rural Georgia SEER Registry. Acknowledging many of the limitations that exist in the process of geocoding registry data, this study specifically assessed the degree of misclassification that was present when using real data from a population-based cancer registry to assign the chosen area-based measure of SES. The primary findings of this research can be summarized as follows. Some degree of positional error in geocoding based research is inevitable and the need for concern about this error is largely dictated by the research question of interest and the exposure measure of interest. Researchers must both understand and report the percent of data geocoded based on ZIP code centroids as the level of potential misclassification within the study results will increase as this percentage increases. PO Box and Rural Route addresses should be used as a last resort for assigning area-based measures of SES as these locations are subject to the highest degree of misclassification. Vendor assigned geocoded coordinates are typically much more accurate in metropolitan relative to rural areas, largely resulting from issues around population density and the effect this has on the methodology for interpolating geocoded coordinates for a specific street address. Finally, effective methods do exist for cleaning and enhancing registry data that is not geocoded to an exact street location, and this should be a first step in any study wishing to utilize area-based measures of SES from geocoded data.

The second study explored the role of SES-specific background mortality in the calculation of relative survival rates for female breast and prostate cancer using

population-based registry data. Current calculations of relative survival in the U.S. do not control for differences in mortality by SES, despite the fact that numerous studies have clearly demonstrated that mortality varies by SES for both non-cancer and cancer causes of death (6-12). Age, race, sex, and SES-specific life tables were developed for Metropolitan Atlanta and the results from utilizing these new tables were compared against the standard methodology for calculating relative survival which uses national life tables that do not incorporate a measure of SES. A consistent pattern of decreased survival with increased area-based poverty was observed in this study. The use of national life tables to estimate background mortality generally overestimated relative survival in low poverty areas and underestimated relative survival in high poverty areas. The use of SES-specific life tables somewhat diminished the observed SES disparities in survival. When modeling the data using additive hazard models, the excess risk of death among persons living in poorer areas was lower when appropriately controlling for SES-specific background mortality. Traditional cause-specific regression models underestimated the role of poverty in breast cancer analyses but produced almost identical results for prostate cancer.

The final study of this dissertation used SEER-Medicare linked data to explore the extent to which various demographic, clinical, and social factors explain the association between SES and prostate cancer that is observed in the SEER-Medicare population. A significant finding of interaction between SES and stage was observed in these analyses. While there was virtually no observed association between SES and survival among men with metastatic prostate cancer, SES played a more important role in men with less

advanced disease. An association of lower SES and poorer survival was observed in these men with non-metastatic disease even in the presence of equal eligibility for care as afforded by the Medicare system. Most of the SES related disparities in prostate cancer survival, however, were explained by factors for which SES may serve as a surrogate. These underlying factors included stage at diagnosis, social support as reflected in marital status and comorbidities.

In conclusion, area-based measures of SES are a valuable addition to population-based cancer registry data and play an important role in analyses of cancer survival utilizing these data. Further research is needed to both explore possible methods of incorporating SES into national relative survival statistics and to examine the SES-survival relationship in cancers other than prostate. Regardless of the research that is conducted, it is imperative that researchers understand the source data from which the area-based measures were derived. A clear understanding of the completeness and accuracy of the assigned geocodes is critical as these factors have a major effect on potential misclassification of the derived area-based measures.

REFERENCES

1. Krieger N, Chen JT, Waterman PD, et al. Race/ethnicity, gender, and monitoring socioeconomic gradients in health: a comparison of area-based socioeconomic measures--the public health disparities geocoding project. American Journal of Public Health 2003;93:1655-71.
2. U.S. Bureau of the Census. 2000 Census Population and Housing, Summary File 3: Technical Documentation, 2002.
3. Krieger N, Chen J, Waterman P, et al. Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? The Public Health Disparities Geocoding Project. American Journal of Epidemiology 2002;156:471-82.
4. Krieger N, Chen JT, Waterman PD, et al. Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: The Public Health Disparities Geocoding Project (US). Journal of Epidemiology & Community Health 2003;57:186-99.
5. Krieger N, Waterman PD, Chen JT, et al. Monitoring socioeconomic inequalities in sexually transmitted infections, tuberculosis, and violence: geocoding and choice of area-based socioeconomic measures--the public health disparities geocoding project (US). Public Health Reports 2003;118:240-60.
6. Faggiano F, Partanen T, Kogevinas M, et al. Socioeconomic differences in cancer incidence and mortality. In: Kogevinas M, Pearce N, Susser M, Boffetta P, eds. Social Inequalities and Cancer: IARC Scientific Publications, 1997:65-176.

7. Geronimus AT. Poverty, time and place: variation in excess mortality acreoss selected US populations, 1980-1990. Journal of Epidemiology and Community Health 1999;53:325-334.

8. Geronimus AT, Colen CG, Shochet T, et al. Urban-rural differences in excess mortality among high-poverty populations: evidence from the Harlem Household Survey and the Pitt County, North Carolina Study of African American Health. Journal of Health Care for the Poor & Underserved 2006;17:532-58.

9. Muller A. Association between income inequality and mortality among US States: considering population at risk.[comment]. American Journal of Public Health 2006;96:590-1.

10. Singh GK, Hiatt RA. Trends and disparities in socioeconomic and behavioural characteristics, life expectancy, and cause-specific mortality of native-born and foreign-born populations in the United States, 1979-2003.[see comment]. International Journal of Epidemiology 2006;35:903-19.

11. Singh GK, Miller BA, Hankey BF, et al. Persistent area socioeconomic disparities in U.S. incidence of cervical cancer, mortality, stage, and survival, 1975-2000. Cancer 2004;101:1051-7.

12. Vinnakota S, Lam NSN. Socioeconomic inequality of cancer mortality in the United States: a spatial data mining approach. International Journal of Health Geographics [Electronic Resource] 2006;5:9.