

RACE, ETHNICITY AND GEOGRAPHIC VARIATION IN COLORECTAL CANCER

MORTALITY AMONG THE U.S. POPULATION 2005-2007

By

Clementine Aubry-Blanchard

Submitted to the

Faculty of the College of Art and Science

of American University

in Partial Fulfillment of

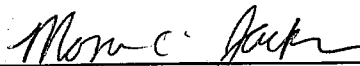
the Requirements for the Degree

of Master of Science

In

Statistics

Chair:



Monica Jackson, Ph.D.



Inga Maslova, Ph.D.



Dean of the College of Art and Science

April 25, 2011

Date

2011

American University
Washington, D.C. 20016

© COPYRIGHT

by

Clementine Aubry-Blanchard

2011

ALL RIGHTS RESERVED

DEDICATION

I would like to dedicate this thesis to my father, who taught me to never worry and never give up, even in the most difficult of situations. And to my grandfather who always believed in me.

RACE, ETHNICITY AND GEOGRAPHIC VARIATION IN COLORECTAL CANCER
MORTALITY AMONG THE U.S. POPULATION 2005-2007

BY

Clementine Aubry-Blanchard

ABSTRACT

Examining geographic variation in racial/ethnic colorectal cancer mortality can help targeting programs and interventions designed to decrease U.S. disparities in colorectal cancer mortality. In this study we apply different spatial methods designed to determine where colorectal cancer (CRC) mortality rates substantially differ from same race/ethnic group's national rates. Using data from a population base, statewide death registry; we examine a cohort of 108 millions men and women from all 50 states and the District of Columbia and 3,143 counties in the United States, whose death was related to colorectal cancer from 2005 to 2007. Mortality rates were adjusted for significant prognostic factors (population size per county, age and race/ethnicity) and evaluated using spatial smoothing and clustering to identify places where CRC death rate in different race/ethnics group was significantly higher or lesser than the statewide rates.

ACKNOWLEDGMENTS

I would like to acknowledge and thank my advisor, Dr. Monica Jackson, for her guidance and dedication to this thesis. I would also like to thank my committee member, Dr. Inga Maslova for her support and wisdom. I would like to give special thanks to Dr. Vickie Shavers at the National Institute of Health and Dr. Ilene B. Rosen at Center for Disease Control, for their contributions to this thesis. Finally, I want to extend my deepest gratitude to August D. Dabney, for his extraordinary advices and guidance, tireless dedication, and boundless emotional support. I am eternally grateful.

TABLE OF CONTENTS

ABSTRACT.....	ii
ACKNOWLEDGMENTS	iii
LIST OF TABLES.....	v
LIST OF ILLUSTRATIONS.....	vi
Chapter	
1. INTRODUCTION	1
Overview of Colorectal Cancer	1
Research Problem	2
2. REASEARCH DESIGN AND METHODS	5
Overview of Study Design.....	5
Data Source.....	6
Descriptive Statistics.....	11
Geographical Analysis	14
3. RESULTS	29
REFERENCES	31

LIST OF TABLES

Table

1.	International Classification of Disease for Colorectal Cancer.....	7
2	File Specification for the Mortality File	8
3	File Specification for the Population File	10
4.	National Population Estimate According to Race and Year , 45+: United States, 2005-2007	12
5.	Number of Death by Race and Year , 45+: United States, 2005-2007.....	13
6.	Cluster Detection Indexes	29

LIST OF ILLUSTRATIONS

Figure

1. U.S. Caucasian Non-Smoothed Colorectal Cancer Mortality Rate by County from 2005-2007. 15
2. U.S. Caucasian Non-Smoothed Colorectal Cancer Mortality Rate Difference Between County Rates and U.S. Rates from 2005-2007. 16
3. U.S. White Non-Hispanic Smoothed Colorectal Cancer Mortality Rate Difference Between County Rates and U.S. Rates from 2005-2007. 22
4. U.S. African American Non-Hispanic Smoothed Colorectal Cancer Mortality Rate Difference Between County Rates and U.S. Rates from 2005-2007..... 23
5. U.S. Hispanic Smoothed Colorectal Cancer Mortality Rate Difference Between County Rates and U.S. Rates from 2005-2007. 24
6. U.S. Smoothed Colorectal Cancer Mortality Rate Disparities between African American Non-Hispanic and White Non-Hispanic from 2005-2007. 25
7. U.S. Smoothed Colorectal Cancer Mortality Rate Disparities between Hispanic and White Non-Hispanic from 2005-2007..... 26

CHAPTER 1

INTRODUCTION

Overview of Colorectal Cancer

Each year more than 50,000 Americans die from colorectal cancer (CRC). Many of these deaths may have been prevented through guideline consistent use of colorectal screening. Epidemiology, genetic, and experimental studies have suggested that colorectal cancer results from complex interactions between predisposition and environmental factors. Previous research has shown that a higher risk of colorectal cancer may be related to obesity, excessive dietary fat consumption, high meat and calcium intake, use of postmenopausal female hormone supplements, alcohol consumption, lack of physical activity, and cigarette smoking in addition to older age, male gender, and black race. (16-18) However, these factors may not fully explain the racial discrepancy associated with colorectal cancer.

The failure to account for geographic variation is posited to overestimate the role of race in health disparities (Baicker, Chandra and Skinner 2005). Data from the Surveillance and Epidemiology Registry (SEER) has shown that there are variation in colorectal cancer mortality rate with respect to racial/ethnic, access to healthcare and utilization of colorectal screening (Altekruse SF 2010) and (Schenck, Klabunde and Davis 2006).

There are several reasons to believe geography plays an important role in CRC mortality rate; First of all it is associated with access to healthcare and utilization of colorectal screening (Turner and Rawlings 2009). Subsequently, geography is linked to the health care delivery environment as well as the local medical culture (Ricketts n.d.). Finally, the neighborhood is a significant source of several types of environmental exposures. The concentrations of racial/ethnic minorities within specific geographic areas therefore have the potential to play a significant role in racial/ethnic health disparities. Unfortunately other published studies who have investigated geographic variation in colorectal cancer outcomes, have either focused on colorectal cancer incidence rate (Lai, et al. 2006) 5-year survival or covered only small areas and/or single years (Henry, Niu and Boscoe 2009). In a recent study, (Shavers, Jackson and Sheppard 2010) investigators found regional differences in colorectal cancer screening by race/ethnicity. This in conjunction with observed reductions in colorectal cancer mortality attributed to increased use of screening suggests a role for geography in racial/ethnic colorectal cancer mortality disparities. Henry et al, found an association between neighborhood income and race with high income predominantly white neighborhoods having the best colorectal cancer survival and low income racially diverse neighborhoods experiencing the worse colorectal cancer survival (Henry, Niu and Boscoe 2009).

Research Problem

Drawing from social epidemiology, medical sociology and medical geography, this study aims to contribute practical public health surveillance information.

Furthermore it discusses the structural pathways involved in CRC mortality given the socio-economic condition that exist in the United-States.

This paper includes a spatial analysis of CRC mortality rate and CRC race specific mortality rate among the United-State's resident during the years 2005-2007. It particularly employs a cluster detection analysis to identify race/ethnic group with higher than expected mortality rate of CRC, as well as disparities of CRC mortality between different race/ethnic groups. Colorectal cancer outcomes are examined based on axes of differences defined by socio-economic status (SES), race and ethnicity. The specific objectives of this study are to:

- Identify geographic area, global cluster, where racial and ethnic disparities in colorectal screening mortality are higher for African American compared to Non Hispanic White and for Hispanic compared to Non Hispanic White.
- Identify areas where race/ethnic specific mortality rates differ from U.S. rates.
- Identify county-level associations between CRC mortality and social and economic variable.

The purpose of this research is to highlight the spatial dimensions of CRC mortality, link it with county level social and economic characteristics, and identify important racial and ethnic disparities among CRC screening rate. Furthermore, in analyzing the spatial and structural patterns of CRC outcomes, this study emphasize the significance of interdisciplinary research in the health field, with the intention of examining how inequalities in health are a product of many processes. These can, directly

or indirectly, influence vulnerabilities people face in disease prevention and health management. It is primordial that these vulnerabilities are identified and mediated at the individual, neighborhood and national level.

CHAPTER 2

RESEARCH DESIGN AND METHODS

Overview of Study Design

This study examine spatial patterns of mortality clusters to identify geographical areas in the United States with statistically significantly higher than expected CRC mortality rates, and measures how socioeconomic status and racial and ethnic affect the CRC outcomes. Based on these analyses, we considered how race/ethnic group, social and economic factor and geographical can influence the CRC outcomes. Alike analyses are important not only to understand how these underlying conditions can end in negative health outcome, but also to be able to provide health agencies with data to allow them to better allocate their resources for CRC screening.

The study collected its data on mortality cases due to CRC in the 50 states and the District of Columbia from the NCHS and then analyzed these cases at the county level using census population data from census year 2005 to 2007. This study employs specific spatial cluster modeling and regression analyses to measure geographical area social and economic influence with CRC mortality outcome. After controlling for difference in population size, age and race/ethnic groups we expected to find significant cluster of higher than expected CRC mortality rate, and cluster of disparities between race/ethnic

group, as well as within one same race/ethnic group. Additionally we expected to find significant relationship with some socioeconomic status factor.

Data Source

National Center for Health Statistics Data

This study primary data source was NCHS, which provided CRC mortality data from all death certificates filed in the 50 States and the District of Columbia during 2005-2007. The National Center for Health Statistics is a division of the United States federal agency the Center for Disease Control and Prevention (CDC). NCHS is under the United States Department of Health and Human Services. Mortality data from death certificates are coded by the states and provided to NCHS through the Vital Statistics Cooperative Program or coded by NCHS from copies of the original death certificate provided to NCHS by the state registration offices. (Statistics 1999) The data set excluded deaths of nonresidents (e.g. deaths of nonresident aliens, nationals residing abroad, and residents of Puerto Rico, the Virgin Islands, Guam, and other territories of the United States) and fetal deaths (Xu, et al. 2010).

The NCHS provided cancer specific mortality data. On September 21, 2010 the office approved the mortality for release for use in this study. The National Association for Public Health Statistics and Information Systems (NAPHSIS) which represents state vital registrars conducted its review of the study prior to the NCHS review and includes both federal and non-federal requests for restricted data files. The views expressed herein are solely ours, and do not necessarily reflect those of the contractor of data.

This study analyzed a total of # of cases non identified CRC death cases. The underlying causes of death were classified by the NCHS in accordance with the tenth revision of the International Statistical Classification of Disease and Related Health Problems. The mortality cases were categorized using ICD-10 codes for all deaths that occurred after 1999. Table 1 list out the specific codes for CRC mortality used in this study.

Table 1. International Classification of Disease for Colorectal Cancer

	ICD-10
Malignant neoplasm of colon	
Malignant neoplasm of cecum	C18.0
Malignant neoplasm of appendix	C18.1
Malignant neoplasm of ascending colon	C18.2
Malignant neoplasm of hepatic flexure	C18.3
Malignant neoplasm of transverse colon	C18.4
Malignant neoplasm of splenic flexure	C18.5
Malignant neoplasm of descending colon	C18.6
Malignant neoplasm of sigmoid colon	C18.7
Malignant neoplasm of overlapping lesion of colon	C18.8
"Malignant neoplasm of colon, unspecified"	C18.9
Malignant neoplasm of rectosigmoid junction	C19
Malignant neoplasm of rectum	C20

Source: National Center for Health Statistics (Compressed Mortality File 1999-2007)

The compressed mortality file also contained demographics and geographic information. The database recorded the person's year of death, race by sex, Hispanic origin, age at death, and cause of death. Geographical location at year of death was coded as FIPS State code and FIPS county code. Race and Hispanic origin are reported

separately on the death certificate in accordance with standards set forth by the Office of Management and Budget (OMB) (Budget 1997). The American Indian or Alaska Native race category includes: North, Central, and South American Indians, Eskimos, and Aleuts. The Asian or Pacific Islander race category includes Chinese, Filipino, Hawaiian, Japanese, and Other Asian or Pacific Islanders. Table 2 lists the file layout provided by the NCHS.

Table 2 File Specification for the Mortality File

Location	Field Size	Item and code outline	Format
		<u>FIPS Codes</u>	
1-2	2	FIPS State code	Numeric
3-5	3	FIPS county code	Numeric
6-9	4	Year of death	Numeric
10	1	<u>Race-Sex</u>	Numeric
		1 White male	
		2 White female	
		3 Black male	
		4 Black female	
		5 American Indian of Alaska Native male	
		6 American Indian of Alaska Native female	
		7 Asian or Pacific Islander male	
		8 Asian or Pacific Islander female	
11	1	<u>Hispanic Origin</u>	Numeric
		1 not Hispanic or Latino	
		2 Hispanic or Latino	
		9 not stated	
12-13	2	<u>Age at Death</u>	Numeric
		01 under 1 day	
		02 1-6 days	
		03 7-27 days	
		04 28-364 days	
		05 1-4 years	

Table 2 Continued

Location	Field Size	Item and code outline	Format
		06 5-9 years	
		07 10-14 years	
		08 15-19 years	
		09 20-24 years	
		10 25-34 years	
		11 35-44 years	
		12 45-54 years	
		13 55-64 years	
		14 65-74 years	
		15 75-84 years	
		16 85+	
		99 Unknown	
14-17	4	ICD-10 code for underlying cause-of-death	Character
18-20	3	113 Cause-of-Death Recode	Numeric
21-24	4	Number of deaths	Numeric

Source: National Center for Health Statistics (Compressed Mortality File 1999-2007)

U.S. Census Population File

The compressed mortality file also provided national, State, and county population estimate. These are bridged-race estimates of the resident population of the United States produced by the U.S. Census Bureau in collaboration with NCHS (N. C. Statistics 2009). The database recorded geographical area, year, race by sex, Hispanic origin, number of live birth, population in thirteen age group, county name, and the type of record type. There are National, State and county estimates recorded; they can be distinguished by using FIPS code and record type. For all years 1999-2007, the national population estimates are derived by summing the county-level estimates, so they are consistent with

both State and county. For the study county and State population estimate were used aggregating for each race/ethnic group and population in age group 45 and up. Table 3 lists the variable in the population file. However, this one did not provide socio-economic variables; these were extracted from the U.S. Census Bureau website. Table 4 list out these variables.

Table 3 File Specification for the Population File

Location	Field Size	Item and code outline	Format
		<u>FIPS Codes</u>	
1-2	2	FIPS State code	Numeric
3-5	3	FIPS county code	Numeric
6-9	4	Year (1999-2007)	Numeric
10	1	<u>Race-Sex</u>	Numeric
		1 White male	
		2 White female	
		3 Black male	
		4 Black female	
		5 American Indian of Alaska Native male	
		6 American Indian of Alaska Native female	
		7 Asian or Pacific Islander male	
		8 Asian or Pacific Islander female	
11	1	<u>Hispanic Origin</u>	Numeric
		1 not Hispanic or Latino	
		2 Hispanic or Latino	
		9 not stated	
12-19	8	Number of live births	Numeric
20-27	8	Population in age group: <1 year	Numeric
28-35	8	Population in age group: 1-4 years	Numeric
36-43	8	Population in age group: 5-9 years	Numeric
44-51	8	Population in age group: 10-14 years	Numeric
52-59	8	Population in age group: 15-19 years	Numeric
60-67	8	Population in age group: 20-24 years	Numeric

Table 3. Continued

Location	Field Size	Item and code outline	Format
68-75	8	Population in age group: 25-34 years	Numeric
76-83	8	Population in age group: 35-44 years	Numeric
84-91	8	Population in age group: 45-54 years	Numeric
92-99	8	Population in age group: 55-64 years	Numeric
100-107	8	Population in age group: 65-74 years	Numeric
108-115	8	Population in age group: 75-84 years	Numeric
116-123	8	Population in age group: 85+ years	Numeric
124-148	25	County name	Character
149	1	<u>Record Type</u>	Numeric
		1 National population record	
		2 State population record	
		3 County population record	

Source: National Center for Health Statistics (Compressed Mortality File 1999-2007)

Descriptive Statistics

For the purpose of the study we combined all male and female, then aggregated deaths counts and population estimate for each county and each race/ethnic group.

Subsequently we divided each death counts by the respective population estimate which resulted in colorectal cancer mortality rate per county and race/ethnic group. National population estimates and number of death are displayed in Table 4 and Table 5 respectively.

Table 4. National Population Estimate According to Race and Year , 45+: United States, 2005-2007

Race	2005	2006	2007	Average
All races	105,266,186	107,619,390	109,952,337	107,612,604
White Non-Hispanic	85,156,766	86,656,304	88,105,193	86,639,421
Black Non-Hispanic	10,980,093	11,307,642	11,634,422	11,307,386
Hispanic	9,129,327	9,655,444	10,212,722	9,665,831

Source: National Center for Health Statistics (Compressed Mortality File 1999-2007)

Table 5. Number of Death by Race and Year , 45+: United States, 2005-2007

Race	2005		2006		2007		Total	
	Counts	Counts per 1000	Counts	Counts per 1000	Counts	Counts per 1000	Counts	Counts per 1000
All races	51,592	0.49	51,836	0.48	51,512	0.47	154,940	1.43
White Non-Hispanic	41,195	0.48	41,242	0.47	41,522	0.47	123,959	1.43
Black Non-Hispanic	6,630	0.60	6,612	0.58	6,682	0.57	19,924	1.76
Hispanic	2,461	0.26	2,652	0.27	2,693	0.26	7806	0.81

Source: National Center for Health Statistics (Compressed Mortality File 1999-2007)

Geographical Analysis

Visualizing Spatial Data

The typical goal of mapping public health data is to provide insight into geographic variation in disease risk, which is the probability that an unfortunate event occurs. In public health, the unfortunate event is usually the contraction of or death from a specific disease. When we make a map a disease counts, proportions, or rates, we are ultimately intending to convey inferences about disease risk. However, a map of raw counts is not the best tool for inference about disease risk, since we expect regions with larger populations to have higher disease counts. We can account for population differences by using rates (disease incidence per person per time) as measures of risk. Higher disease rates reflect greater chances for contracting the disease, and thus, viewed this way, rates reflect a person's risk for disease. For the purpose of the study we would like to find areas (1) where there are racial disparities in colorectal cancer mortality and (2) where race specific mortality rates differ from U.S. rates. For this reason two groups of maps will be necessary to geographical variation, the first group will try to assess where racial disparities exist between African American (AA) and Non Hispanic White (NHW) and between Hispanic and NHW. And the second group will show where CRC mortality rates differ from same race national rate for all four race/ethnic group as well as Hispanic and non Hispanic. Consequently maps of each race/ethnic group's CRC mortality rate will be produce in order to find areas with spatial clustering.

However, a map of rates may still obscure the spatial pattern in disease risk, particularly if the rates are based on populations of very different size. Since the variability in the estimated local rates depends on population size, some rates may be better estimated than others, and this may obscure the spatial pattern. Rates based on small populations or on small numbers of death cases are likely to be elevated artificially, reflecting a lack of data rather than true elevated mortality risk. As an example, let us consider a large area with 2 million residents and 2 observed deaths. Suppose that the area of study consist of sub regions each containing 100 residents. For the entire area the crude mortality rate is $1/1,000,000$, while two of the sub regions contain a crude rate of $1/100$. This is referred to as the small number problem (Waller and Gotway 2004). The maps below illustrate that problem.

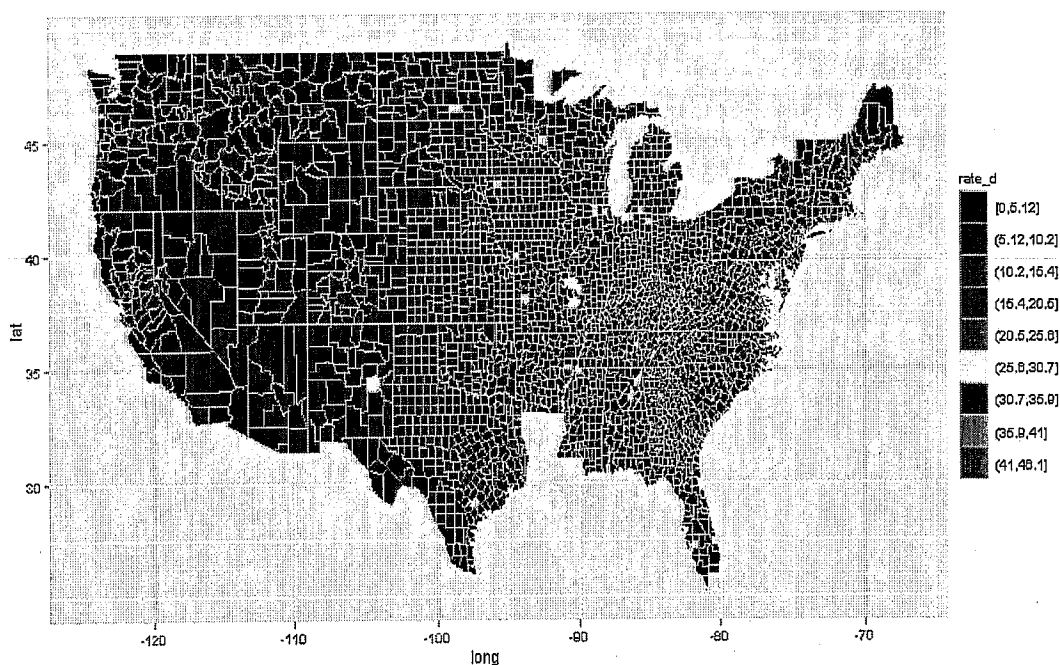


Figure 1 U.S. Caucasian Non-Smoothed Colorectal Cancer Mortality Rate by County from 2005-2007.

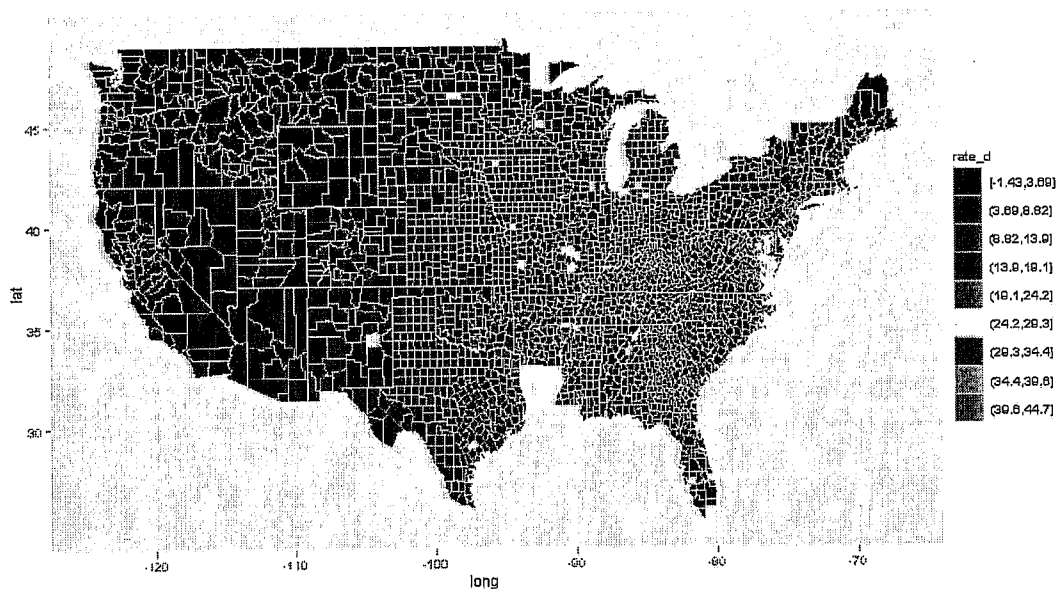


Figure 2 U.S. Caucasian Non-Smoothed Colorectal Cancer Mortality Rate Difference Between County Rates and U.S. Rates from 2005-2007.

Mapping Smoothed Rates

One solution to this problem is spatial smoothing, a method that not only allows the researcher to stabilize rates based on small numbers by combining available data at the resolution of interest, but also allows to reduce noise in the rates caused by different population sizes, thus increasing our ability to discern systematic patterns in the spatial variation of the underlying risk. The idea is to borrow information from neighboring regions to produce a more stable and less noisy estimate of the rate associated with each region and thus separate out the spatial pattern from the noise.

There are several approaches to spatial smoothing. One “borrows” information from nearby regions to stabilize local estimates through the use of various weighting schemes. Another, more formal approach uses probability models to obtain smoothed estimates consisting of a compromise between the observed rate for each region and an estimate from a larger collection of cases and persons at risk. The compromise combines the rate from each region, which can be statistically unstable due to the rarity of the disease and the relatively small number of people at risk.

Empirical Bayes Smoothing

Clayton and Kaldor (1987) propose a Bayesian approach to this problem which defines the analytic form of the compromise estimator. Bayesian statistics in general treats all unknown model parameters as random variables, and the goal of inference is to define the distributions of the variables, thereby providing point and interval estimates, predictions, and probability calculations. Bayesian inference depends on the prior distribution. The use of a “non-informative” prior distribution results in a posterior distribution very similar to the likelihood function, while an overly “informative” prior may result in a posterior far removed from the likelihood function. We illustrate how incorporation of particular prior distributions achieves our goal to borrow information from other areas to stabilize local rates while maintaining a very sensible probabilistic structure to the problem (Clayton and Kaldor 1987).

For starters, we build a probability model describing our data. Assume that the death counts Y_i represent random variables, each following a Poisson distribution with

mean equal to $n_i \xi_i$, where ξ_i denotes the risk of a person residing in region i dying from the disease during the study period. Given this local probability of disease, we have

$$Y_i | \xi_i \stackrel{ind}{\sim} \text{Poisson}(n_i \xi_i), \quad (1)$$

Under this model, we are assuming that the Y_i are conditionally independent given the ξ_i . This does not mean that the Y_i are mutually independent; rather, this implies that any spatial correlation observed in the Y_i is a function of spatial trends in either the population sizes n_i (considered to be fixed, known quantities) or in the local individual risks ξ_i for $i = 1, \dots, N$.

In a Bayesian analysis, the likelihood function is defined by the conditional distribution defined in equation (1). Since the Y_i are conditionally independent given the ξ_i parameters, the likelihood takes a particularly simple form and is defined as the product, across all regions $i = 1, \dots, N$, of the conditional distributions given in equation (1).

The ξ_i is treated as random variables by the Bayesian analysis, and we next define a prior distribution for each ξ_i . To start, we denote the prior mean by $E_\xi(\xi_i) = m_{\xi_i}$ and the prior variance by $Var_\xi(\xi_i) = v_{\xi_i}$ (Marshall 1991). The mean and variance of the observed local count, Y_i , conditional on the value of ξ_i , as $n_i \xi_i$ is given by equation (1). Thus, the conditional mean and variance of the local rate observed, r_i , of the mortality are

$$E(r_i | \xi_i) = E[Y_i | n_i] | \xi_i = \xi_i$$

and

$$Var(r_i | \xi_i) = Var[Y_i | n_i] | \xi_i = \xi_i / n_i,$$

respectively.

In order to find the unconditional mean of the rate observed in region i , r_i , we need to take the expectation over ξ_i of the conditional expectation:

$$E_r(r_i) = E_\xi E(r_i | \xi_i) = E_\xi(\xi_i) = m_{\xi_i}$$

where E_r and E_ξ represent the expectation with respect to the marginal distributions of r and ξ , respectively. The unconditional variance of r_i equals the sum of the variance of the conditional mean with respect to ξ_i and the expectation of the conditional variance:

$$Var_r(r_i) = Var_\xi(\xi_i) + E_\xi \left(\frac{\xi_i}{n_i} \right) = v_{\xi_i} + \frac{m_{\xi_i}}{n_i}.$$

Deriving the best linear Bayes estimator of ξ_i by minimizing the expected total squared-error loss yields (Marshall 1991).

$$\begin{aligned} \hat{\xi}_i &= m_{\xi_i} + C_i(r_i - m_{\xi_i}) \\ &= C_i r_i + (1 - C_i)m_{\xi_i}, \quad (2) \end{aligned}$$

where $C_i = \frac{v_{\xi_i}}{v_{\xi_i} + \frac{m_{\xi_i}}{n_i}}$ is the ratio of the prior variance to the data variance. This ratio is called the shrinkage factor since it defines how much the crude rate, $r_i = Y_i/n_i$, “shrinks” toward the prior mean. Note that the estimator defined in equation (2) corresponds to a weighted average of the crude estimate and the prior mean. When the population size, n_i , is small, $C_i \rightarrow 0$, and the Bayes estimator is close to the prior mean, m_{ξ_i} . However, when the expected count is large, $C_i \rightarrow 1$, and the Bayes estimator approaches the rate

observed. In short, the Bayes estimator provides an approach that borrows strength from the prior mean, where the amount of strength borrowed depends on the stability of the crude local estimate as measured by the prior variance.

To compute the estimates, we require values for m_{ξ_i} and v_{ξ_i} . In empirical Bayes estimation, the unknown parameters are estimated from the data. In this case, following (Marshall 1991) we assume that $m_{\xi_i} \equiv m_{\xi}$, and $v_{\xi_i} \equiv v_{\xi}$, or the same prior mean and variance for all regions) since the model is otherwise over specified, (i.e. there is more unknown parameters than there are data values). Following this assumption Marshall (1991) uses the method of moments to estimate m_{ξ} , v_{ξ} , and C_i . The method-of-moments estimator of the overall mean, m_{ξ} , is simply the weighted sample mean,

$$\widetilde{m}_{\xi} = \frac{\sum_{i=1}^N r_i n_i}{\sum_{i=1}^N n_i}. \quad (3)$$

The weighted sample variance is

$$s^2 = \left(\sum_{i=1}^N (r_i - \widetilde{m}_{\xi})^2 \right) / \left(\sum_{i=1}^N n_i \right)$$

and has expected value (ignoring estimation of m_{ξ}) $v_{\xi} + m/\bar{n}$, where $\bar{n} = \sum_{i=1}^N n_i/N$.

Thus, the method-of-moments estimator v_{ξ} is

$$\widetilde{v}_{\xi} = s^2 - \left(\frac{\widetilde{m}_{\xi}}{\bar{n}} \right). \quad (4)$$

In order to avoid negative variance estimates, we will use zero as our estimates of v_{ξ} if this quantity is negative. Substituting \widetilde{m}_{ξ} and \widetilde{v}_{ξ} from equation (3) and (4) respectively,

into the expression for C_i gives the method-of-moments estimator of the Bayes shrinkage factor as

$$\tilde{C}_i = \begin{cases} \frac{s^2 - \tilde{m}_\xi / \bar{n}}{s^2 - \tilde{m}_\xi / \bar{n} + \tilde{m}_\xi / n_i} & \text{if } s^2 \geq \tilde{m}_\xi / \bar{n} \end{cases}$$

Substituting these values into equation (2) gives us the empirical Bayes estimator

$$\hat{\xi}_i = \tilde{m}_\xi + \tilde{C}_i(r_i - \tilde{m}_\xi). \quad (5)$$

After smoothing the data using the empirical Bayes estimator we were able to start mapping the mortality rates by county and start observing patterns and clusters. Below are, for each race, a map of the difference between the race/ethnic specific mortality rates from U.S. rates.

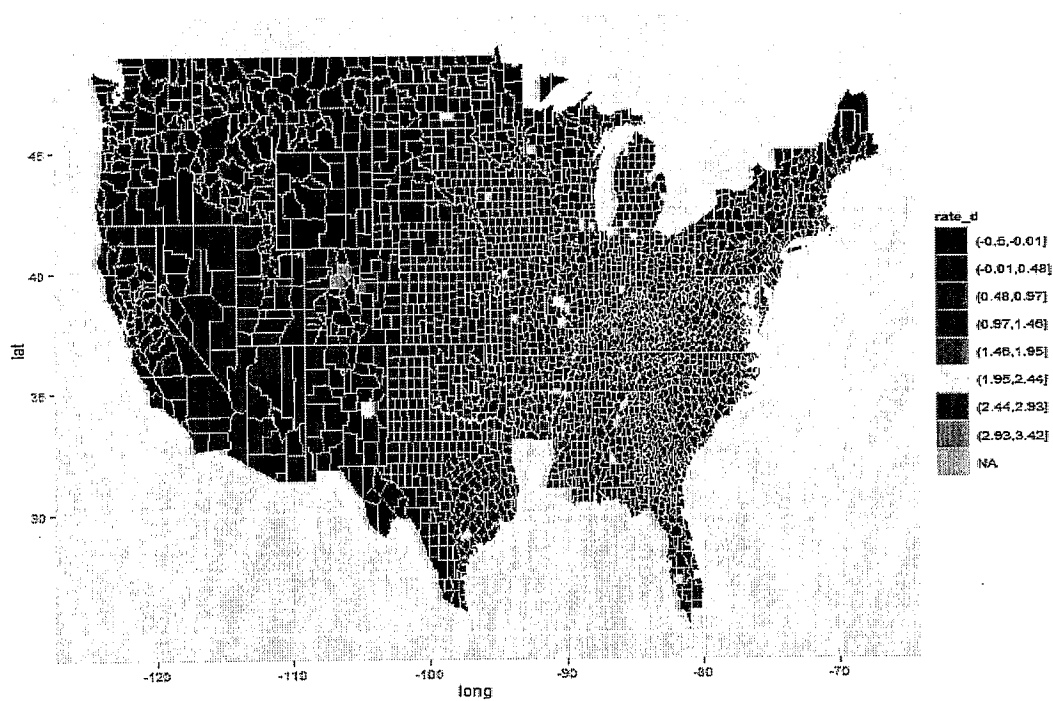


Figure 3. U.S. White Non-Hispanic Smoothed Colorectal Cancer Mortality Rate Difference Between County Rates and U.S. Rates from 2005-2007.

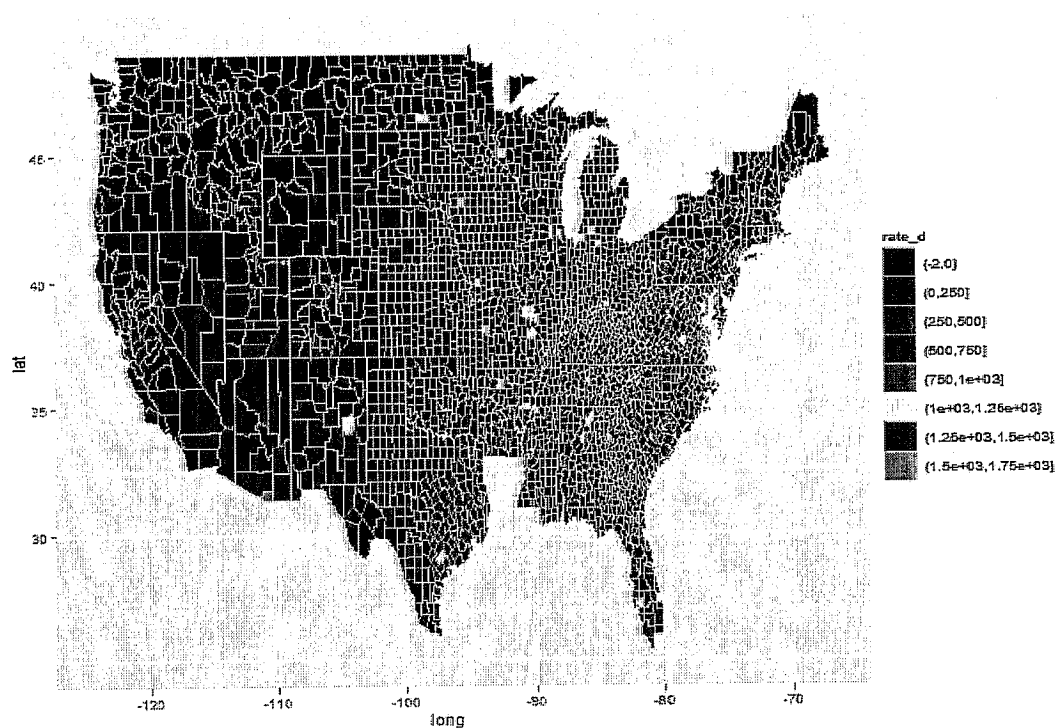


Figure 4. U.S. African American Non-Hispanic Smoothed Colorectal Cancer Mortality Rate Difference Between County Rates and U.S. Rates from 2005-2007.

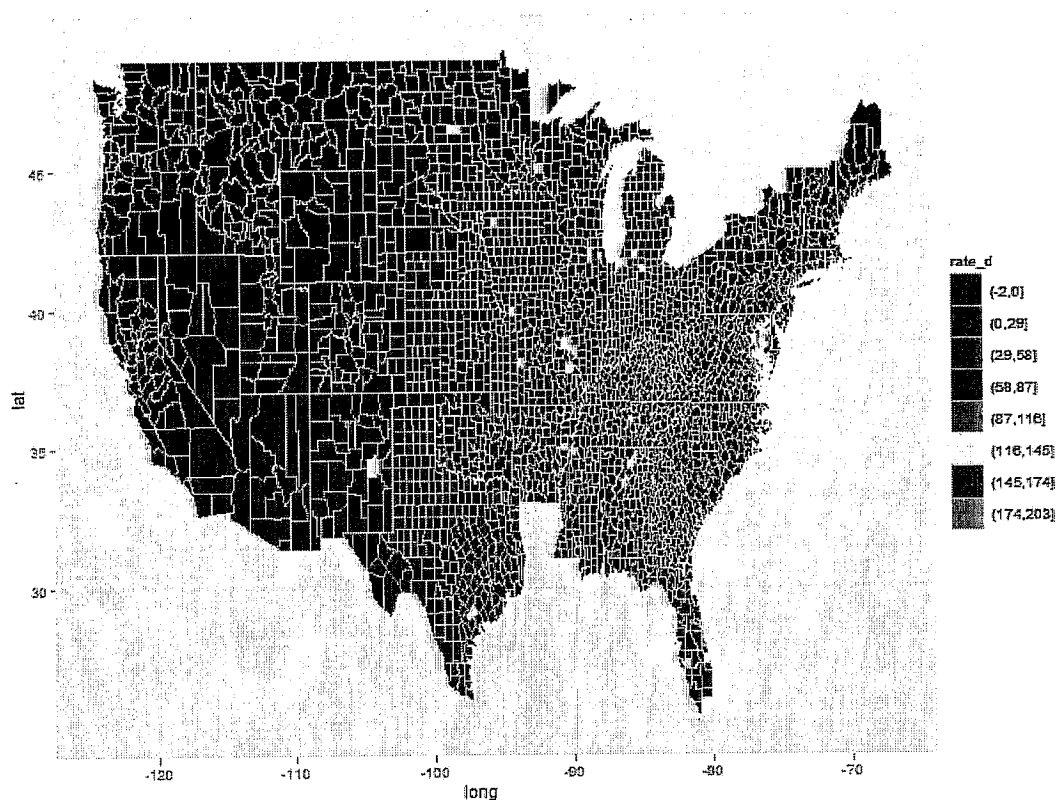


Figure 5. U.S. Hispanic Smoothed Colorectal Cancer Mortality Rate Difference Between County Rates and U.S. Rates from 2005-2007.

In order to allow for a visual of the differences. All the county who were falling below the U.S average were section in one interval, the counties above the U.S. average were divided into equal interval. We can observe a cluster in regions where the population of that specific rate is above the U.S National average.

Below is a map of the difference in rates between African American Non Hispanic and White Non-Hispanic, and finally a map of the difference between Hispanic

and White Non-Hispanic. For similar reasons as above, the rate who were falling below the White Non-Hispanic rates were put in their own interval, and the rates above were divided into equal intervals. We can observe similarly in the South where African American Non-Hispanic population is above average, the CRC mortality rate is above the one of White Non-Hispanic race. Similarly in the South West regions Hispanic population is above the U.S. average and therefore shows sign of a CRC mortality rate above the one of White Non-Hispanic. Aside from a population disparities, this may also be explain by other socio-economic factors.

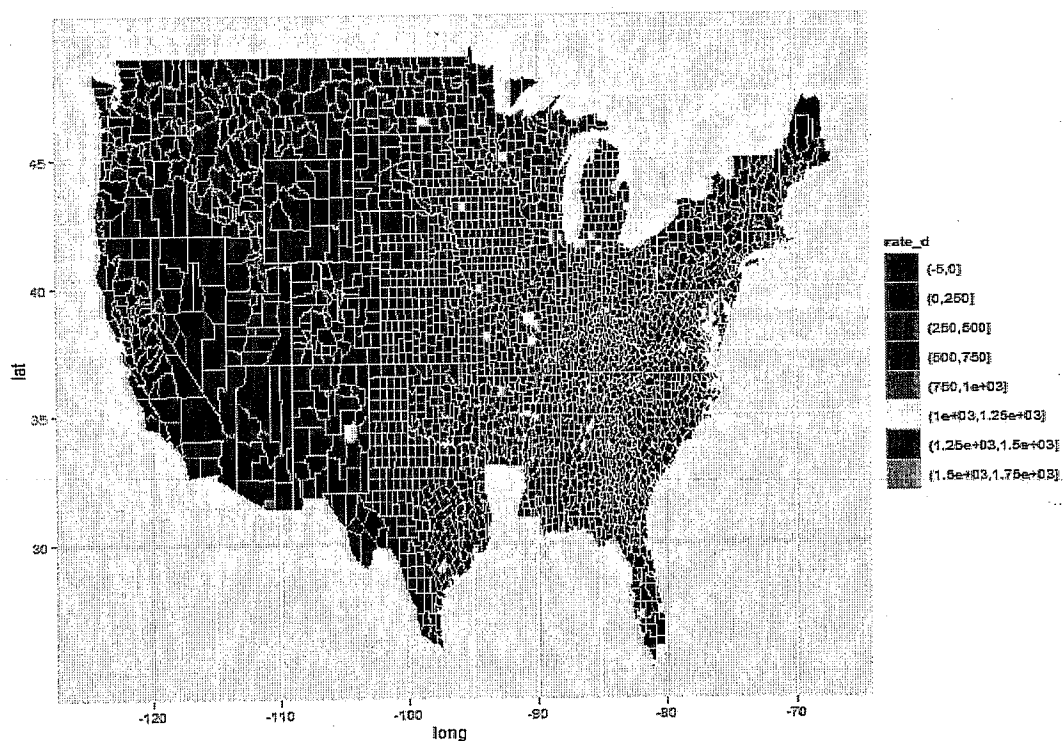


Figure 6 U.S. Smoothed Colorectal Cancer Mortality Rate Disparities between African American Non-Hispanic and White Non-Hispanic from 2005-2007.

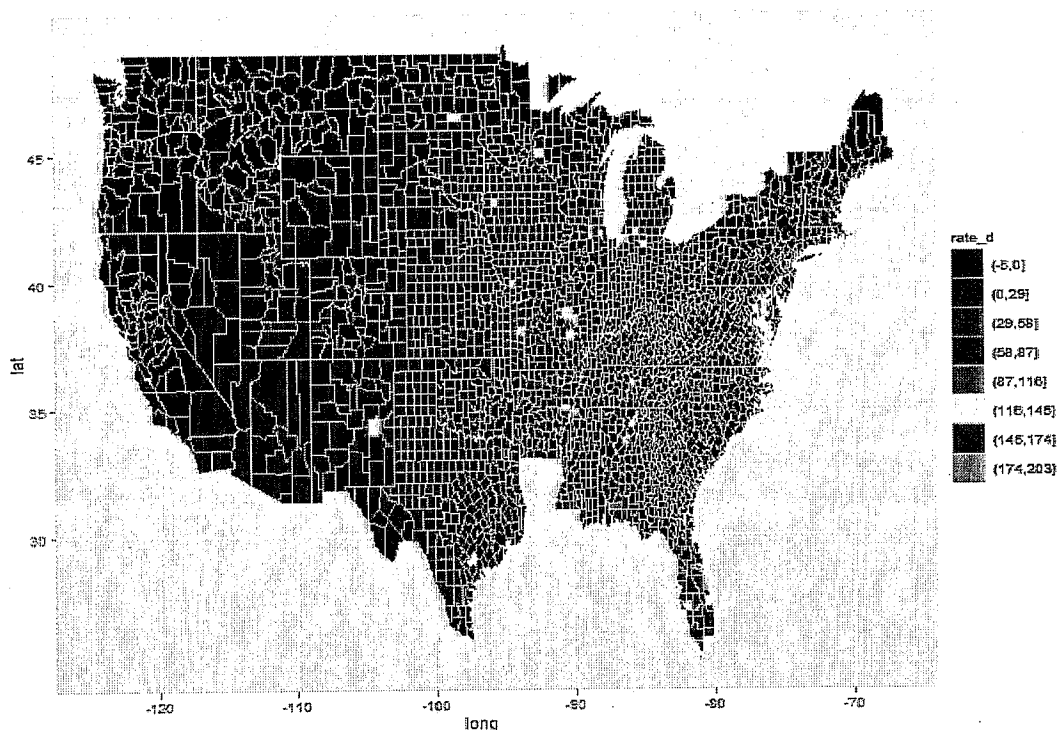


Figure 7 U.S. Smoothed Colorectal Cancer Mortality Rate Disparities between Hispanic and White Non-Hispanic from 2005-2007.

Clustering Detection

Moran's I

The first index we consider is Moran's I. Moran's I is largely used, and variations of it relate to likelihood ratio tests and best invariant tests for particular models of correlation for normally distributed random variables. Moran's I follows the basic from global indexes of spatial autocorrelation with similarity between regions i and j defined as the product of the respective difference between Y_i and Y_j with the overall mean:

$$sim_{ij} = (Y_i - \bar{Y})(Y_j - \bar{Y})$$

where $\bar{Y} = \sum_{i=1}^N Y_i / N$. In addition, we divide this basic form by the sample variance observed in the Y_i 's, yielding

$$I = \left(\frac{1}{s^2} \right) \frac{\sum_{i=1}^N \sum_{j=1}^N w_{ij} (Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i=1}^N \sum_{j=1}^N w_{ij}} \quad (6)$$

where

$$s^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2$$

Thus, I is a random variable having a distribution defined by the distributions of and interactions between the Y_i . We obtain the value of I observed by inserting observations into equation (6). When neighboring regions tend to have similar values, I will be positive. If neighboring regions tend to have different values, I will be negative. When there is no correlation between neighboring values, the expected value of I is

$$E(I) = -\frac{1}{N-1} \quad (7)$$

approaching zero as N increases. Unlike a traditional correlation coefficient, values for Moran's I does not have to be restricted to the interval [-1,1].

Tango's Index

Tango (1984) introduces an index of disease clustering in time using interval count data, based on equal length time intervals subdividing the entire study period with event

counts observed for each interval. Later, he generalizes the index for applications involving unequal time intervals and/or interval-specific covariates. Finally, Tango (1995) recasts the generalized statistics in a spatial setting. This version of Tango's index is represented here.

First rather than the set of regional counts $Y_1 \dots Y_N$, we consider the set of regional proportions,

$$\left(\frac{Y_1}{Y_+}, \dots, \frac{Y_N}{Y_+}\right)$$

where $Y_+ = \sum_{i=1}^N Y_i$, the total number of observed cases. Next we obtain the vector of expected proportions under the constant risk null hypothesis, namely the vector of population proportions

$$\left(\frac{n_1}{n_+}, \dots, \frac{n_N}{n_+}\right)$$

Where $n_+ = \sum_{i=1}^N n_i$ denotes the total population at risk. Note that Tango's Index assumes that both Y_+ and n_+ are known. When we condition a set of independent Poisson counts on their total, the distribution of the set of counts follow a multinomial distribution. Thus, under the constant risk hypothesis, the population proportions provide the expected cell probabilities for a multinomial distribution. Tango's index compares the case proportions observed to those expected under the constant risk null hypothesis.

We define Tango's index as

$$T_{ti} = \sum_{i=1}^N \sum_{j=1}^N w_{ij} \left(\frac{Y_i}{Y_+} - \frac{n_i}{n_+}\right) \left(\frac{Y_j}{Y_+} - \frac{n_j}{n_+}\right) \quad (8)$$

CHAPTER 3

RESULTS

After we applied both these index on our data for each of our race/ethnic group, we find evidence of possible cluster. The table below describe the findings.

Table 6. Cluster Detection Indexes

Race	Moran's I	Tango's Index	P-value
White Non-Hispanic	0.2561	4.227e-06	0
Black Non-Hispanic	0.2920	2.744e-05	0
Hispanic	0.3316	3.441e-04	0

The Moran's I index being positive lead us to believe the pattern is clustered. This backs up our findings from the map, as we saw more significant cluster in the Hispanic map. All three Moran's I index are greater than 0 which tell us there exist clusters. The Tango's Index are also positive however relatively small which contradict our findings

with Moran's I . This may tell us that Moran's I is a better index to use when attempting to detect clusters among so many regions.

REFERENCES

- Altekruse SF, Kosary CL, Krapcho M, Neyman N, Aminou R, Waldron W, Ruhl J, Howlander N, Tatalovich Z, Cho H, Mariotto A, Eisner MP, Lewis DR, Cronin K, Chen HS, Feuer EJ, Stinchcomb DG, Edwards BK. "SEER Cancer Statistics Review 1975-2007." *National Cancer Institute*. 2010.
http://seer.cancer.gov/crs/1975_2007/.
- Baicker, Katherine, Amitabh Chandra, and Jonathan S Skinner. "Geographic variation in health care and the problem of measuring racial disparities." *Perspective in Biology and Medicine* 48, no. 1 (2005): S42-S53.
- Budget, Office of Management and. "Revisions to the standards for the classification of Federal data on race and ethnicity. ." *Federal Register* 62FR58781-58790. October 30, 1997.
- Clayton, David, and John Kaldor. "Empirical Bayes Estimates of age-standardized relative risks for use in disease mapping." *Biometrics* 43 (1987): 671-682.
- Henry, Kevin A, Xiaoling Niu, and Francis P Boscoe. "Geographic disparities in colorectal cancer survival." *International Journal Health Geographic* 8, no. 1 (2009): 48.
- Lai, Sue-Min, Kevin B Zhang, Robert J Uhler, Jovanka N Harrison, Gayle G Clutter, and Melanie A Williams. "Geographic variation in the incidence of colorectal cancer in the United-States, 1998-2001." *Cancer* 107, no. 5 (2006): 1172-1180.
- Marshall, Roger J. "Mapping disease and mortality rates using empirical Bayes estimators." *Applied Statistics* 40, no. 2 (1991): 283-294.
- National Center for Health Statistics. Compressed Mortality File. 1999-2007 (machine readable data file and documentation, CD-ROM Series 20, No. 2M). Hyattsville, Maryland. 2010.
- Ricketts, Thomas C. "Geography and health disparity in the United-States." *Woodrow Wilson International Center*.
<http://www.wilsoncenter.org/topics/docs/Ricketts.pdf> (accessed February 22, 2011).
- Schenck, Anna P, Carrie N Klabunde, and William W Davis. "Racial differences in colorectal cancer test use by Medicare Consumers." *Am J Prev Med* 30, no. 4 (2006): 320-326.

Shavers, Vivkie L, Monica C Jackson, and Vanessa B Sheppard. "Racial/Ethnic patterns of uptake of colorectal screening, National Health Interview Survey 2000-2008." *Journal of the national medical association* 102, no. 7 (2010): 621-635.

Statistics, National Center for Health and. "Technical appendix." *Vital statistics of the United States: Mortality*. 1999. <http://www.cdc.gov/nchs/data/statab/techap99.pdf>.
 Statistics, National Center for Health. "Postcensal estimates of the resident population of the United States for July 1, 2000-July 1, 2008, by year, county, age, bridge race, Hispanic origin, and Sex." May 14, 2009.
<http://www.cdc.gov/nchs/about/major/dvs/popbridge/popbridge.htm>.

Turner, Margery Austin, and Lynette A Rawlings. "Promoting neighborhood diversity: benefits, barriers and strategies." *Urban Institute*. August 1, 2009.
<http://www.urban.org/url.cfm?ID=411955> (accessed February 22, 2011).

Waller, Lance A, and Carol A Gotway. *Applied Spatial Statistics for public health data*. John Wiley & Sons, 2004.

Xu, Jiaquan, Kenneth Kochanek, Sherry L Murphy, and Betzaida Tejada-Vera. "National Vital Statistics Reports." *Deaths: Final Data for 2007*. May 20, 2010.
http://www.cdc.gov/nchs/data/nvsr/nvsr58/nvsr58_19.pdf.