2018

The Importance of Person and Place in Predicting Prostate Cancer Incidence and Mortality among United States Veterans Seeking Veterans Health Administration Care

Peter Georgantopoulos
University of South Carolina

The Importance of Person and Place in Predicting Prostate Cancer Incidence and Mortality among United States Veterans Seeking Veterans Health Administration Care

by

Peter Georgantopoulos

Bachelor of Science
University of California, Irvine, 2000

Master of Public Health
Saint Louis University, 2004

Master of Art
Washington University in St. Louis, 2010

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Epidemiology

The Norman J. Arnold School of Public Health

University of South Carolina

2018

Accepted by:

James R. Hébert, Major Professor

Jan M. Eberth, Committee Member

Bo Cai, Committee Member

Gowtham Rao, Committee Member

Charles L. Bennett, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

DEDICATION

To my parents Demetrios and Effie Georgantopoulos who have provided and continue to provide unconditional love, guidance, and support for me and my siblings.

ACKNOWLEDGEMENTS

The completion of this research was the hardest professional endeavor that I have ever undertaken. The achievement of this task was only made possible by the wonderful mentorship of my committee, the academic and intellectual support of the Epidemiology and Biostatistics faculty and staff at the Arnold School of Public Health at the University of South Carolina, and the unwavering love and support from my family and friends.

ABSTRACT

Introduction: There are several unique characteristics in the epidemiology of prostate cancer (PrCA) that make it an interesting and important cancer to study. The first is that while prostate cancer is the most common cancer that men develop, it is one of the least common cancers that men die from. This indolent nature of PrCA has led to the idiom among health scientists that "men are more likely to die *with* PrCA than *due* to PrCA". Just like other cancers, several individual-level risk factors (e.g., family history of the disease, age, and race) are well established for both PrCA incidence and mortality.

It is becoming more common for health scientists to utilize innovative modeling techniques to describe the current epidemiologic characteristics of PrCA given its unique etiological nature and established individual-level risk factors. One such avenue of research is to understand how geography impacts the epidemiology of PrCA. The vast majority of studies incorporating a geographical component will model an area-level characteristic(s) and nest PrCA subjects within a geographical area where that area is treated as a random component (either as a random intercept or random effect). Yet another, more recent, approach is to account for the spatial autocorrelation of the geographical area of interest.

Modeling individual-level geographical components (e.g., distance each subject travels to a healthcare facility), area-level geographical components (area-level risk factors and geographical area as a random effect) and the spatial autocorrelation of the

geographical area of interest within a succinct modeling framework. It is this gap in knowledge in simultaneously accounting for multiple geographical levels and components in PrCA epidemiology that this research was undertaken.

Methods: To evaluate how geographic context impacts the epidemiology of PrCA, this research was divided into three parts. The first investigated the incidence of PrCA, the second the mortality, and the third the mortality-to-incidence ratio (MIR).

Electronic medical records from the United States (US) Veteran's Health Administration (VHA) served as the data sources for all individual-level information and aggregate demographic information. The US 2010 Decennial Census and 2007-2011 5-Year American Community Survey (ACS) were used for all area-level information. The Social Vulnerability Indexes (SoVI®) was obtained from the University of South Carolina Hazards and Vulnerability Institute. The 2004 Rural-Urban Commuting Area Codes (RUCA) were obtained from the University of Washington. The geographical area of interest were all ZIP code tabulated area linked (ZCTA) ZIP codes in the state of South Carolina (SC) during the timeframe January 1, 1999 to December 31, 2015.

Results: It was found that PrCA incidence among SC veterans who receive care at VHA facilities from 1999 – 2015 were at 35% and 39% increased risk if they reside in ZCTA-linked ZIP codes with SoVI® scores that were medium and high, respectively, as compared to those veterans residing in ZIP codes with low SoVI® scores. Also, the best-fitting model for PrCA incidence accounted for the random effect of SC ZCTA-linked ZIP codes and accounted for the spatial autocorrelation between SC ZCTA-linked ZIP codes.

It was found that PrCA mortality among SC veterans who received are at VHA facilities from 1999 – 2015 were at 25% increased risk of death if they traveled more than 55 miles to receive care. Also, the best-fitting model for PrCA mortality accounted for the random effect of SC ZCTA-linked ZIP codes.

It was found that the PrCA mortality-to-incidence ratio (MIR) was 0.17 for SC veterans who receive are at VHA facilities from 1999 – 2015. Stratified MIRs for standard risk factors for PrCA were found to vary by racial category, age at first VA visit, ZCTA-linked ZIP code level social vulnerability and ZIP code rurality. Collectively, it was found that MIRs by ZCTA-linked ZIP codes did not vary significantly in SC; however, two distinct clusters of ZCTA-linked ZIP code MIRs was found.

Discussion: This research found that the epidemiology of PrCA among veterans who received care at VHA facilities from 1999 – 2015 varies by geographic location and context. This research has successfully used three distinct measures to characterize where a veteran resides: 1) linear distance to his most frequented VHA facility, 2) neighborhood characteristics of the ZIP code he resides in, and 3) the spatial autocorrelation of the ZIP code he resides in.

The successful demonstration of the approach undertaken in this research has the potential to be replicated and expanded to use different geographical boundaries (e.g. counties), incorporate temporal factors, and evaluate other cancers and chronic diseases. This research approach has the potential to 1) allow health scientists to target areas of higher than- and lower than expected risk for PrCA, 2) allow clinicians to know if where a patient resides increases the risk for PrCA, and 3) provides further and detailed

evidence to policy makers to understand that where a veteran resides can influence his/her health.

TABLE OF CONTENTS

DEDICATION iii

ACKNOWLEDGEMENTS iv

ABSTRACT v

LIST OF TABLES xi

LIST OF FIGURES xii

LIST OF ABBREVIATIONS xiv

CHAPTER 1 INTRODUCTION AND SPECIFIC AIMS 1

1.1 Introduction 1

1.2 Specific Aims 3

CHAPTER 2 BACKGROUND AND MATERIALS 5

2.1 Prostate Organ 5

2.2 Prostate Cancer 5

2.3 Prostate Cancer Epidemiology 11

2.4 Data Sources 21

2.5 Software Programs 23

CHAPTER 3 PATIENT AND AREA-LEVEL PREDICTORS OF PROSTATE CANCER AMONG SOUTH CAROLINA VETERANS: A SPATIAL ANALYSIS 29

3.1 Abstract 29

3.2 Introduction 30

3.3 Methods 32

3.4 Results 40

3.5 Discussion 43

CHAPTER 4 NEIGHBORHOOD INFLUENCES ON PROSTATE CANCER MORTALITY AMONG SOUTH CAROLINA VETERANS 55

4.1 Abstract 55

4.2 Introduction 56

4.3 Methods 58

4.4 Results 68

4.5 Discussion 71

CHAPTER 5 PROSTATE CANCER MORTALITY-TO-INCIDENCE RATIO AMONG SOUTH CAROLINA VETERANS: 1999 – 2015 85

5.1 Abstract 85

5.2 Introduction 86

5.3 Methods 88

5.4 Results 93

5.5 Discussion 94

CHAPTER 6 CONCLUSIONS 102

6.1 Research Accomplishments 102

6.2 Research Strengths 105

CHAPTER 7 REFLECTIONS 108

REFERENCES 118

LIST OF TABLES

Table 2.1. Summary of how peer-reviewed epidemiological publications that accounted for spatial variation in prostate cancer research 25

Table 3.1. Descriptive Statistics of South Carolina veteran population who seek healthcare at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria 50

Table 3.2. Univariate $\chi 2$ analysis comparing all prostate cancer patients to non-prostate cancer patients for all categorical variables among the South Carolina veteran population who seek healthcare at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria 51

Table 3.3. Odds ratios and 95% credible intervals for multivariate models comparing prostate cancer patients to non-cancer patients among the South Carolina veteran population who seek healthcare at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria 52

Table 4.1. Descriptive statistics of South Carolina veteran population who were treated for prostate cancer at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria 78

Table 4.2. Univariate comparisons between censored status and risk factors among South Carolina veteran population who were treated for prostate cancer at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria 80

Table 4.3. Hazard ratios and 95% credible intervals for multivariate models that included prostate cancer tumor classification among South Carolina veteran population who were treated for prostate cancer at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria..... 82

Table 4.4. Hazard ratios and 95% credible intervals for multivariate models that did not include prostate cancer tumor classification among South Carolina veteran population who were treated for prostate cancer at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria 83

LIST OF FIGURES

Figure 2.1. Age-adjusted prostate cancer epidemiological trends in the United States of America from 1975 – 2015 26

Figure 2.2. Overall prostate cancer incidence rate per 100,000 in the United States of America from 2011 – 2015 by state 27

Figure 2.3. Overall prostate cancer mortality rate per 100,000 in the United States of America from 2011 – 2015 by state 28

Figure 3.1. STROBE Diagram for the generation of the analytical cohort comparing prostate cancer patients to non-cancer patients among the South Carolina veteran population who seek healthcare at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria 53

Figure 3.2. Systematic plan for Bayesian model evaluation of the analytical cohort comparing prostate cancer patients to non-cancer patients among the South Carolina veteran population who seek healthcare at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria..... 54

Figure 4.1. STROBE Diagram for the generation of the analytical cohort for survival modeling among the South Carolina veteran population with prostate cancer who seek healthcare at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria 84

Figure 5.1. STROBE Diagram for the generation of the analytical cohort for survival modeling among the South Carolina veteran population with prostate cancer who seek healthcare at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria 98

Figure 5.2. Overall and categorical mortality-to-incidence ratios among South Carolina veterans diagnosed with prostate cancer who receive care at US Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria 99

Figure 5.3. Prostate cancer mortality-to-incidence ratios among South Carolina veterans who receive care at US Veteran Health Administration facilities by ZIP code tabulation areas from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria..... 100

Figure 5.4. Prostate cancer mortality-to-incidence ratios results of local Moran's I among South Carolina veterans who receive care at US Veteran Health Administration facilities by ZIP code tabulation areas to identify MIR ZIP code tabulation area clusters from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria..... 101

LIST OF ABBREVIATIONS

ACS ..... American Community Survey

CBOC ..... Community Based Outpatient Clinics

CDC ..... Centers for Disease Control and Prevention

ICD9 ..... International Statistical Classification of Diseases version 9

MIR ..... Mortality-to-Incidence Ratio

NCI ..... National Cancer Institute

PrCA ..... Prostate Cancer

RUCA ..... Rural Urban Commuting Areas

SC ..... South Carolina

SEER ..... Surveillance, Epidemiology, and End Results

SES ..... Socioeconomic Status

TNM ..... Classification of Malignant Tumors

US ..... United States of America

VA ..... United States Department of Veterans Affairs

VACCR ..... Veterans Affairs Central Cancer Registry

VHA ..... Veterans Health Administration

VINCI ..... Veterans Affairs Informatics and Computing Infrastructure

ZCTA ..... Zip Code Tabulation Area

ZIP ..... Zone Improvement Plan

CHAPTER 1

INTRODUCTION AND SPECIFIC AIMS

1.1 Introduction

Cancer is the second leading cause of death in the United States of America (USA) accounting for approximately 22% of all deaths.[1] While prostate cancer (PrCA) is the most common epithelial cancer, accounting for 19% of all new cancer cases in men, only 9% of cancer deaths in men are due to PrCA.[2] This represents the smallest overall mortality to incidence ratio for all cancers for both men and women in the USA as a whole.[2]

Between 2011 and 2015 the incidence of PrCA in the US was 109.0 per 100,000 men while the mortality of PrCA during the same timeframe was 19.5 per 100,000 men.[3] However, both the incidence and mortality of PrCA in the US vary spatially.[3] This can clearly be seen upon an examination of mapping either the incidence or mortality of PrCA within the US by state.; the incidence of PrCA varied by state from a high of 137.5 per 100,000 (Louisiana) to a low of 78.6 per 100,000 (Arizona), and the mortality of PrCA varied from a high of 25.2 per 100,000 (Mississippi) to a low of 13.7 per 100,000 (Hawaii).[3] Furthermore, known risk factors for both the incidence and mortality of PrCA such as race, income, education, and other socio-demographic factors also exhibit spatial variation upon closer examination.[4,5] To better understand how spatial variation

impacts PrCA incidence, mortality, and PrCA risk factors; a large sample size encompassing a wide geographical area is necessary.

The United States Veterans Health Administration (VHA) meets the sample size requirements and has nationwide representation. The VHA is one of the largest healthcare providers in the USA. A key feature for the VHA is that it is not a fee-for-service healthcare provider. Due to this, individuals' financial status is not considered in providing treatment at the VHA. This makes the healthcare delivered by the VHA unique when compared to other healthcare providers in the USA.

Attempting to eliminate or at least minimize the impact of an individual's financial status is critical in the quality of care that one receives. The financial status of individuals is an important component to consider and account for when investigating any cancer in general. In regard to PrCA, the overall incidence of PrCA is higher in men who are at higher socioeconomic status (SES) levels as compared to men who are at lower SES levels.[6-10] However, the incidence of advanced PrCA and higher mortality rates are more likely to occur in lower SES levels.[6-9,11] This indicates that men who are more affluent have better healthcare access, therefore more likely to be diagnosed with PrCA earlier resulting in higher survivorship as compared to less affluent men. This issue is further complicated when race is accounted for because Black men are also more likely to have higher PrCA mortality rates as compared to White men.[6,8,11]

While means-based provision of care is a commendable goal, disparities still may arise through a variety of insidious risk factors that create differences in exposure to as yet unidentified risk factors, early detection and receipt of treatment. These encompass

putative factors ranging from subtle differences related to racism and social class to biological differences that have not been fully elucidated yet. The current state of the science indicates that there may be some biological basis for racial differences observed for a variety of diseases, including PrCA and related comorbidities.[12]

There also may be a racial difference due to societal norms that even VHA healthcare cannot account or compensate for. For example, how does proximity to a VHA facility influence disease incidence and mortality because those traveling over a certain distance receive mileage reimbursements? The VHA cannot dictate where their patients live; however, does where a VHA patient live influence disease incidence and mortality?

1.2 Specific Aims

The aims of this dissertation are to characterize whether geographical risk factors affect the epidemiology for PrCA among VHA patients in SC while controlling for known PrCA risk factors. The aims are:

> *Aim 1: Examine the role of potential geographical risk factors at the individual and neighborhood levels while controlling for known risk factors in predicting incidence of PrCA among US veterans residing in SC.*

The first aim seeks to determine how best to predict incident PrCA within the SC VHA population based on individual-level risk factors, ZIP code characteristics that a VHA patient lives (neighborhood level), and geographical considerations (distance traveled to the treating VHA facility and spatial autocorrelation).

Aim 2: Examine the role of potential geographical risk factors at the individual and neighborhood levels while controlling for known risk factors in predicting mortality of PrCA among US veterans residing in SC.

The second aim seeks to determine how best to predict PrCA mortality within the SC VHA population based on individual-level risk factors, ZIP code characteristics that a VHA patient lives (neighborhood level), and geographical considerations (distance traveled to the treating VHA facility and spatial autocorrelation).

Aim 3: Determine whether risk factors for PrCA age-adjusted mortality-to-incidence ratios among US veterans seeking VHA care vary spatially.

The third aim will use the age-adjusted mortality-to-incidence ratios (MIR) of PrCA within each SC ZIP code. The MIR is a measure that standardizes mortality differences to incidence differences between populations. This third aim will seek to determine the spatial variation in PrCA risk factors influencing PrCA age-adjusted MIRs within SC ZIP code.

CHAPTER 2

BACKGROUND AND MATERIALS

2.1 Prostate Organ

The prostate is an organ that is physiologically part of the male reproductive system. It is an exocrine gland that is approximately the size of a chestnut.[13] It is situated directly below the bladder, surrounding the urethra between the bladder and the penis.[13] The function of the prostate is to contribute the enzyme fibrolysin, citric acid, acid phosphatase, and several key minerals.[13] This contribution accounts for 15 – 30% of the seminal fluid by volume.[13]

2.2 Prostate Cancer

Prostate cancer (PrCA) is the leading type of cancer, accounting for 19% of all new cancer cases in men; however, it is a relatively indolent disease reflected in the fact that only 9% of cancer deaths in men is due to PrCA.[2] This is the smallest overall mortality to incidence ratio for any cancer, in both men and women.[2] Adenocarcinomas, which arise from exocrine or endocrine glands, account for 99% of all PrCA cases.[14]

Depending on the assessment of the virulence (i.e., metastatic potential) of an individual PrCA, it may not be necessary to treat the PrCA but instead to monitor its development. In many PrCA patients, the cancer progresses extremely slowly. Because of this, most (> 80%) men with diagnosed PrCA will die due to other causes rather than due to PrCA.[2,15,16] Because of these considerations, it is evident why the American

Urological Association (AUA) is specific in its recommendations for PrCA screening for the general population. However, the US Preventive Services Task Force views this same evidence and recommends against PrCA screening for the general population.[16] Both organizations, though, recommend men discuss their personal, individual risk factors for PrCA with their physicians as to guide subsequent decisions with regard to their individual PrCA susceptibility and decisions regarding screening conditional on this assessment.

2.2.1 Screening

As with all cancers, the anatomic and pathophysiological progression of PrCA dictates survival and treatment options. In general, the thinking is that the earlier the disease is detected and treated, the greater the probability of survival. Therefore, a key aspect of PrCA research is correctly identifying and characterizing the pathological phase of the PrCA – with the imperative that the screening test distinguishes virulent PrCA from any other prostate-related condition (including benign prostatic hypertrophy [BPH] and indolent cancer). However, a clinician must suspect the possibility of PrCA first, and then conjecture about disease virulence. One source of information is obtained via a digital rectal exam (DRE), which involves palpating the prostate. The purpose of the DRE is to identify an enlarged prostate, a “firm/hard” prostate, or abnormal prostate which could be indications of PrCA. However, simple enlargement of the prostate is relatively useless for distinguishing PrCA from other conditions such as BPH. Another source of information is the prostate specific antigen (PSA) level. The AUA notes that DRE as a

primary screening procedure lacks scientific evidence but is a useful test given increased PSA levels.[16]

PSA, a protein specifically secreted by the prostate, is present in the blood, and can be quantified through assays using blood samples. In general, as men get older, their normal PSA baseline increases.[17] However, an abnormally high PSA level, which varies depending on age, is an indication for further testing for possible PrCA.[17] The use of PSA tests as a screening tool for PrCA in the early first half of the 1990's is the reason why there was a spike in PrCA incidence rates (Figure 2.1).

The (AUA) recommends that men who are not at increased risk should have PSA tests between 55 and 69 years old every two to four years.[16] For those who are at increased risk for PrCA, the AUA recommends discussion with a physician.[16] The AUA notes that the scientific evidence indicates that screening 1,000 men will result in one fewer death due to PrCA over the course of a decade.[16] However, there are issues with using PSA tests to screen for PrCA. In the United States it was found that PrCA based on PSA tests were over diagnosed by 42%.[18,19] These studies are frequently cited as key scientific evidence that determine recommended guidelines for PSA testing. The disadvantages of PSA testing include a 75.9% false positive rate; which leads to a 12% chance of having at least one false positive (> 4.0 ng/ml) following three PSA tests every four years[20], a 13% chance of having at least one false positive (> 4.0 ng/ml) following four PSA tests per year, and a 5.5% chance of having a biopsy as a result of the false positive.[21] In addition, complications due to PrCA treatment include erectile dysfunction, urinary incontinence, loss of bowl control; and complications generally associated with

any surgery (including death) as well as those specifically arising from removing the prostate gland or having a prostate biopsy.[15,16]

Given these considerations, making the most informed decision in the proper management of PrCA becomes more difficult. In other words, the identification of any PrCA is relatively easy as compared to how to treat it, which is strongly related to determining virulence which, in turn, is more difficult. The salient question is: Is it a PrCA that will progress slowly, if at all, for the remainder of the man's life; and therefore the need for medical and surgical intervention is minimal (or even not necessary)? Or, is it a PrCA that will progress rapidly if untreated resulting in the death of the man; and therefore the risks of medical and surgical intervention are smaller than the risk of dying from PrCA?

Being able to differentiate a PrCA that the man will die with from a PrCA that the man will die from will vastly improve clinical decision making and lower the medical cost for treatment. Significant research already has been conducted and continues to be conducted to refine differentiating between a PrCA that will cause death from one that will not.[12,22-26]

2.2.2 Prostate Cancer Classification

As with any solid tumor, the classification of PrCA is detailed. However, because each case of PrCA (as with any cancer) is unique; collapsing of this classification system becomes necessary for epidemiological research. Two classifications systems are commonly used for PrCA epidemiological studies.

The first is based on the size of the primary tumor and how far the cancer has spread, and is known as TNM staging (which refers to the size of the Tumor, Nodal involvement, and Metastatic spread). If the cancer cells are located only in the organ where they originated then the cancer is termed local (or localized). If the cancer cells have spread to the nearby tissues around the organ where they originated (e.g., lymph Nodes) then the cancer is termed regional. If the cancer cells have spread to other organs then the cancer is termed distant (i.e., Metastasized).[27] Over time, and without any intervention, a cancer typically progresses from local to regional to distant. Epidemiologic studies typically will use TNM stage as a categorical variable in multivariable models. Generally, either localized staging is used (no nodal involvement or metastatic spread) as the reference group, which is then compared to regional and distant stages respectively; or localized and regional staging are combined into one group to be used as the reference group and compared to distant stages.

The TNM system is clinically focused. However, if the sample size is sufficiently large in an epidemiological study, it may be possible to incorporate additional substrata and components of the TNM classification levels. This may provide additional insights into understanding the risk factors based on as much available clinical information on the progression of the cancer as possible.

The second classification system is based on the extent of cancer cell differentiation seen, referred to as grade, and evaluated via the Gleason score. The Gleason score is a composite score from the summation of the two most prominent differentiations seen among the cancer cells in the tissue. Each differentiation pattern is

graded on a scale of 1 to 5; were 1 is the lowest score and closely resembles normal tissue and 5 is the highest score. A score of 5 indicates virtually no differentiation, and therefore no resemblance to normal tissue. Because the Gleason score sums the dominant and secondary patterns, the range of possible scores is from 2 to 10. Epidemiological studies and multivariable modeling will use the Gleason score as a continuous variable or may categorize it. Typically, two or three levels are used when the Gleason score is categorized. If two levels are used, then the cutoff is either ≤ 6 or ≤ 7. If three levels are used, then the cutoffs are ≤ 5, 6 – 7, and ≥ 8.

Gleason scores tend to be positively correlated with stage. As PrCA progresses, both the Gleason score and stage progress. However, progression of the PrCA will first be observed in the Gleason score, because it has more distinct levels than the staging of a cancer and is a better indication of its "biological behavior." In other words, as the PrCA increases its Gleason score value by one, it may still remain in the same stage. Therefore, the Gleason score is a more sensitive and specific measure of PrCA progression rather than the stage of the PrCA. Because PrCA typically progresses slowly, the Gleason score becomes a more useful indication for PrCA progression than the staging value.

2.2.3 Treatment

Treatment for PrCA involves active surveillance, surgery, chemotherapy, radiation therapy, or a combination of any of these.[27] Surgery involves either radical prostatectomy (removal of the prostate), if the cancer is localized, or pelvic lymph node dissection (in addition to prostatectomy) if it is more developed (e.g. regional).[27] Initial

treatment is based on the evident progression of the PrCA and further treatment is based on the response of the PrCA to previous treatments.[27] Courses in radiation therapy and chemotherapy are diverse and vary given the development of the PrCA.[27] The National Comprehensive Cancer Network Guidelines lists 15 types of therapies based on initial development of the PrCA and how treatment of the PrCA progresses.[27]

2.3 Prostate Cancer Epidemiology

2.3.1 Overall

Over the last two decades, the incidence and mortality of PrCA appears to have stabilized with a 91.1% Five Year Survival if diagnosed in 2015 (Figure 2.1).[28] Furthermore, it is clear from Figure 2.1 that two additional interrelated phenomena can be observed. The first is that while there was a spike in PrCA incidence in the first half of the 1990's, the mortality actually began the decreasing trend that it continues to this day. By identifying PrCA earlier in its development and treating it, this results in decreasing the probability of mortality. The second observation is that as PrCA was beginning to be identified and treated earlier in its natural history, the 5-year survival increased from 84.4% in 1989 to 95.2% in 1993. These positive trends in PrCA is rightly credited to the advent, use, and promotion of PrCA screening by the healthcare community. However, these are overall trends and more work needs to be done to eliminate disparities in PrCA incidence and mortality rates (e.g. as seen across racial categories, rural/urban categories, SES status, etc.).

While PrCA screening and treatment has significantly improved the probability of surviving PrCA, the risk of developing PrCA still remains. As of 2013, one in six men will develop PrCA, and the risk of developing PrCA increases with increasing age.[2]

2.3.2 Spatial Variability

The importance of accounting for spatial variation among diseases and the application of spatially oriented techniques has been increasing rapidly within the last decade.[29-32] Recognizing the importance of location as it relates to a disease has always been a key cornerstone of epidemiological research. Being able to fully assess that importance and impact, however, is a relatively new development for the epidemiology community.[33-37] PrCA is no exception. As seen in Figure 2.2, the incidence rate of PrCA in the US varies by state.[38] The state with highest incidence rate of PrCA between 2011 and 2015 was Louisiana with a rate of 137.5 per 100,00.[38] The state with lowest incidence rate of PrCA between 2011 and 2015 was Arizona, with a rate of 78.6 per 100,000.[38]

As seen to the Figure 2.3, the mortality rate of PrCA in the US also varies by state.[38] The state with highest mortality rate of PrCA between 2011 and 2015 was Mississippi with a rate of 25.2 per 100,00.[38] The state with lowest mortality rate of PrCA between 2011 and 2015 was Hawaii, with a rate of 13.7 per 100,000.[38]

Figures 2.2 and 2.3 indicate that there are areas in the US were both the incidence rate and mortality rate of PrCA are both high (e.g. the Southeast and Northwestern States). There also are states where incidence is only moderate, but mortality is very high given incidence (e.g., South Carolina, where White men have

relatively low PrCA incidence, Black men have very high incidence and extremely high mortality).

As indicated in the prior sentence the spatial disparity seen in PrCA incidence and mortality rates is confounded by racial/ethnic disparities as well. This can be seen in the variations in incidence rates of PrCA in the US by race and ethnicity across states.[39] Among White Americans from 2011 to 2015, the highest incidence rate was in New Jersey (124.2 per 100,000) and the lowest was in Arizona (74.9 per 100,000) with high concentrations in the Mid-Northern States and North-Mid Atlantic States.[39] Among Blacks from 2011 to 2015, the highest incidence rate was in Delaware (212.3 per 100,000) and the lowest was in Idaho (106.4 per 100,000) with high concentrations in the Southern States and the Eastern States.[39] Among Hispanic Americans from 2011 to 2015, the highest incidence rate was in New York (134.7 per 100,000) and the lowest was in Alaska (42.8 per 100,000) with high concentrations in the North-Mid Atlantic States.[39]

The mortality rates of PrCA in the US also varies by race and ethnicity across states.[39] Among White Americans from 2011 to 2015, the highest mortality rate was in Idaho (23.8 per 100,000) and the lowest was in Florida (15.6 per 100,000) with high concentrations in the Mid-Northern States and Western States.[39] Among Black Americans from 2011 to 2015, the highest mortality rate was in Arkansas (61.9 per 100,000) and the lowest was in Iowa (23.5 per 100,000) with high concentrations in the Southern States and Prairie States.[39] Among Hispanic Americans from 2011 to 2015, the

highest mortality rate was in New Mexico (19.4 per 100,000) and the lowest was in North Carolina (8.5 per 100,000).[39]

It needs to be noted that the incidence and mortality rates for Black Americans and Hispanic Americans are not displayed due to rate stability. This suppression is due to there being fewer than 16 cases in that state.[39] To overcome this issue for this dissertation research; multiple continuous years (1999 – 2015) will be used focusing on one state (i.e. South Carolina) and including only White and Black racial categories.

While it is clear that PrCA incidence and mortality varies across states, it is only in recent years have researchers examined how spatial variation influences the incidence and mortality of PrCA. All of these studies have been at the state level: Florida[40-42], Georgia[43,44], South Carolina[45-47], Texas[48], Virginia[49], Maryland[50], and Massachusetts and Commecticut[51] (Table 2.1).

Several of these studies did not account for the spatial variation in inferential models, but rather identified whether spatial variation was present.[43-45,48,51] Of these, one noted areas of higher PrCA mortality-to-incidence ratios[45], three used a Scan Statistic that identifies adjacent areas of similar PrCA incidence[43,51] and similar PrCA mortality[48], one used a Z-test to identify areas with higher PrCA mortality-to-incidence ratios[44], and one used observed-to-expected ratios to identify areas with higher PrCA incidence[51]. Furthermore, these studies used different geopolitical boundaries for their respective analyses; two used census tracts[43,51], two used heath district/regions[44,45], and one used counties[48].

Seven studies accounted for the spatial variation present in inferential models (Table 2.1).[40-42,46,47,49,50] Of these, two studies used kriging to predict the incidence of prostate cancer.[40,41] Kriging is a geographical technique used to predict an unknown value in a given area of interest given the known values of the areas around the area of interest.[33-37] Both of these studies used Florida counties as the geopolitical boundary of interest.[40,41]

Two of the seven used multilevel modeling but did not directly account for the spatial variation in the models they used.[42,50] In the first of these two studies; individual-level risk factors, county-level environmental factors, and census-level demographic factors were used in multilevel, multivariable logistic regression models of late-stage and poor-grade differentiation of PrCA in Florida.[42] Data from the Florida Cancer Registry was used.[42] Both the late-stage and poor-grade differentiation logistic regression models had similar results for the individual-level risk factors; increasing age, being Black American as compared to White American, being a current smoker as compared to a non-smoker, and being a former smoker as compared to a non-smoker were risk factors while being married was a protective factor.[42] Differences in risk factors appeared at the census tract level between the late-stage and poor-grade differentiation models.[42] Increasing census tract median income and increasing percent college educated were protective in the late-stage model while only increasing percent college educated was protective in the poor-grade differentiation model.[42] The county-level environmental factors; water quality, superfund site, and toxic release were not significant for both models.[42] In addition; a Black American-White American ratio of age-

adjusted PrCA incidence and PrCA late-stage incidence was determined for each county.[42] This racial ratio of PrCA incidences identified Florida counties that had higher ratios than the state average ratio.[42] This indicated Florida counties where the incidence of PrCA and late-stage PrCA incidence were higher among Black Americans than the state average when compared to White Americans.[42] The highest age-adjusted PrCA incidence ratio was seen in Union County where Black Americans were 4.14 times more likely to develop PrCA than White Americans.[42] The highest age-adjusted late-stage PrCA incidence ratio was seen in Holmes County where Black Americans were 8.69 times more likely to be diagnosed with late-stage PrCA than White Americans.[42]

In the second of the two multilevel studies, the incidence for late-stage and high-grade PrCA diagnosis from 1992 to 1997 were modeled in Maryland.[50] Using the Maryland Cancer Registry and the 1990 US Census, the investigators developed a social resource index using census tract, census block groups, and county level information (i.e., neighborhood factors).[50] Four multivariable models were constructed.[50] The first two evaluated the incidence of being diagnosed with late-stage PrCA with one solely modeling individual level risk factors, and the other modeling individual and neighborhood level factors.[50] The second two evaluated the incidence of being diagnosed with high-grade PrCA, with one modeling only individual-level risk factors, and the other modeling individual and neighborhood-level factors.[50] For all four models, no census tract information was significant.[50] The majority of risk factor estimates were consistent across both late-stage and high-grade models and across individual-level and multilevel models.[50] There were three noteworthy risk factors.[50] The first is that an

increase in county level social resources is protective for late-stage PrCA.[50] The second is an increase in county level social resources is a risk factor for high-grade PrCA.[50] These associations may seem to contradict each other; however, the investigators note that factors associated with diagnosis of high-grade disease are influenced by the distribution of resources in the community that mitigate other known risk factors for PrCA (i.e., stress, environmental factors, diet, occupation, etc.) and not by the quantity of resources.[50] The third noteworthy risk factor is that while increasing income (as measured by median income at the census block group) is protective against high-grade PrCA; increasing income among Black Americans (as measured by median income and race at the census block group) is a risk factor.[50]

One of the seven multilevel studies used individual-level information to describe the study population but used only neighborhood factors in its multivariable models.[49] The investigators used the Virginia Cancer Registry from 1990 to 1999 and the 1990 US Census.[49] They developed multivariable models by census tract and county, and stratified by race (Black American and White American).[49] For the county-level models, increasing age was a risk factor for PrCA among both Black Americans and White Americans.[49] Among Black Americans, the interaction between residing in a county were 10 to 19% are below the poverty level and in a majority non-rural county is a risk factor for PrCA.[49] The census tract models produced more statistically significant predictors for PrCA incidence.[49] Increasing age was still a risk factor for PrCA among both Black Americans and White Americans.[49] Residing in majority non-rural census tract was a risk factor for PrCA among both Blacks and Whites.[49] Increasing median household income

was also a risk factor for PrCA among both Black Americans and White Americans, but the effect was more than double among Black Americans as compared to Whtie Americans.[49] Residing in a majority non-rural census tract was significant for both Black Americans and White Americans as median income increases, but in opposite directions; it was a risk factor for Black Americans and protective for White Americans.[49] Unique risk factors for PrCA among White Americans were decreasing census tract poverty levels, increased percent of greater than high school education, increased percent of female heads of households, and increased physician to population ratio while increasing percent of being poor in the census tract was protective.[49] No statistically significant unique risk factors for PrCA among Black Americans were evident.[49] It is important to note that the investigators discussed the differences among risk factors between the census tract multivariable model and the county-level multivariable model; as well as, how the results of these neighborhood-level models differ from individual level factors and the ongoing scientific discussion of using socio-economic status measures as predictors for PrCA (especially among Blacks).[49]

The final two studies accounted for the spatial variation in PrCA incidence and mortality within multivariable models.[46,47] Both studies used South Carolina Cancer Registry Data, county as the geopolitical boundary, and both studies were from the same team of investigators.[46,47] In the first study, the investigators evaluated three model types: baseline (logistic regression), proportional odds, and adjacent-categories logit model.[46] Within each model type, they evaluated whether the distribution was spatially correlated or uncorrelated.[46] In addition, the investigators also evaluated the

three model types without accounting for the location of the cases (i.e., models without any spatial information).[46] Two outcomes were considered: comparing risk factors for being diagnosed with a distant stage PrCA as compared to a local stage PrCA, and risk factors for being diagnosed with a regional stage PrCA as compared to a local stage PrCA.[46] There were 21 models that were evaluated within a Bayesian framework.[46] Given this, the model with the lowest Deviance Information Criteria (DIC) was used to select the best fitting model.[46]

The model comparing late-stage to early-stage PrCA was a logistic regression model yet spatially uncorrelated by county.[46] However, the uncorrelated spatial variation component remained in the model as a random effect.[46] The statistically significant risk factors for being diagnosed with distant stage PrCA as compared to early-stage PrCA include older age, and being Black American as compared to White American; while ever being married decreased the risk of distant stage PrCA (with currently married more protective than separated, divorced, or widowed).[46] The model comparing regional to local PrCA was a logistic regression model with spatially correlated counties.[46] Increasing age is a statistically significant protective factor for regional stage as compared to local stage PrCA.[46] Being Black-American and any ever married category was not statistically significant factors.[46] For both models, the year of diagnosis was statistically significant when compared to the earliest year, 1997.[46] Furthermore, with each successive year, the protective effect increases.[46] In addition, the investigators tested the robustness of these models by altering the prior distribution, which had inconsequential impacts of the results and conclusions.[46] Finally,

and most notably, because the investigators' model comparing regional-stage PrCA to local-stage PrCA accounted for spatial variation, they were able to identify counties with higher (and lower) than expected incidences of regional stage PrCA.[46]

The second study from this team of investigators incorporated the spatial variation in South Carolina PrCA patients between counties obtained again for the South Carolina Cancer Registry.[47] In a unique and innovative modeling paradigm for cancer research, the investigators were able to model the risk factors for time to death (i.e. a survival analysis) while conditioning on the risk factors for the incidence of PrCA.[47] In other words, given the risk factors for PrCA incidence, the risk factors for PrCA mortality were determined.[47] As in their earlier study (already summarized), the authors used a Bayesian perspective with multiple model combinations to determine were spatial variation in PrCA is best accounted for.[47] As before, the model with the lowest DIC was selected, which was a model were PrCA incidence was spatially uncorrelated yet remained in the model as a random effect estimate, but PrCA mortality was spatially correlated.[47] Risk factors for PrCA incidence were being Black American as compared to White American, and being more likely to be diagnosed with regional stage PrCA than with local stage PrCA.[47] Protective factors for PrCA incidence were being married, separated, divorced, or widowed as compared to single; and being diagnosed with distant stage or unstaged PrCA.[47] In regard to PrCA staging, the investigators found that if someone is diagnosed with PrCA, the PrCA will most likely be in the regional stage.[47]

Given the risk factors associated with PrCA incidence, the risk factors associated with PrCA mortality were being Black American as compared to White American; being

diagnosed with a regional, distant, or unstaged PrCA as compared to a local-stage PrCA (where distant stage had an almost 6.5 times increased risk); and increasing age.[47] The only protective factor was being married as compared to single.[47] Furthermore, this joint model fitted the data exceptionally well for men between 56 and 80 years old.[47]

By being able to account for the factors associated with PrCA incidence, the random distribution of PrCA incidence by counties, the factors associated with PrCA mortality, and the spatial correlation of PrCA mortality; the investigators identified counties with higher than expected and lower than expected risk of PrCA mortality.[47]

It is clear, based on this research that accounting for the spatial component of PrCA is not only important but necessary in identifying areas of high risk. It is also clear that if the spatial component cannot be incorporated into the models for any reason than nesting the risk factors based on geopolitical boundaries is the next logical approach. In other words, beginning and ending on the individual may not provide an accurate assessment of PrCA risk.

2.4 Data Sources

The first data source for this dissertation will be the United States Veteran's Affairs Central Cancer Registry (VACCR). The VACCR is a national cancer registry that collects information on all cancer diagnoses among veterans seeking care in VA facilities. Information collected are divided into five broad categories; patient identification and demographics, cancer identification and diagnostic information, staging and extent of disease at diagnosis, first course of treatment, and outcomes.

The second data source will be the VA MedSAS® data files, which are files developed from electronic medical records for the purposes of analytical inquiries.[52] These files are divided into inpatient and outpatient files groups with data available beginning in 1999.[52] The outpatient files are further divided into two file types; one file type documents an outpatient event and while the other file type documents an outpatient visit.[52] The inpatient files are further divided into two groupings.[52] The first describes the type of care received while the second describes the type of file (i.e., information) present.[52] The type of care group is divided into acute care, extended care, observational care, and non-VHA care; while the type of file group is divided in the main file, bed section file (i.e. a medical record describing the care received at a specific inpatient room), procedure file, and surgery file.[52] These two inpatient subgroups collective describe the 15 inpatient MedSAS® files in the VHA.[52]

Information such as scrambled Social Security Numbers, ICD-9 Diagnosis and Procedure codes, race, date of birth, date of death, number of visits, type of visit, clinical service (e.g. oncology, cardiology, etc.), length of medical stay, ZIP code, county, etc. will be used to create a comprehensive cohort file of all VHA patients receiving care between 1999 and 2015.

The third data source that will be used in this dissertation are the electronic VA Vital Status files.[53] These two file types will provide detailed demographic information (e.g. date of birth, date of death, race, and ethnicity).[53]

The fourth and fifth data sources will be the 2010 United States Census and the 2007 □ 2011 American Community Survey.[54,55] Data from these sources will be used for

SC ZIP code level characteristics, and to determine a Social Vulnerability Index (SoVI®) for the continental United States based on ZIP code level information[56-58].

2.5 Software Programs

The three key software programs that will be used for this dissertation will be Microsoft SQL Server®, SAS®, R®, and WinBUGS®.[59-61] Microsoft SQL Server® will serve as the initial program in data management. SAS® will serve as the program for secondary data management, and primarily for data cleaning and data analysis for all three Aims.[59] R® will serve as the primary program involving the creation of maps depicting spatial information for all three Aims.[61] The R® program for the VA will provide the geospatial information for counties, ZIP codes, ZCTA's, and location of VHA facilities in the continental US[61] In addition, R® will serve in the analysis for AIM 3.[61] R® will serve as required in data analysis. In the possibility that AIMS 1 and 2 will be analyzed in a Bayesian framework, then WinBUGS® integrated through R® will serve as the primary data analysis programs.[60,61]

In addition to the aforementioned central programs, this dissertation will also use a remote interface program to the VA servers known as VA Informatics and Computing Infrastructure (VINCI) for all data management, cleaning, and analyses involving VHA data. VINCI was developed by the VA as a resource to VA researchers. The three major advantages of VINCI are security, the wide variety of programs available, and computing power. All VHA data will reside within the VA servers which are behind the VA firewalls and meet the highest standards of data security for patient-level information. No researcher with access to VHA data can remove VHA data from within

the VA servers. VINCI therefore allows access to the data for research purposes but only summary information can be removed (i.e., model estimates, graphs, etc.). Furthermore, VINCI provides access to all the programs that are required for the proper completion of this dissertation. Lastly, the VA servers possess the computing power to effectively and efficiency store and analyze tens of millions of rows of information; as opposed to a personal high-end computer which will have the ability to analyze tens of thousands of rows of information.

Table 2.1. Summary of how peer-reviewed epidemiological publications accounted for spatial variation in prostate cancer research

Type	Article	State	Time	Geographical Unit	Data Source(s)	Analytical Technique
non-inferential	SE Wagner, et. al. [44]	GA	2003 – 2007	Health Districts	GA Cancer Registry	Z-test to identify districts of higher than expected PrCA MIR's*.
non-inferential	SE Wagner, et. al. [43]	GA	1998 – 2008	Census Tracts	GA Cancer Registry	Scan Statistic to identify clusters of census tracts with similar PrCA incidence.
non-inferential	JR Hébert, et. al. [45]	SC	2001 – 2005	Health Regions	SC Cancer Registry	Mapping SC health regions' PrCA MIR's with respective 95% Confidence Intervals.
non-inferential	CE Hsu, et. al. [48]	TX	1980 – 2001	Counties	Texas Vital Records, 1990 and 2000 US Census	Scan Statistic to identify clusters of counties with excess PrCA mortality.
non-inferential	LM DeChello, et. al. [51]	CT & MA	1994 – 1998	Census Tracts	CT Tumor Registry, MA Cancer Registry, 1990 US Census	Scan Statistic to identify clusters of observed-to-expected ratios among census tracts with higher PrCA incidence.
inferential	P Goovaerts and H Xiao [40]	FL	1981 – 2007	Counties	FL Cancer Data Systems	Kriging to extrapolate PrCA incidence based on jointpoint regression model.
inferential	P Goovaerts and H Xiao [41]	FL	1981 – 2007	Counties	FL Cancer Data Systems	Kriging to extrapolate late-stage PrCA incidence using longitudinal models.
inferential	H Xiao, et. al. [42]	FL	1990 – 2001	Counties and Census Tracts	FL Cancer Data Systems and 2000 US Census	Counties and Census Tracts as nesting variables for multilevel multivariable models predicting PrCA incidence.
inferential	H Zhou, et. al. [46]	SC	1997 – 2001	Counties	SC Central Cancer Registry	Determine best fitting multivariable spatial model predicting PrCA incidence. Spatial variation accounted for in models.
inferential	H Zhou, et. al. [47]	SC	1997 – 2001	Counties	SC Central Cancer Registry	Determine best fitting multivariable spatial model predicting time to death while conditioning on the risk factors for the incidence of PrCA. Spatial variation accounted for in models.
inferential	MN Oliver, et. al. [49]	VA	1990 – 1999	Counties and Census Tracts	VA Cancer Registry and 1990 US Census	Counties and Census Tracts as nesting variables for multilevel multivariable models predicting PrCA incidence.
inferential	AC Klassen, et. al. [50]	MD	1992 – 1997	Counties and Census Block Groups	MD Cancer Registry and 1990 US Census	Counties and Census Block Groups as nesting variables for multilevel multivariable models predicting PrCA incidence.

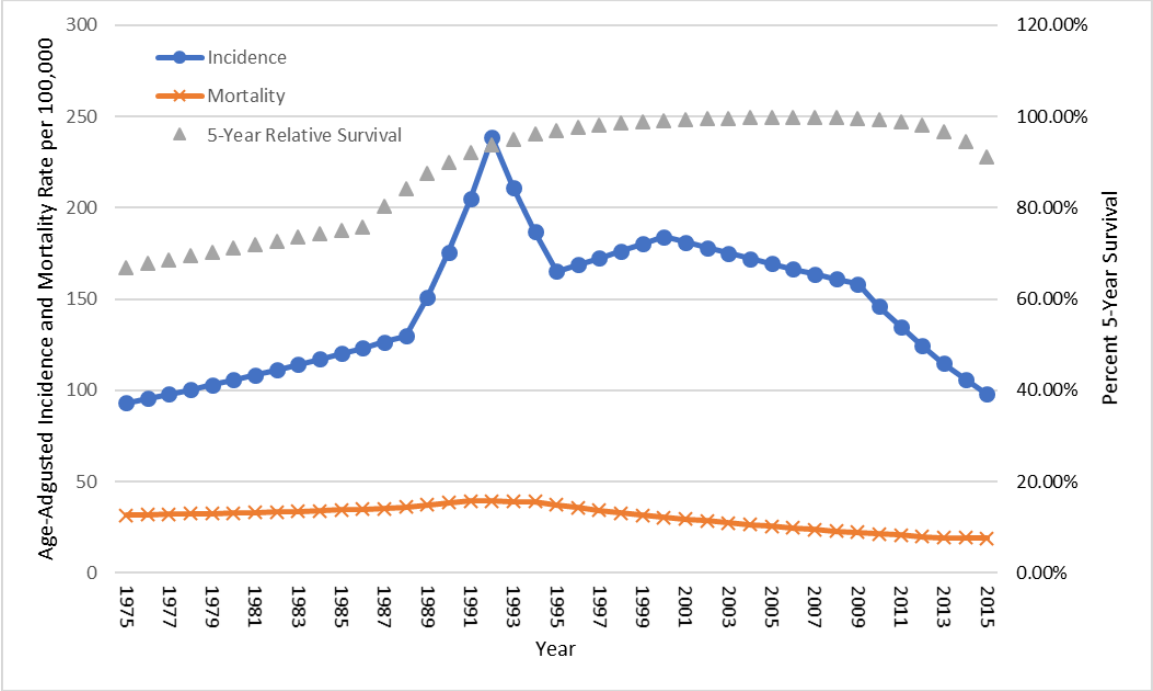

Figure 2.1. Age-adjusted prostate cancer epidemiological trends in the United States of America from 1975 – 2015[1]

[1]Figure 2.1 was created using publicly available data from the National Cancer Institute (NCI) Surveillance, Epidemiology, and End Results Program Stat Fact Sheets: Prostate Cancer webpage (http://seer.cancer.gov/statfacts/html/prost.html).

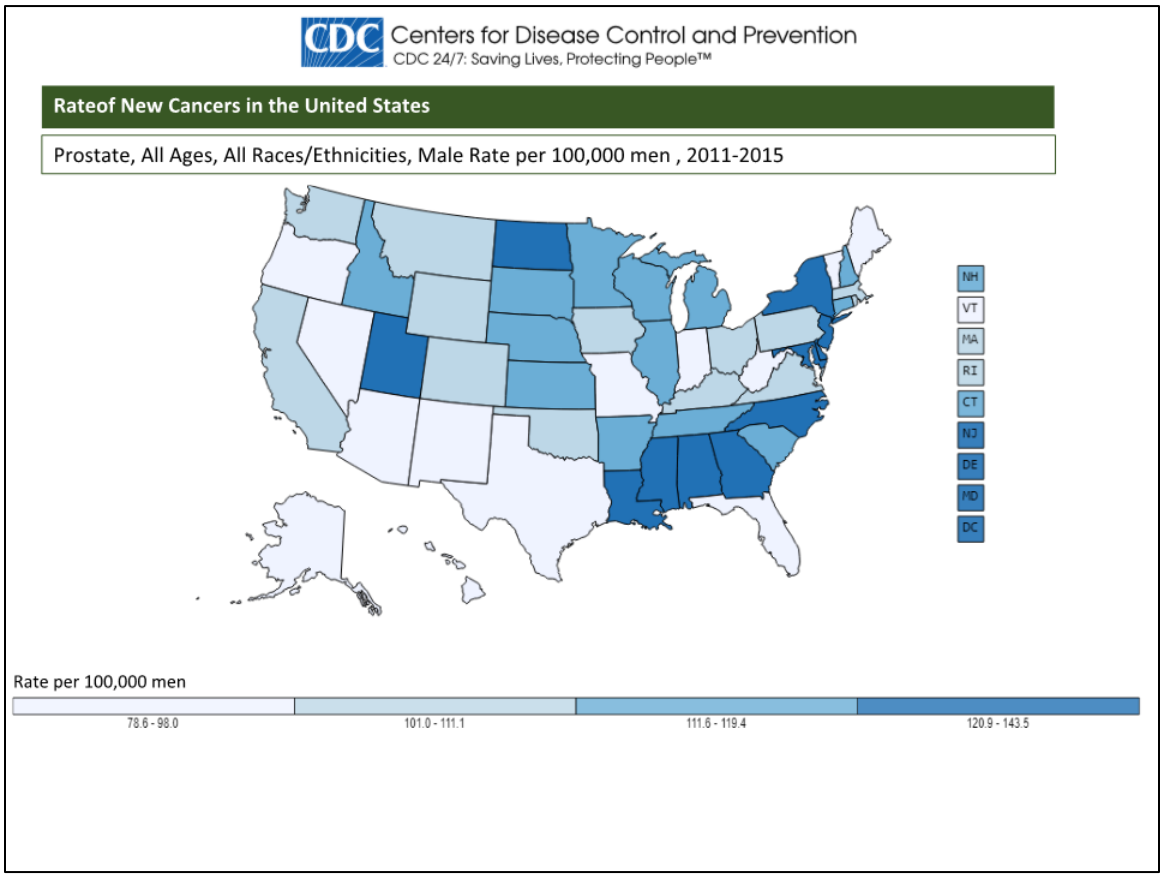

Figure 2.2. Overall prostate cancer incidence rate per 100,000 in the United States of America from 2011 – 2015 by state[2]

[2]Figure 2.2 was created and export using the US Centers for Disease Control and Prevention (CDC) Data Visualizations webpage at (https://gis.cdc.gov/Cancer/USCS/DataViz.html)

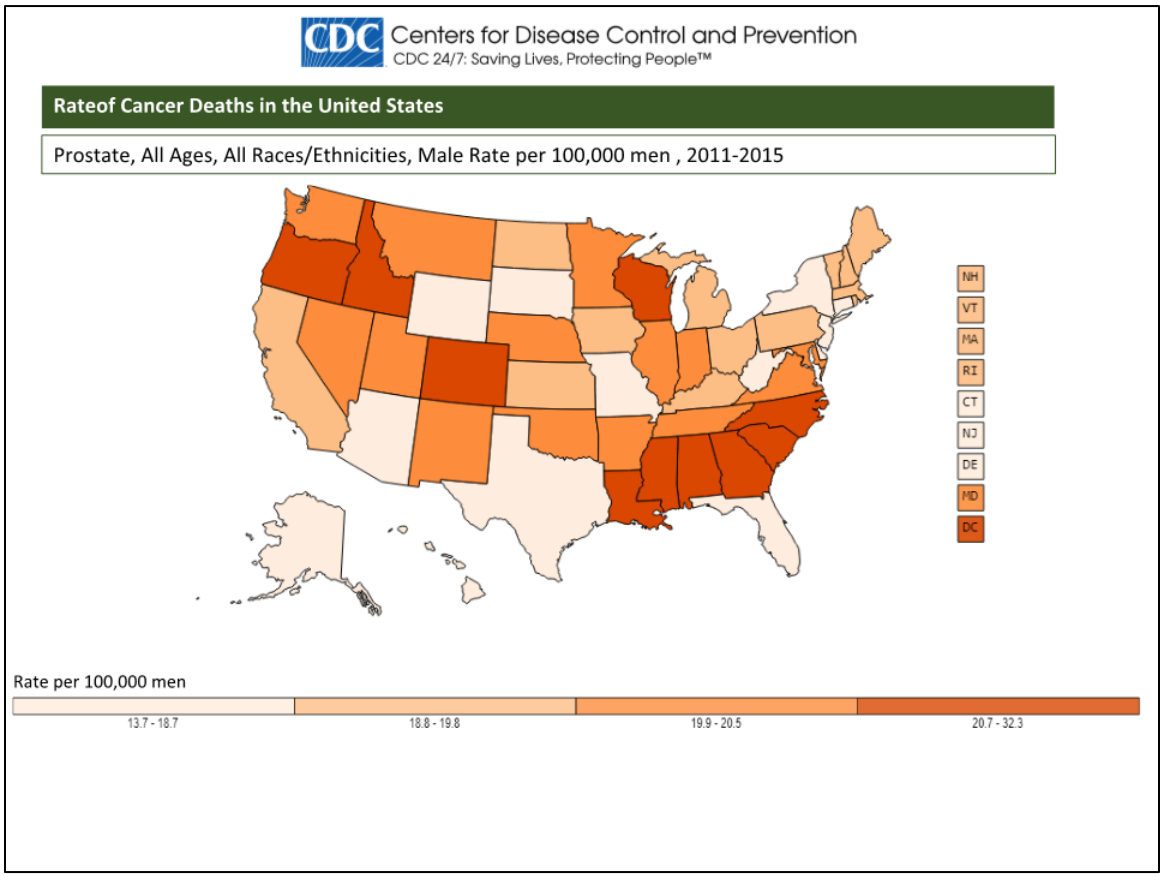

Figure 2.3. Overall prostate cancer mortality rate per 100,000 in the United States of America from 2011 – 2015 by state[3]

[3]Figure 2.3 was created and export using the US Centers for Disease Control and Prevention (CDC) Data Visualizations webpage at (https://gis.cdc.gov/Cancer/USCS/DataViz.html)

CHAPTER 3

PATIENT AND AREA-LEVEL PREDICTORS OF PROSTATE CANCER AMONG SOUTH CAROLINA VETERANS: A SPATIAL ANALYSIS

3.1 Abstract

Background: Racial and socio-economic status (SES) disparities exist in prostate cancer (PrCA) incidence and mortality. Less is known regarding how geographical factors, including neighborhood social vulnerability, affect PrCA risk. The Veterans Administration Medical System provides a unique means for studying PrCA epidemiology among diverse individuals with ostensibly equal access to healthcare.

Methods: From the US Veteran's Health Administration electronic medical records database from January 1999 to December 2015, we identified 3,736 PrCA patients and 104,017 cancer-free controls from South Carolina (SC). Data were analyzed using a Bayesian multivariate conditional autoregressive model, which accounted for individual-level factors, area-level factors, spatial random effects and autocorrelation.

Results: As expected, after accounting for age (6-fold and 13-fold increases in men 40-50 years and >50 years, respectively), race was an important risk factor, with 3-fold higher odds among Blacks in the fully adjusted model [ORadj: 2.98 (2.77, 3.20)]. After accounting for all other factors, residing in a ZIP code with an intermediate level of Social Vulnerability Index (SoVI®) versus the lowest, least-vulnerable, ZIP codes,

increased PrCA risk by 35% [ORadj=1.35 (1.11, 1.65); while residing in the most socially vulnerable ZIP codes increased risk by 39% [ORadj=1.39 (1.11, 1.75).

Conclusions: While accounting for known risk factors for prostate cancer, including age, race, and marital status, we found geographic areas in SC characterized by higher than average social vulnerability with higher rates of incident PrCA among veterans. Outreach for screening, education and care coordination may be needed for veterans in these areas.

3.2 Introduction

Prostate cancer (PrCA) is the most common cancer in men, accounting for 22% of all new cancer cases in American men.[2] Identifying novel risk factors associated with PrCA risk beyond established risk factors has been met with little success. Well known risk factors include race, education, marital status, family history, diet, and age.[2,7,14]

In the general United States (US) population, PrCA incidence displays a non-uniform geographical distribution, where in 2015 the national average was 99.1 new PrCA cases per 100,000.[62] The top three highest states being New Jersey (127.4.7 per 100,000), Mississippi (126.7 per 100,000), and Louisiana (125.7 per 100,000) while the three lowest rates being Alaska (61.0 per 100,000), Nevada (69.5 per 100,000), and New Mexico (72.6 per 100,000).[62] These differences across states may be potentially associated with macro-level factors such as area-level income, educational attainment, rurality, and environmental factors.[42,49,50] Examining these potential area-level differences while accounting for individual-level risk factors for PrCA may provide

invaluable insight into better identifying individuals at higher than normal risk just by knowing where they live.

Epidemiologic research is focused on evaluating putative disease risk factors, through the generation and testing of hypotheses related to risk factors across the socioecological spectrum including environmental exposures and social determinants of health. Identifying areas with higher- or lower-than-expected numbers of PrCA cases using spatial techniques[43-45,48,51], employing spatial nested models[42,49,50] or autoregressive models[40,41], and using neighborhood factors of the geographical unit of investigation[42,49,50] can provide a more thorough understanding of PrCA by expanding the list of risk factors associated with PrCA.[40,41,46,47] In addition, accounting for spatial autocorrelation may allow researchers to improve on model fit, which in turn will allow them to determine geographical regions that are at increased or decreased risk for PrCA.[37]

The goal of this study is to identify the impact of area-level characteristics including social vulnerability on incident PrCA diagnosis while controlling for established individual-level risk factors among veterans who receive care in the VA health care system, the largest universal access, integrated delivery system in the nation. We limited our analysis to veterans with PrCA cancer who were diagnosed and treated at a VA facility, who reside in South Carolina, a state with a higher incidence of PrCA cancer than the national average.[38] A key consideration is that while the majority of men who receive care in the VA are of lower socioeconomic status, all have guaranteed access to PrCA screening and treatment in particular and health care in general.

3.3 Methods

3.3.1 Study Design

This study employed a retrospective cohort design. The timeframe was from 01 January 1999 to 31 December 2015. The unit of investigation was the US Census Bureau-defined ZIP code tabulation areas (ZCTA), which consists of aggregated Zone Improvement Plan (ZIP) codes developed by the US Postal Service.[63] All geographical related information was either linked to the shapefile (i.e. ZIP code patient level information linked to ZCTA's in the shapefile), directly obtained at the ZCTA level (i.e. neighborhood-level information), or directly obtained from the 2015 ZCTA shapefile from the US Census Bureau (i.e. individual ZCTA boundaries).[64]

3.3.2 Data Sources

Data were obtained from the United States Department of Veteran Affairs and the United States Census Bureau. From the United States Department of Veteran Affairs the following datasets were used: all MedSAS® datasets, Master Vital Status dataset, Mini Vital Status dataset, and Primary Oncology dataset from the VA Cancer Registry.[52,53] The VA Master Vital Status and VA Mini Vital Status datasets provided patient-level information (i.e., each unique VA patient is listed only once in these files) while the VA MedSAS® datasets provided visit-level information (i.e., with ≥1 record/subject).[52,53] From the United States Census Bureau the following datasets were used: United States 2010 Decennial Census, the 2007 – 2011 Five-Year American Community Survey, and the 2015 ZCTA shapefile for the US.[4,5,64]

3.3.3 Exclusion Criteria

Subjects were excluded based on the following criteria (see Figure 3.1 for the STROBE Diagram establishing the analytic cohort)[65]: data in the Primary Oncology Dataset from the VA Cancer Registry could not be linked to MedSAS® datasets[52], a non-prostate cancer diagnosis in the Primary Oncology Dataset from the VA Cancer Registry, prostate cancer diagnosis in the Primary Oncology Dataset from the VA Cancer Registry prior to January 1, 1999, date of birth differs by more than 365 days between VA MedSAS® datasets[52] and VA Vital Status datasets[53], age during the timeframe did not wholly or partially fall between 40 and 70 years old, missing ZIP code information, racial classification other than white or black, and not residing in a South Carolina ZIP code.

Overall, 964,047 unique subjects from a total of 1,159,188 (83.17%) unique cancer diagnoses were included in the Primary Oncology Dataset. Of these subjects, 80.47% could be linked to unique patients in the MedSAS® datasets.[52]

There were 235,782 unique PrCA cases in the Primary Oncology Dataset that could be linked to the VA MedSAS® Files. There were 230,401 (97.72%) PrCA cases were exclude due to not residing in a South Carolina ZIP code. Of the 5,384 PrCA cases residing in a South Carolina ZIP code; 1,648 (30.61%) were ineligible due to diagnosis prior to 01/01/1999, date of birth discrepancy, age at first VA visit not satisfied, missing ZIP code, and/or not being white or black. There were 9,976,241 unique subjects in the VA MedSAS® datasets that were not linked to any subjects in the VA Cancer Registry.[52] These subjects were classified as not having cancer. There were 9,798,934 (98.22%) non-cancer cases were excluded due to not residing in a South Carolina ZIP code. Of the

177,306 non-cancer cases residing in a South Carolina ZIP code; 73,289 (41.33%) were ineligible due to date of birth discrepancy, age at first VA visit not satisfied, missing ZIP code, and/or not being white or black.

The final South Carolina analytical level cohort consisted of 3,736 subjects with PrCA and 104,017 non-cancer subjects (Figure 3.1).

3.3.4 South Carolina Geographical Area

South Carolina veterans receiving care at VHA facilities resided in 413 of South Carolina's 424 ZIP codes (97.41%). The 2015 ZCTA shapefile from the US Census Bureau was used South Carolina ZCTA's were in turn used to link the selected ZCTA's to ZIP codes with identical designations in the VA MedSAS® Files.[52,64]

3.3.5 Primary Outcome Variable

Primary PrCA patients between January 1, 1999 and December 31, 2015 were identified from the VA Primary Oncology Dataset.[52,53] Non-cancer controls were identified by removing all cancer patients in the VA Primary Oncology Dataset from the VA MedSAS® Files.[52,53] Non-cancer controls were then limited to those patients whose first VA visit occurred between January 1, 1999 and December 31, 2015.

3.3.6 Individual-Level Independent Variables

The date of the first VA visit for each subject was obtained from the MedSAS® datasets and the date of birth (DOB) was obtained from the VA Mini Vital Status dataset.[52,53] Four age strata variables were created: age <40, age from 40 – 50, age from >50 – 60, age from >60 – 70 years. The information located within the VA Master Vital Status file for race and ethnicity were classified into White, Black, Hispanic, Asian,

Hawaiian/Pacific Islander, and Native/Alaskan America, Other, and Unknown. Subjects other than white or black racial category were exclude because they accounted for 9.90% of the PrCA and non-cancer patients.[53] The marital status field at the time of diagnosis in the Primary Oncology Dataset was used for PrCA cases, while marital status field in the MedSAS® Outpatient dataset was used for controls to determine marital status for the cohort.[52] Marital status was categorized as Married, Previously Married (Divorced, Widowed, Separated), and Never Married/Unknown. The unique VA facility numerical codes in the MedSAS® datasets were used to identify which facility a subject visited.[52] The number of different VA facilities , the most frequented VA facility, and the number of times each subjected visited their most visited VA facility were determined.

The ZIP code listed the most times for each subject's total visits in the MedSAS® datasets was used for all ZIP code related information for each subject.[52] For PrCA cases, the ZIP code at diagnosis was obtained from the Primary Oncology dataset. For PrCA cases, the ZIP code at diagnosis was used for cases without ZIP code information from the MedSAS® datasets (n = 6,784).[52]

There were 131 unique VA facilities (i.e. VA hospitals and VA Community Based Outpatient Clinics) determined from VA MedSAS® datasets.[52] The distance between the most frequented VA facility and ZCTA-linked ZIP code listed the most times was represented as a continuous variable and obtained using the MedSAS® datasets.[52] This ZIP code information was linked to the 2015 ZCTA shapefile from the US Census Bureau.[64] Using the shapefile, a contiguity straight line origin-to-destination matrix distance matrix between VA facilities and ZCTA code centroids to determine the

distance traveled by a veteran to his most frequented VA facility.[52,64] The distance between the most frequented VA facility and the ZCTA-linked ZIP code listed the most times was categorized into: 0 – 25 miles, >25 – 55 miles, and >55 miles. These cutoffs were selected *a priori* to represent how many miles can approximately be driven within 30 minutes, 1 hour, and more than 1 hour.

3.3.7 Neighborhood-Level Independent Variables

The Social Vulnerability Index (SoVI®): The SoVI®, initially developed by the Hazards and Vulnerability Research Institute (HVRI) at the University of South Carolina, is a composite measure of neighborhood-level factors obtained from publicly available population based datasets of factors associated with the health of individuals within those neighborhoods.[58] The SoVI® uses ZCTA measurements obtained from the 2010 US Census and the 2007 – 2011 5-Year American Community Survey.[4,5] The SoVI® is a relative measure represented on a continuous scale from negative infinity to positive infinity in which the greater a geographical unit's SoVI® score is the less prepared for, respond to, and recover from a disaster that area is compared to a geographical unit with a lower SoVI® score.[66] A SoVI® score was assigned for each subject's most frequently listed ZIP code. The SoVI® measurements were stratified into 3 categories using cutoffs of 1 standard deviations: Low ($-\infty$, -2.04], Medium (-2.04, 2.79], and High (2.79, ∞).

Specific Zip Code Level Factors: The 2007 – 2011 Five-Year American Community Survey was used to determine the percent of people within each ZCTA with at least a college degree with tertile cutoffs of [0%, 12.86%], [12.86%, 22.60%], [22.60%, 100%],

percent of people within each ZCTA living in poverty with tertile cutoffs of [0%, 7.96%], [7.96%, 15.86%], [15.86%, 100%], percent of Black Americans within each ZCTA with tertile cutoffs of: [0%, 0%], [0%, 2.73%], [2.73%, 100%], and percent of people within each ZCTA that were at least 65 years of age with tertile cutoffs of: [0%, 11.95%], [11.95%, 16.92%], [16.92%, 100%].[5]

3.3.8 Statistical Analyses

Descriptive statistics were calculated for the analytic cohort (i.e., t-test for all continuous independent variables and a chi-squared tests for all categorical independent variables). Frequentist multivariate models using SAS® were first developed to inform key steps in development and evaluation of the final models.[59] These models were developed for all individual level risk factors, and nested models incorporating both individual- and neighborhood-level risk factors with patient ZCTA-linked ZIP code residence as a random effect were created for model selection purposes. Bayesian models were developed for the null (empty) models; individual-level models; nested models incorporating both individual, area-level risk factors, and accounted for patient ZCTA-linked ZIP code residence as a random effect; and the multivariate conditional autoregressive models incorporating both individual, area-level risk factors, accounted for patient ZCTA-linked ZIP code residence as a random effect, and accounted for the spatial autocorrelation between ZCTA-linked ZIP codes using queen contiguity based spatial weights (i.e. all bordering ZCTAs for a given ZCTA.will have a weight of 1 while all other ZCTAs will have a weight of 0).[34,36,37]

All categorical variables were evaluated using created dummy variables. Any independent individual-level variables where 95% confidence intervals could not be computed or extremely large point estimates with extremely wide 95% confidence intervals were excluded since those variables had almost all observations in one stratum. Independent individual-level variables not excluded were used in a manual backward stepwise approach to generate the final multivariate individual-level model were the added variable remained if it changed the risk factor estimates of the other variables by at least 10%.

A manual backward stepwise approach was used to determine the final nested-level model. Each neighborhood level independent risk factor was evaluated separately with the final multivariate individual-level model. The model with the lowest Akaike information criterion (AIC) was selected as the initial nested-level model.[67] Subsequently, the neighborhood factor with the next lowest AIC was added to the initial nested-model. If any risk factor estimates changed by more than 10% than the additional neighborhood factor remained. This was repeated for all neighborhood level factors.

The median odds ratio (MOR) also was calculated for each model. The median odds ratio uses the area-level variance in the model and calculates a statistics that can be interpreted as the median difference in odds between the ZCTA-linked ZIP code with the highest compared to the lowest risk for two individuals with the same evaluated risk factors (i.e. the MOR is the risk estimate for the unexplained variation in the model).[68]

Therefore, the model with the best fit will also have the lowest MOR because risk factors in the model account for more of the variance in the outcome.

A Global Moran I's Test was conducted on the ZCTA-linked ZIP code level residuals of the final models to determine existence of spatial autocorrelation.[37] If the Global Moran I's Test was statistically significant, indicating the presence of spatial autocorrelation, then models accounting for spatial autocorrelation would be constructed using risk factors included in the final nested-level models. If the Global Moran I's Test was statistically nonsignificant, indicating no presence of spatial autocorrelation in the data, then the nested-level model would become the final overall model for that comparison.

Based on the frequentist models, the final Bayesian models were developed. A Bayesian framework was selected for all final models because we assumed the factors associated ZIP code residence are not independent and random between neighboring ZIP codes and therefore will have a structured variance that needs to be accounted for. Incorporating a conditional autoregressive (CAR) component into the models is the most popular way to account for this structured variance. The most appropriate framework when incorporating a CAR component in the models is to use a Bayesian framework. It was decided prior to modeling building that since models incorporating a CAR component would have to be conducted in a Bayesian framework then all other final models would also be conducted within a Bayesian framework so as to compare between final models.

All Bayesian models used uninformed priors. The initial statistical inference from each Bayesian model was based on 15,000 iterations (i.e. samples) after 5,000 burn-in period (Figure 3.2). The convergence of the sample chains was evaluated using the Geweke diagnostic, which compares the first 10% of kept iterations to the last 50% of kept iterations.[69] The mean value was used as the point estimate for each individual and area level risk factors as well as the MOR estimates. The 2.5% and 97.5% iteration cutoffs were used as the ranges for the 95% credible intervals (CrI) for each risk factor.

Bayesian models created were compared to each other using the Deviation Information Criteria (DIC).[60] The model with the lowest DIC was selected as the best-fitted model for that comparison in that analytic cohort.

3.3.9 Logistical Aspects

All data storage, management, and analyses were conducted with the VA Informatics and Computing Infrastructure (VINCI) servers. All data management was done using Microsoft SQL Server. All univariate and frequentist modeling analyses were conducted using SAS® version 9.4 software.[59] All Bayesian modeling was conducted using WinBUGS® version 1.4.3 and R version 3.3.2.[60,61] All evaluation of Bayesian modeling was conducted using R.[61] All statistical tests used an α-level of 0.05. This project had Institutional Review Board Approval from the University of South Carolina (Pro00036431) and the William Jennings Bryan Dorn VA (10404).

3.4 Results

PrCA patients accounted for 3.5% of the final analytical cohort. The plurality of patients were between >60 to 70 years of age (37.7%). Most prostate cancer patients

were white (65.5%), married (62.3%), and travelled more than 55 miles to a VA facility for care (52.0%) (Table 3.1). The median distance traveled to a VA facility was 60.0 miles.

3.4.1 Univariate Statistics

PrCA patients were more likely than non- PrCA patients to be older for their first VA visit (age from > 50 – 60 years: 45.1% vs. 34.0%; age from > 60 – 70 years: 41.0% vs. 37.6%). PrCA patients were also more likely to be Black (54.5% vs. 33.6%), less likely to be have been married at diagnosis (58.4% vs. 62.5%), more likely to be have been previously married (32.7% vs. 22.4%), less likely to never have been married or have an unknown marital status (9.0% vs. 15.1%), and more likely to travel between 25 and 55 miles to a VA facility to receive care than their non-cancer counterparts (Table 3.2).

3.4.2 Multivariable Bayesian Modeling

Seven models of increasing complexity using a Bayesian framework were fit (Table 3.3). All independent variables were significant for the first model. The second model was the nested null model where the subject's ZCTA-linked ZIP code was accounted for. This model had a MOR of 1.37 (1.30, 1.44) as well as the highest DIC of all seven models. The third model accounted for the individual-level factors as well as nesting the subjects in their ZCTA-linked ZIP codes. The fourth model included area-level factors including percent of the population living in poverty and categorized SoVI® scores. While Model 3 and Model 4 have identical DICs, Model 4 had a smaller MOR indicating a better fitted model. The fifth model was the spatial null model with a spatial

random effect and autocorrelation term. Model 5 fit the data better than Model 2, with both a smaller DIC and MOR. The sixth model was the spatial model with only the individual-level factors. The seventh model was the spatial model with the individual-level factors as well as the ZCTA poverty level and the categorized SoVI® scores for each model. The fifth model was the spatial null model with a spatial random effect and autocorrelation term. Model 5 fit the data better than Model 2, with both a smaller DIC and MOR. The sixth model was the spatial model with only the individual-level factors. The seventh model was the spatial model with the individual-level factors as well as the ZIP code poverty level and the categorized SoVI® scores for each model.

Model 7, the most complex model, had the smallest DIC and MOR [1.14 (1.11, 1.19)] of all the models, indicating that it is the model with the best fit. PrCA patients were more likely to be older when they had their first VA visit as compared to non-cancer patients, with the risk increasing with increasing age; 40 – 50 years old (ORadj: 5.63 (3.94, 8.39), >50 – 60 years old [ORadj: 13.11 (9.29, 19.53)], and >60 – 70 years old [ORadj: 13.31 (9.44, 19.81)]. PrCA patients were more likely to be Black [ORadj: 2.98 (2.77, 3.20)]. PrCA patients were more likely to have been married previously as compared to being currently married [ORadj: 1.47 (1.37, 1.58)]. However, those who had never been married or their marital status was unknown were at lowest risk relative to those who were married [ORadj: 0.55 (0.49, 0.62)].

Distance between 25 and 55 miles from the patient's ZCTA-linked ZIP code centroid to the most frequented VA facility was statistically insignificant in Model 7 [ORadj: 1.06 (0.92, 1.21)]. This stratum for this variable was also statistically insignificant

in Model 4 while statistically significant for Models 1, 3, and 6. In model 7, PrCA patients were less likely to be living more than 55 miles from the patient's ZCTA-linked ZIP code centroid to the most frequented VA facility vs. their non-cancer controls [ORadj: 0.86 (0.75, 0.98)]. This stratum for this variable was also statistically significant in Model 4 while statistically insignificant for Models 1, 3, and 6.

PrCA patients were more likely to live in ZCTA-linked ZIP codes with a SoVI® class 1 - 2 standard deviations from the ZCTA-linked ZIP codes with the lowest Social Vulnerability Index [ORadj: 1.35 (1.11, 1.65)] and 3 standard deviations from the ZCTA-linked ZIP codes with the lowest SoVI® score [ORadj: 1.39 (1.11, 1.75)]. These estimates remained consistent and statistically significant for all models that accounted for them.

Residing in ZCTAs with the highest poverty level (>= 15.86%) increased the risk of PrCA by 24% in Model 4 [ORadj: 1.24 (1.01, 1.48)]. However, this ZCTA poverty level was statistically insignificant after accounting for the spatial autocorrelation between ZCTAs in Model 7 [ORadj: 1.13 (0.93, 1.35)].

3.5 Discussion

Our study identified an association between a veteran's ZCTA-linked ZIP code of residence and prostate cancer risk. This study used a unique approach to account for individual-level risk factors, area-level risk factors, and spatial random effects and autocorrelation within one modeling framework.

The distance traveled to receive care demonstrates a unique dynamic. When distance is the only measure describing the location of a veteran (i.e., Model 1), it is

statistically significant for those veterans traveling between 25 and ≤55 miles (i.e., a VA healthcare facility within an approximate 30-minute to 1-hour commute). That stratum remains statistically significant (i.e., Models 3 and 6), until accounting for ZCTA poverty level and SoVI® (i.e., Models 4 and 7).

Interestingly, veterans traveling >55 miles (i.e. a VA healthcare facility more than an approximate 1-hour commute), became a statistically protective effect when ZCTA poverty level and SoVI® were accounted for (i.e., Models 4 and 7). This protective effect may be an indication of those veterans living in rural areas.[8] The primary VA healthcare facilities in South Carolina are located in urban areas, which are limited to South Carolina. The rural/urban distinction of a ZIP code proceeds towards rural as the distance between VA facilities in South Carolina and a veteran increase. Obertova et. al. concluded in their systematic review article on the urban-rural disparity in PrCA that those living in rural communities could be less likely to seek preventive/wellness care visits, which in turn, are less likely to be screened for PrCA and thereby less likely to be diagnosed with prostate cancer resulting in a fewer PrCA cases being associated with that area.[8]

A potential dynamic that will be explored in future research is what type of VA facility a veteran receives his care; is it a VA hospital or a VA Community Based Outpatient Clinics (CBOC). CBOC's serve an important function in providing care to veterans living in less populated regions; however, CBOC's are limited in what type of care they can provided. This analysis did not differentiate between VA hospitals and

CBOCs. However, based on the plausible conclusions inferred by these results, determining if differences exist between veterans who receive care at a VA hospital and a VA CBOC may be beneficial in improving how health care is delivered to this population.

Our results demonstrated that more than one area-level risk factor was associated with higher rates of incident PrCA in the VA population. The models confirmed that area poverty was associated with higher incidence of PrCA.[6,70] However, of key interest, is our finding that models accounting for the spatial autocorrelation between ZCTA-linked ZIP codes diminished the impact of area poverty while the SoVI® estimates remained similar. While a poverty level of >15.9% went from being statistically significant in Model 4 to statistically non-significant in Model 7, risk estimates for the SoVI® remained statistically significant and stable across those models. A credible explanation is that poverty levels for a community cannot be constrained by political/administrative boundaries such as ZIP codes, therefore when spatial autocorrelation is accounted for as in Model 7, it inherently accounts for similarities such as poverty level. However, the SoVI® is a composite measure that includes multiple community-level measures that cannot be fully accounted for by spatial autocorrelation alone. This opens the possibility of further exploration of how the measures that comprise the SoVI® individually and collectively impact the outcome of interest.

While addressing societal issues of poverty and community social vulnerability is outside the scope of the VA, these results indicate that veterans living within an approximate 30-minute to 1-hour commute to a VA healthcare facility maybe at

increased risk for being diagnosed with an incident case of prostate cancer, especially if they reside in a ZIP code with a high percentage of the population in poverty and an increased SoVI® class. These results may indicate 1) a false sense of health assurance knowing that they live within 1-hour of VA facility, but do not choose to utilize the VA facility for preventive care/wellness/routine care; 2) a veteran who requires increased healthcare services chooses to live closer to a VA facility; or 3) an indication of the urban/rural dynamic within the veteran population seeking care at VHA facilities.[8]

We also showed that established prostate cancer risk factors, such as marital status and being Black, were statistically significant factors associated with PrCA.[2,71-73] In addition, their respective point estimates and 95% credible intervals were nearly identical for all models.

Findings on age at first VA visit can provide insight into understanding age of prostate cancer diagnosis for veterans. Veterans diagnosed with incident prostate cancer were more likely to begin their VA healthcare later in life than those veterans without incident prostate cancer or who did not develop incident prostate cancer during the 16-year study timeframe. Furthermore, the risk of incident prostate cancer more than doubled from the 40 – 50 to the >50 – 60-year-old strata, at which point there was slight differences in incident prostate cancer risk between the >50 – 60 and >60 – 70-year-old strata. This risk of age at first VA visit may also be caused by the recent influx of new veterans who served in Operations Enduring Freedom and Iraqi Freedom resulting in a larger population of younger veterans in the VA system who are still at risk for developing chronic diseases such as prostate cancer but may cause a bias when used as

part of a larger comparison group to evaluate those chronic diseases at this point in time.

3.5.1 Strengths and Limitations

This study used electronic medical records for patient-level variables. As such, issues that are inherent in using such information are present.[74] The data files used have been extensively utilized and queried in other VA-based peer-reviewed research and are of high reliability and validity.[75-78] Furthermore, biases that may be present between PrCA patients and non-cancer patients will be non-directional because data management and cleaning did not differentiate between those patients. Of course, use of pre-existing datasets precludes examining data outside of its intended scope. For example, we had no information on putative prostate cancer risk factors including diet, physical activity, cardiorespiratory fitness, and a host of other psychosocial factors.

The use of ZCTA's as a proxy for ZIP codes as the geographical unit of analysis has been demonstrated to have boundary representation issues especially for non-populated ZCTA's as compared to ZIP codes.[79] However, limiting the population under investigation to South Carolina, which is a small, stably-populated state as compared to the rest of the nation should mitigate the boundary issue. Therefore, the only areas impacted by boundary issues are those ZCTAs bordering the states of Georgia and North Carolina, which is always present in all spatial analyses where there is an adjacent area not included in the analysis.

South Carolina veterans who were not excluded and yet received healthcare services at a VA facility not in South Carolina were still included in the analyses.

Receiving care at a South Carolina VA facility was not an exclusion criterion for that reason. The closet VA healthcare facility may be located across the state line for many border areas of the South Carolina. This occurrence does not impact the analyses in determining if residence-specific characteristics are associated with PrCA.

More importantly, the ZCTA was selected as the geographical unit of analysis because it was the most appropriate choice to assess the spatial impact of prostate cancer epidemiology for two reasons. The first is that it was the smallest of the two geographical units available (the other being county). The second is a consequence of the first reason. The more homogenous the population was, the more likely to differentiate between ZCTA's in the models and therefore identify ZCTA's with higher and lower than expected risks for PrCA. As the geographical unit area increases, the population becomes more heterogenous in characteristics that could not be assessed. This increases the likelihood of introducing an unaccounted for directional bias.

No South Carolina ZIP code maps were presented identifying crude counts, crude rates, or ZCTA specific risk estimates because ten of the 413 ZIP codes (2.4%) had, at most, five VA patients. Therefore, any presentation identifying location even at the aggregate level would violate the Health Insurance Portability and Accountability Act of 1996.[80] The research approach undertaken should not be discounted due to the limitation of what can be presented in the public domain. This research has the most benefit if used as an internal resource in identifying ZIP codes where the risk for a specific disease is above or below what is expected after controlling for known risk factors. In this regard, additional resources may be better directed toward

understanding what is driving a higher-than-expected disease risk in a specific area. This, in turn, could lead to targeting interventions by public health researchers that have the potential to benefit a specific community in need to better improve their health outcomes. Furthermore, clinicians can use this research in factoring in a patient's residence as a potential risk factor for specific diseases.

Finally, in addition to known risk factors for PrCA (race and marital status), we identified the unique factors such as area of residence and waiting to begin VA healthcare services until later in life as risk factors for incident PrCA. We demonstrated that location-specific characteristics as a risk factor for PrCA can be evaluated and accounted for at three levels; the individual level, community level, and the spatial autocorrelation among the ZIP codes. Expanding on the known risk factors for PrCA will allow clinicians to better assess a man's likelihood for developing PrCA thereby improving the preventive care he receives.

Table 3.1. Descriptive Statistics of South Carolina veteran population who seek healthcare at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria

Categorical Variables	N = 107,753 (%)
Prostate Cancer	
Yes	3,736 (3.47)
No	104,017 (96.53)
Age at First VA Visit (years)	
< 40	8,259 (7.66)
40 – 50	21,848 (20.28)
>50 – 60	37,052 (34.39)
>60 – 70	40,594 (37.67)
Race	
White	70,703 (65.62)
Black	37,050 (34.38)
Marital Status	
Married	67,173 (62.34)
Previously Married	24,520 (22.76)
Never Married/Unknown	16,060 (14.9)
Preventive Care Visit	
Yes	104,644 (97.11)
No	3,109 (2.89)
Digital Rectal Exam	
Yes	104,023 (96.54)
No	3,730 (3.46)
Distance Traveled (Miles)	
0 – 25	31,200 (28.96)
>25 – 55	20,515 (19.04)
>55	56,038 (52.01)

Table 3.2. Univariate χ2 analysis comparing all prostate cancer patients to non-prostate cancer patients for all categorical variables among the South Carolina veteran population who seek healthcare at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria

Risk Factors	All PrCA Patients (n = 3,736)	Non-Cancer Patients (n = 104,017)	p-value
Age of first VA visit (years)			<0.001
< 40	31 (0.83%)	8,228 (7.91%)	
40 – 50	490 (13.12%)	21,358 (20.53%)	
>50 – 60	1,684 (45.07%)	35,368 (34.00%)	
>60 – 70	1,531 (40.98%)	39,063 (37.55%)	
Race*			<0.001
White	1,662 (44.49%)	69,041 (66.37%)	
Black	2,074 (55.51%)	34,976 (33.63%)	
Marital Status			<0.001
Married	2,181 (58.38%)	64,992 (62.48%)	
Previously Married	1,222 (32.71%)	23,298 (22.40%)	
Never Married/Unknown	333 (8.98%)	15,727 (15.12%)	
Distance Traveled (Miles)			<0.001
0 – 25	1,071 (28.67%)	30,129 (28.97%)	
>25 – 55	910 (24.36%)	19,605 (18.85%)	
>55	1,755 (46.98%)	54,283 (52.19%)	

Table 3.3. Odds ratios and 95% credible intervals for multivariate models comparing prostate cancer patients to non-cancer patients among the South Carolina veteran population who seek healthcare at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria

Risk Factors	Model 1*	Model 2*	Model 3*	Model 4*	Model 5*	Model 6**	Model 7***
Age of first VA visit (years)							
< 40	Reference	N/A	Reference	Reference	N/A	Reference	Reference
40 – 50	5.60 (3.88, 8.36)	N/A	5.72 (4.06, 8.13)	5.62 (4.03, 7.61)	N/A	5.55 (3.94, 7.91)	5.63 (3.94, 8.39)
>50 – 60	13.38 (9.40, 19.93)	N/A	13.48 (9.67, 19.03)	13.14 (9.51, 17.67)	N/A	13.03 (9.33, 18.62)	13.11 (9.29, 19.53)
>60 – 70	13.60 (9.51, 20.25)	N/A	13.67 (9.79, 19.30)	13.36 (9.64, 17.90)	N/A	13.23 (9.45, 18.82)	13.31 (9.44, 19.81)
Black	3.07 (2.87, 3.29)	N/A	3.02 (2.81, 3.24)	2.96 (2.75, 3.18)	N/A	3.03 (2.82, 3.25)	2.98 (2.77, 3.20)
Marital Status							
Married	Reference	N/A	Reference	Reference	N/A	Reference	Reference
Ever Married	1.50 (1.40, 1.61)	N/A	1.49 (1.38, 1.60)	1.47 (1.37, 1.58)	N/A	1.48 (1.38, 1.59)	1.47 (1.37, 1.58)
Never Married /Unknown	0.57 (0.50, 0.64)	N/A	0.56 (0.49, 0.63)	0.55 (0.49, 0.62)	N/A	0.56 (0.50, 0.63)	0.55 (0.49, 0.62)
Distance traveled from the most listed ZIP code (Miles)							
0 – 25	Reference	N/A	Reference	Reference	N/A	Reference	Reference
>25 – 55	1.21 (1.10, 1.32)	N/A	1.20 (1.07, 1.34)	1.08 (0.96, 1.22)	N/A	1.16 (1.01, 1.33)	1.06 (0.92, 1.21)
>55	0.96 (0.89, 1.04)	N/A	0.95 (0.86, 1.04)	0.88 (0.80, 0.98)	N/A	0.92 (0.80, 1.06)	0.86 (0.75, 0.98)
Percent of population in ZIP code living in poverty							
0 – 7.96	N/A	N/A	N/A	Reference	N/A	N/A	Reference
>7.96 – 15.86	N/A	N/A	N/A	1.04 (0.87, 1.24)	N/A	N/A	0.94 (0.78, 1.14)
>15.86	N/A	N/A	N/A	1.24 (1.04, 1.48)	N/A	N/A	1.13 (0.93, 1.35)
SoVI®							
Low	N/A	N/A	N/A	Reference	N/A	N/A	Reference
Medium	N/A	N/A	N/A	1.27 (1.05, 1.56)	N/A	N/A	1.35 (1.11, 1.65)
High	N/A	N/A	N/A	1.27 (1.01, 1.58)	N/A	N/A	1.39 (1.11, 1.75)
Median Odds Ratio	N/A	1.37 (1.30, 1.44)	1.20 (1.15, 1.26)	1.17 (1.12, 1.22)	1.23 (1.16, 1.31)	1.17 (1.12, 1.22)	1.14 (1.11, 1.19)
DIC	30,539	32,079	30,430	30,430	32,067	30,411	30,408

Model 1: Individual Level Risk Factors; Model 2: Nested Null Model; Model 3: Nested Model with Individual Level Risk Factors; Model 4: Full Nested Model; Model 5: Spatial Null Model; Model 6: Spatial Model with Individual Level Risk Factors; Model 7: Full Spatial Model

N/A: not applicable; *All 15,000 iterations after a burn-in of 5,000 iterations were evaluated; **Every fifth iteration of 25,000 iterations after a burn-in of 15,000 iterations were evaluated; ***All 25,000 iterations after a burn-in of 15,000 iterations were evaluated

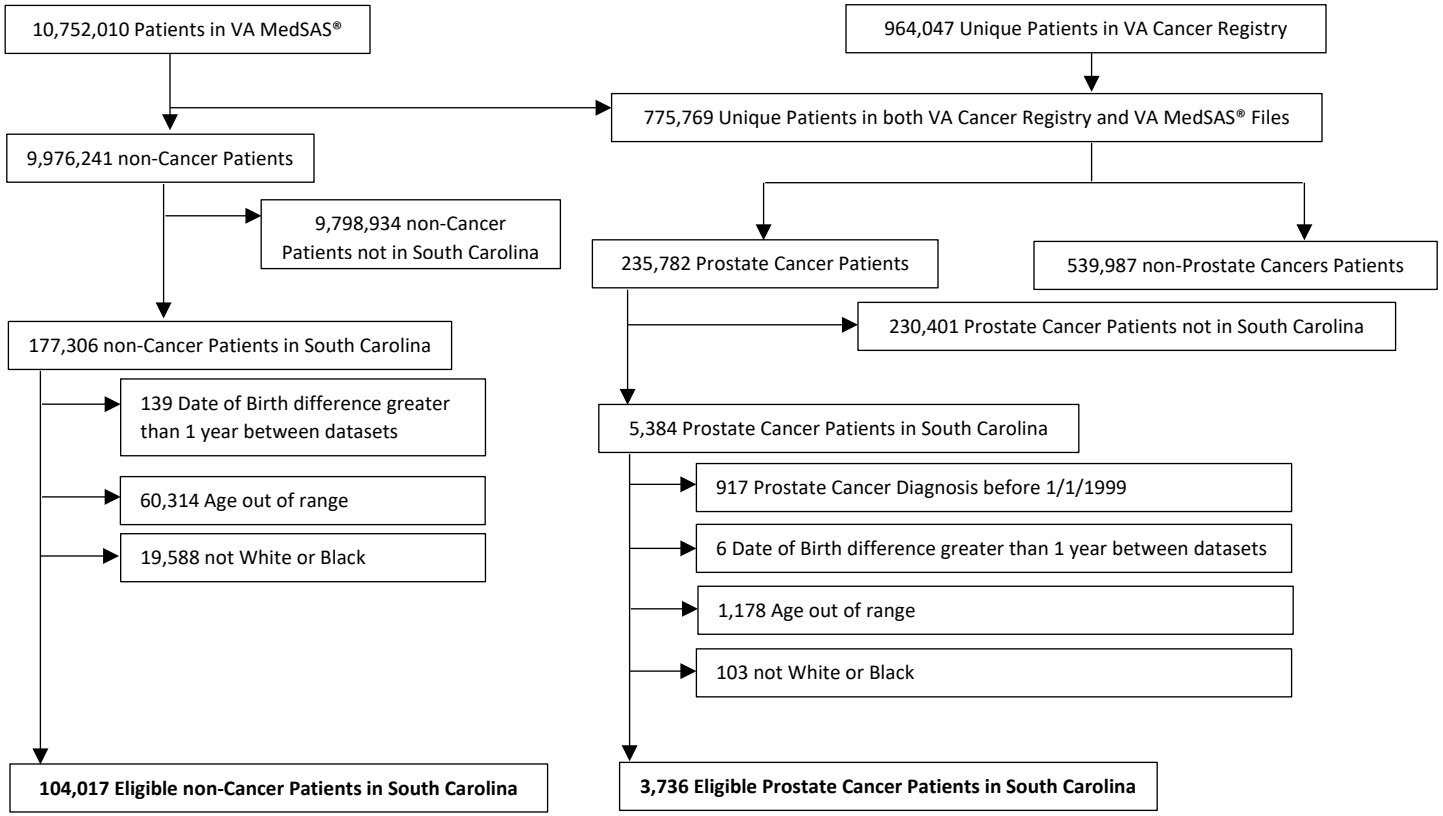

Figure 3.1. STROBE Diagram for the generation of the analytical cohort comparing prostate cancer patients to non-cancer patients among the South Carolina veteran population who seek healthcare at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria

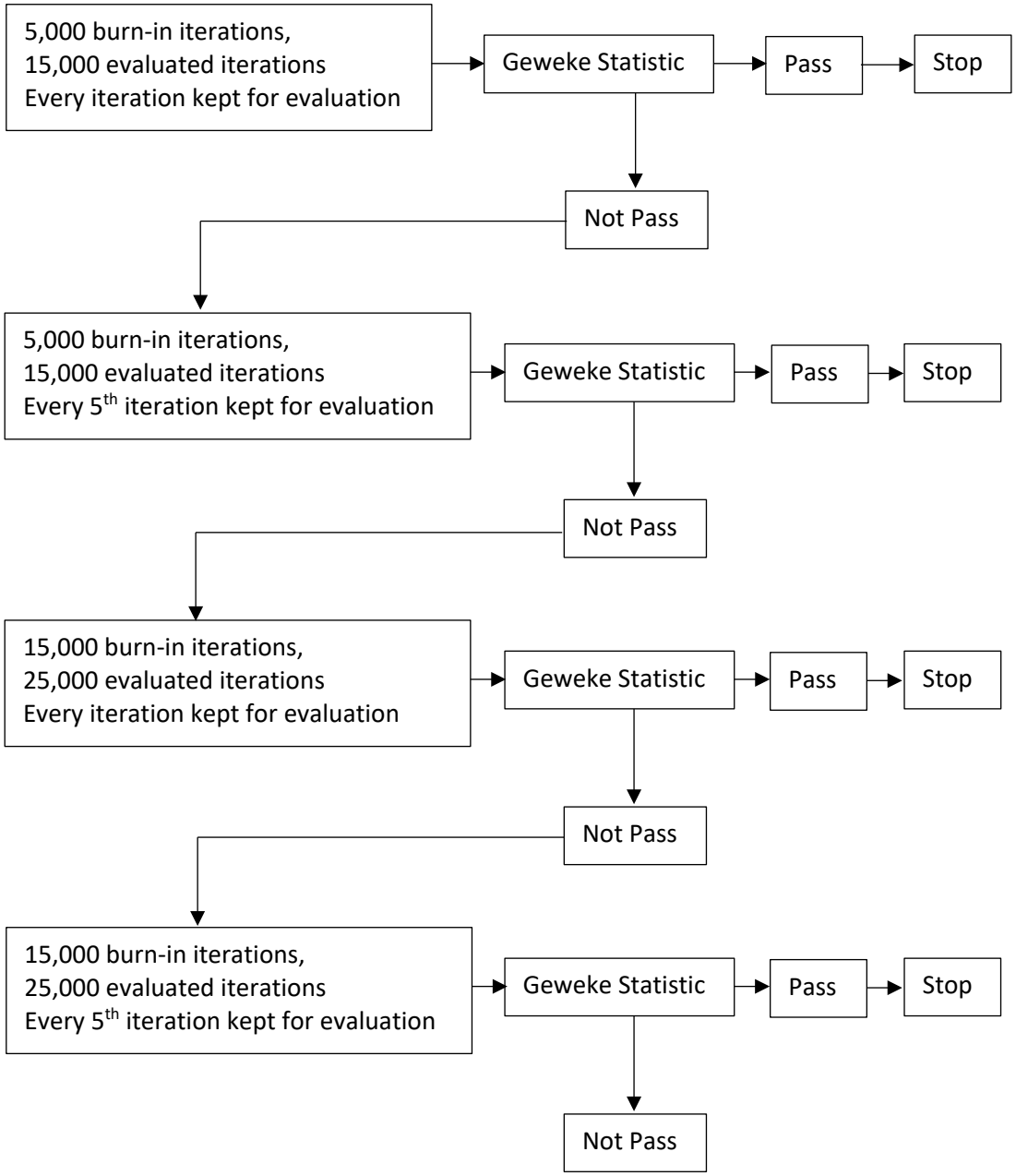

Figure 3.2. Systematic plan for Bayesian model evaluation of the analytical cohort comparing prostate cancer patients to non-cancer patients among the South Carolina veteran population who seek healthcare at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria

CHAPTER 4

NEIGHBORHOOD INFLUENCES ON PROSTATE CANCER MORTALITY AMONG SOUTH CAROLINA VETERANS

4.1 Abstract

Background: Established risk factors for prostate cancer (PrCA) mortality include age at diagnosis, TNM tumor stage, and Gleason score, family history and race. Only recently have factors related to PrCA patients' residence on PrCA mortality begun to receive serious attention. We used electronic medical records from the Veterans Health Administration (VHA) to assess geographical factors and established risk factors for PrCA mortality.

Methods: US VHA electronic medical records from January 1999 to December 2015, were used to identify 3,073 PrCA patients residing in 346 ZIP codes within South Carolina (SC). Data were analyzed using a Bayesian multivariate conditional autoregressive modeling framework with mortality as the outcome, while accounting for individual-level factors, area-level factors (i.e. ZIP code tabulation area-level), spatial random effects and spatial autocorrelation.

Results: Residing >55 miles from the VHA facility was associated with a 25% elevated risk of dying, [OR_{adj}: 1.25 (1.04, 1.50)], after accounting for age at diagnosis (1.65-fold and 2.34-fold increases in men >60 – 70 years and those >70 years, respectively). Gleason score of ≥7 and TNM distant stage were important risk factors in

the fully adjusted model [OR_{adj}: 1.51 (1.26, 1.82)] and [OR_{adj}: 4.71 (3.14, 6.92)], respectively. Accounting for ZIP code tabulation area (ZCTA) of residence as a random effect in the multivariate model resulted in the best-fitting model.

Conclusions: While accounting for known risk factors for PrCA, including age at diagnosis, Gleason score, TNM stage, and marital status, we found that residing more than 55 miles from the VA facility increased the risk for PrCA mortality. Outreach for screening, education and care coordination may be needed for veterans in these areas.

4.2 Introduction

In most men, prostate cancer (PrCA) is a relatively indolent disease, with an average mortality-to-incidence ratio of just 0.17 (i.e., 17% of all PrCA cases died of the disease) among White Americans in the US. However, there are considerable racial differences. For example, in South Carolina Black-American men have an MIR of 0.26 (95% CI: 0.241-0.277) while their White-American counterparts have an MIR slightly lower than the national average (MIR = 0.16, 95% CI: 0.155-0.277).[45] Because PrCA is the most common cancer among men[2] and large racial disparities exist with respect to disease aggressiveness[2], there is a clear clinical and scientific imperative to understand what drives PrCA burden. Risk factors for PrCA mortality include socioeconomic status (SES)[6,9,11,81], residing in rural localities[8], PrCA tumor Gleason score, and stage at diagnosis[81]. SES is a broad category that encompasses such factors as income and education level. Besides racial differences, PrCA mortality displays a spatial variation in the United States.[38] The PrCA mortality rate in the US in 2014 was 19.1/ 100,000

men/year .[82] This rate varied between the 50 states from a low of 14.4/100,000 men/year in Hawaii to a high of 24.7 in Idaho.[82]

Of all major racial groups in the United States, Blacks have the highest PrCA incidence and mortality. For example, in South Carolina, PrCA incidence in Blacks is about 80% higher than in Whites and mortality is over 2 ½ times higher.[2,45] The United States Veteran's Health Administration (VHA) provides a unique nationwide cohort of patients, namely veterans meeting certain eligibility criteria (e.g. served in the US armed forces and received a discharge other than dishonorable)[83]. SES factors such as education level, race and income that are typically associated with health outcomes in the United States are minimized within the VHA system since VHA care is not based on a fee-for-service arrangement and all eligible veterans have equal access to services. Veterans who receive VHA healthcare therefore become an ideal cohort to assess the impact of SES factors typically associated with access to healthcare or related factors, but have no apparent etiological link to that disease. Furthermore, recent developments in epidemiological modeling techniques allows public health researchers to evaluate multiple risk factors across multiple levels (e.g., individual and community) while also accounting for specific types of dependence in the data (e.g. spatial autocorrelation). In addition to these advantages, it is important to note that the VA system over-represents Blacks in the population[84], and that South Carolina has one of the highest proportions (31%) of Black residents in the country[85].

This study seeks to identify individual and community-level risk factors associated with PrCA mortality, while also accounting for the unique (random effects)

and interdependent (autocorrelation) geographic context. The VHA system's electronic medical records provide a nationwide comprehensive medical history of all veterans who receive care at all their facilities. We focused on South Carolina because its PrCA mortality rate is higher than the national rate[82] yet is a being a smaller state by area with a large veteran population residing throughout the entire state and where the percent of Black veterans as part of the larger veteran population in SC is larger than the percent of Blacks in the general population.[4,5] Because of its unique social and economic history, including historical context of slavery, South Carolina has the added advantage that Blacks live in all parts of the state, including rural areas.

4.3 Methods

4.3.1 Study Design

This study employed a retrospective cohort design. The timeframe was from 01 January 1999 to 31 December 2015. The unit of investigation was the US Census Bureau-defined ZIP code tabulation areas (ZCTA), which consists of aggregated Zone Improvement Plan (ZIP) codes developed by the US Postal Service.[63] All geographical related information was either linked to the shapefile (i.e. ZIP code patient level information linked to ZCTA's in the shapefile), directly obtained at the ZCTA level (i.e. neighborhood-level information), or directly obtained from the 2015 ZCTA shapefile from the US Census Bureau (i.e. individual ZCTA boundaries).[64]

4.3.2 Data Sources

Data were obtained from the United States Department of Veteran Affairs and the United States Census Bureau. From the Unites States Department of Veteran Affairs

the following datasets were used: all MedSAS® datasets, Master Vital Status dataset, Mini Vital Status dataset, and Primary Oncology dataset from the VA Central Cancer Registry (VACCR).[52,53] The VA Master Vital Status and VA Mini Vital Status datasets provided patient-level information (i.e., each unique VA patient is listed only once in these files) while the VA MedSAS® datasets provided visit-level information (i.e., with ≥1 record/subject).[52,53] From the United States Census Bureau the following datasets were used: United States 2010 Decennial Census, the 2007 – 2011 Five-Year American Community Survey, and the 2015 ZCTA shapefile for the US.[4,5,64]

4.3.3 Exclusion Criteria

Subjects were excluded based on the following criteria (see Figure 4.1 for the STROBE Diagram establishing the analytic cohort)[65]: data in the Primary Oncology Dataset from the VA Cancer Registry could not be linked to MedSAS® datasets[52], a non-prostate cancer diagnosis in the Primary Oncology Dataset from the VA Cancer Registry, prostate cancer diagnosis in the Primary Oncology Dataset from the VA Cancer Registry prior to January 1, 1999, date of birth differs by more than 365 days between VA MedSAS® datasets[52] and VA Vital Status datasets[53], age during the timeframe did not wholly or partially fall between 40 and 70 years old, missing ZIP code information, racial classification other than White or Black, not residing in a South Carolina ZIP code, not having a Gleason Score listed in the VACCR, missing PrCA diagnosis date, and negative time to event due to data entry error (e.g. diagnosis date occurred after date of death).

Overall, 964,047 unique subjects from a total of 1,159,188 (83.17%) unique cancer diagnoses were included in the Primary Oncology Dataset. Of these subjects, 80.47% could be linked to unique patients in the MedSAS® datasets[52].

There were 235,782 unique PrCA cases in the Primary Oncology Dataset that could be linked to the VA MedSAS® Files[52]. There were 230,401 (97.72%) PrCA cases were exclude due to not residing in a South Carolina ZIP code. Of the 5,384 PrCA cases residing in a South Carolina ZIP code; 1,648 (30.61%) were ineligible due to diagnosis prior to 01/01/1999, date of birth discrepancy, age at first VA visit not satisfied, missing ZIP code, and/or not being White or Black,, 659 (12.2%) did not have a Gleason Score, three (0.06%) did not have a diagnosis date, and three (0.06%) had a negative time to event. The final analytical cohort consisted of 3,073 eligible South Carolina veterans with PrCA (Figure 4.1).

4.3.4 South Carolina Geographical Area

South Carolina veterans with PrCA that were not excluded resided in 346 of South Carolina's 424 ZIP codes (81.6%). The 2015 ZCTA shapefile from the US Census Bureau was used South Carolina ZCTA's were in turn used to link the selected ZCTA's to ZIP codes with identical designations in the VA MedSAS® Files.[52,63]

4.3.5 Time to Event Outcome

Among subjects who died due to any cause, time in days between PrCA diagnosis and death was calculated for each subject using data from the VA Mini Vital Status dataset and the VACCR.[53] Among subjects who were alive at the end of the study timeframe, time in days was calculated between PrCA diagnosis and the end of the

study timeframe (i.e. 12/31/2015). Subjects were considered right censored if they were alive at the end of the study.

4.3.6 Individual-Level Independent Variables

The date of the first VA visit for each subject was obtained from the MedSAS® datasets and the date of birth (DOB) was obtained from the VA Mini Vital Status dataset.[52,53] Four age strata variables were created: age <40, age from 40 – 50, age from >50 – 60, age from >60 – 70 years. The information located within the VA Master Vital Status file for race and ethnicity were classified into White, Black, Hispanic, Asian, Hawaiian/Pacific Islander, and Native/Alaskan America, Other, and Unknown.[53] Subjects other than White or Black racial category were exclude because they accounted for 2.53% of the PrCA. The marital status field at the time of diagnosis in the Primary Oncology Dataset was used for PrCA cases. Marital status was categorized as Married, Previously Married (Divorced, Widowed, Separated), and Never Married/Unknown. The date of the first VA visit for each subject was obtained from the MedSAS® datasets and the date of PrCA diagnosis was obtained from the VACCR.[52] Four age strata variables were created: age ≤50, age from >50 – 60, age from >60 – 70, age from >70 years. Gleason Scores for each subject was obtained from the VACCR.

Gleason Scores were stratified into a dichotomous variable where a Gleason Score ≥7 was classified as high-risk and a Gleason Score of <7 was classified as low-risk. TNM Stages for each subject was obtained from the VACCR. TNM Staging was classified into Local, Regional, Distant, and Unknown. TNM Staging was also dichotomized into Distant and all other staging. The MedSAS® datasets were used to identify PrCA patients

with an International Classification of Disease version 9 Procedure Code for radical prostatectomy (i.e. 60.5).[52] Radical prostatectomy was dichotomized into having received the procedure and not having received the procedure. The VACCR dataset was used to identify and PrCA patients who received hormone therapy for their treatment. PrCA patients were dichotomized into having received hormone therapy and not having hormone therapy. The VACCR dataset was used to identify PrCA patients who received radiation therapy for their treatment. PrCA patients were dichotomized into having received radiation therapy and not having radiation therapy. The VACCR dataset was used to identify PrCA patients who received chemotherapy for their treatment. PrCA patients were dichotomized into having received chemotherapy and not having chemotherapy.

The modified Elixhauser Comorbidity Index was created excluding cancer diagnoses.[86] Of the 32 comorbidities, two were excluded (any malignancy and metastatic solid tumor). The modified Elixhauser Comorbidity Index was used as a continuous variable indicating the number of unique comorbidities with a possible range of 0 to 30. A comorbidity was required to be listed on at least two distinct VA visits in the VA MedSAS® datasets.[52]

The unique VA facility numerical codes in the MedSAS® datasets were used to identify which facility a subject visited.[52] The number of different VA facilities used during the study timeframe, the most frequented VA facility during the study timeframe, and the number of times each subjected visited their most visited VA facility during the study timeframe were determined. The ZIP code listed the most times for

each subject's total visits in the MedSAS® datasets was used for all ZIP code related information for each subject.[52] For PrCA cases, the ZIP code at diagnosis was obtained from the Primary Oncology dataset. For PrCA cases, the ZIP code at diagnosis was used for cases without ZIP code information from the MedSAS® datasets (n = 6,784).[52]

There were 131 unique VA facilities (i.e. VA hospitals and VA Community Based Outpatient Clinics) determined from VA MedSAS® datasets.[52] The distance between the most frequented VA facility and ZCTA-linked ZIP code listed the most times was represented as a continuous variable and obtained using the MedSAS® datasets.[52] This ZIP code information was linked to the 2015 ZCTA shapefile from the US Census Bureau.[64] Using the shapefile, a contiguity straight line origin-to-destination matrix distance matrix between VA facilities and ZCTA code centroids to determine the distance traveled by a veteran to his most frequented VA facility.[52,64] The distance between the most frequented VA facility and the ZCTA-linked ZIP code listed the most times was categorized into: 0 – 25 miles, >25 – 55 miles, and >55 miles. These cutoffs were selected a priori to represent how many miles can approximately be driven within 30 minutes, 1 hour, and more than 1 hour.

4.3.7 Neighborhood-Level Independent Variables

The Social Vulnerability Index (SoVI®): The SoVI®, initially developed by the Hazards and Vulnerability Research Institute (HVRI) at the University of South Carolina, is a composite measure of neighborhood-level factors obtained from publicly available population based datasets of factors associated with the health of individuals within those neighborhoods.[58] The SoVI® uses ZCTA measurements obtained from the 2010 US

Census and the 2007 – 2011 5-Year American Community Survey.[4,5] The SoVI® is a relative measure represented on a continuous scale from negative infinity to positive infinity in which the greater a geographical unit's SoVI® score is the less prepared for, respond to, and recover from a disaster that area is compared to a geographical unit with a lower SoVI® score.[66] A SoVI® score was assigned for each subject's most frequently listed ZIP code. The SoVI® measurements were stratified into 3 categories using cutoffs of 1 standard deviations: Low (-∞, -2.04], Medium (-2.04, 2.79], and High (2.79, ∞).

Specific Zip Code Level Factors: The 2007 – 2011 Five-Year American Community Survey was used to determine the percent of people within each ZIP code with at least a college degree with tertile cutoffs of [0%, 12.86%], [12.86%, 22.60%], [22.60%, 100%], percent of people within each ZIP code living in poverty with tertile cutoffs of [0%, 7.96%], [7.96%, 15.86%], [15.86%, 100%], percent of Black Americans within each ZIP code with tertile cutoffs of: [0%, 0%], [0%, 2.73%], [2.73%, 100%], and percent of people within each ZIP code that were at least 65 years of age with tertile cutoffs of: [0%, 11.95%], [11.95%, 16.92%], [16.92%, 100%].[5]

4.3.8 Statistical Analyses

Descriptive statistics were calculated for the analytical cohort. Kaplan-Meier curves were created to check the proportional hazards assumption for each variable. Chi-square tests and Fisher's exact tests for categorical variables, and t-tests for continuous variables which used Gleason Score categories at the dependent variable were created.

Frequentist multivariate models using SAS® were first developed to inform key steps in development and evaluation of the final models.[59] These models were developed for all individual level risk factors, and nested models incorporating both individual- and neighborhood-level risk factors with patient ZCTA-linked ZIP code residence as a random effect were created for model selection purposes. Bayesian models were developed for the null (empty) models; individual-level models; nested models incorporating both individual, area-level risk factors, and accounted for patient ZCTA-linked ZIP code residence as a random effect; and the multivariate conditional autoregressive models incorporating both individual, neighborhood-level risk factors, accounted for patient ZCTA residence as a random effect, and accounted for the spatial autocorrelation between ZCTA-linked ZIP codes using queen contiguity based spatial weights (i.e. all bordering ZCTAs for a given ZCTA.will have a weight of 1 while all other ZCTAs will have a weight of 0).[34,36,37]

All categorical variables were evaluated using created dummy variables. Any independent individual-level variables where 95% confidence intervals could not be computed or extremely large point estimates with extremely wide 95% confidence intervals were excluded since those variables had almost all observations in one stratum. Independent individual-level variables not excluded were used in a manual backward stepwise approach to generate the final multivariate individual-level model were the added variable remained if it changed the risk factor estimates of the other variables by at least 10%. Multivariate models were created including and excluding PrCA stage.

A manual backward stepwise approach was used to determine the final nested-level model. Each neighborhood level independent risk factor was evaluated separately with the final multivariate individual-level model. The model with the lowest Akaike information criterion (AIC) was selected as the initial nested-level model.[67] Subsequently, the neighborhood factor with the next lowest AIC was added to the initial nested-model. If any risk factor estimates changed by more than 10% than the additional neighborhood factor remained. This was repeated for all neighborhood level factors.

The median odds ratio (MOR) was calculated for the unstructured variance when the model used ZIP codes as a random effect and was calculated for the structured variance when the model accounted for the spatial autocorrelation between ZCTA-linked ZIP codes. The median odds ratio uses the specific aforementioned variances in the model and calculates a statistic that can be interpreted as the median difference in odds between the ZCTA-linked ZIP code with the highest compared to the lowest risk for two individuals with the same evaluated risk factors (i.e. the MOR is the risk estimate for the unexplained specific variation in the model).[68] Therefore, the model with the best fit will also have the lowest MOR because risk factors in the model accounting for more of the variance in the outcome.

A Global Moran I's Test was conducted on the ZCTA-linked ZIP code level residuals of the final models to determine existence of spatial autocorrelation.[37] If the Global Moran I's Test was statistically significant, indicating the presence of spatial autocorrelation, then models accounting for spatial autocorrelation would be

constructed using risk factors included in the final nested-level models. If the Global Moran I's Test was statistically nonsignificant, indicating no presence of spatial autocorrelation in the data, then the nested-level model would become the final overall model for that comparison. If the Global Moran I's Test was statistically nonsignificant for the nested null model then the individual only level would be the final model.

Based on the frequentist models, the final Bayesian models were developed. A Bayesian framework was selected for all final models because we assumed the factors associated ZIP code residence are not independent and random between neighboring ZIP codes and therefore will have a structured variance that needs to be accounted for. Incorporating a conditional autoregressive (CAR) component into the models is the most popular way to account for this structured variance. The most appropriate framework when incorporating a CAR component in the models is to use a Bayesian framework. It was decided prior to modeling building that since models incorporating a CAR component would have to be conducted in a Bayesian framework then all other final models would also be conducted within a Bayesian framework so as to compare between final models.

All Bayesian models used uninformed priors. The initial statistical inference from each Bayesian model was based on 15,000 iterations (i.e. samples) after 5,000 burn-in period. The convergence of the sample chains was evaluated using the Geweke diagnostic, which compares the first 10% of kept iterations to the last 50% of kept iterations.[69] The mean hazard ratio (HR) value was used as the point estimate for each individual and area level risk factors as well as the MOR estimates. The 2.5% and 97.5%

iteration cutoffs were used as the ranges for the 95% credible intervals (CrI) for each risk factor.

Bayesian models created were compared to each other using the Deviation Information Criteria (DIC).[60] The model with the lowest DIC was selected as the best-fitting model for that comparison in that analytical cohort.

4.3.9 Logistical Aspects

All data storage, management, and analyses were conducted with the VA Informatics and Computing Infrastructure (VINCI) servers. All data management was done using Microsoft SQL Server. All univariate and frequentist modeling analyses were conducted using SAS® version 9.4 software.[59] All Bayesian modeling was conducted using WinBUGS® version 1.4.3 and R version 3.3.2.[60,61] All evaluation of Bayesian modeling was conducted using R.[61] All statistical tests used an α-level of 0.05. This project had Institutional Review Board Approval from the University of South Carolina (Pro00036431) and the William Jennings Bryan Dorn VA (10404).

4.4 Results

There were 3,073 PrCA patients who met the inclusion criteria, and therefore comprise the analytical cohort. Of those, 534 (17.4%) died during the study timeframe. (Table 4.1) A slight majority of PrCA patients had a Gleason Score of at most a 6; 1,578 (51.35%) while a large majority of PrCA patients had a localized tumor TNM stage; 2,803 (91.21%) (Table 4.1). A majority of PrCA patients were Black (1,736 [56.49%]), married (1,767 [57.50%]), diagnosed with PrCA between > 60 – ≤ 70 (1,742 [56.69%]), travelled at most 55 miles to their most frequented VA facility (1,588 [51.68%]), did not receive

hormone therapy (2,268 [73.80%]), did not receive chemotherapy (3,067 [99.80%]), did not receive radiation therapy (1,703 [55.43%]), and did not have a radical prostatectomy (3,068 [99.84%]).

4.4.1 Univariate Analyses

There was a statistically significant difference in the average duration of the time accrued during the study timeframe between those that died (4.96 years) and those who survived (5.89 years) (Table 4.2). PrCA patients who died were statistically more likely to have a higher Modified Elixhauser Comorbidity Index (4.67 vs. 3.76), a higher Gleason score (60.11% vs. 46.27%), and more likely to have a TNM tumor stage of distant (5.81% vs. 0.87). (Table 4.2) PrCA patients who died were statistically less likely to have had radiation therapy (40.45% vs. 45.45%) but statistically more likely to have had hormone therapy (35.58% vs. 23.59%) (Table 4.2). PrCA patients who died were statistically less likely to have been diagnosed at a younger age (23.22% vs. 30.92%), less likely to have been married at the time of diagnosis (51.87% vs. 58.68%), and statistically more likely to travel more than 55 miles to their most frequented VA facility (55.06% vs. 46.91%) (Table 4.2). The proportional hazards assumptions was satisfied for all independent risk factors.

4.4.2 Multivariable Bayesian Modeling

Two sets of seven models of increasing complexity using a Bayesian framework were fitted (Tables 4.3 – 4.4). One set included TNM tumor stage at the time of diagnosis (Table 4.3) while the other did not TNM tumor stage (Table 4.4). No ZCTA level

risk factors resulted in statistically significant additions to the individual level models during the frequentist model building phase.

During the final Bayesian model building phase, all independent variables were significant for the first model in both sets. The second model was the nested null model where the subject's ZIP code was accounted for as random effects. The third model accounted for the individual-level factors as well as nesting the subjects in their ZIP codes as random effects. The fourth model was the null model for the spatial autocorrelation between ZIP codes. Model 5 accounted for the individual level risk factors and the spatial autocorrelation between ZIP codes. Model 6 was the nested null model that accounted for the random effects of the ZIP codes and the spatial autocorrelation between ZIP codes. The final model, Model 7, accounted for the individual level risk factors, the random effects of the nested ZIP codes, and the spatial autocorrelation between ZIP codes. For both sets of models, Models 4 – 7 had the highest DIC's (Tables 4.3 – 4.4).

Model 3 had the lowest DIC for both sets of models with Model 3 from the first set (i.e. including TNM tumor stage) having the lowest DIC between the Model 3's of the two sets. The individual level risk factors and the unstructured MOR for Model 3 in both sets were very similar. The final, best fitted model was Model 3 from the first set.

PrCA patients were at increased risk of death if they were diagnosed later in life as compared to earlier in life (50 – 60 years of age): between > 60 and 70 years of age (HR: 1.65 [95%CrI: 1.35 – 2.04]) and greater than 70 years (HR: 2.34 [95%CrI: 1.80 – 3.02]). PrCA were at increased risk of death if the tumor Gleason Score was at least 7 as

compared to at most 6 (HR: 1.51 [95%CrI: 1.26 – 1.82]), and a TNM stage of distant (HR: 4.71 [95%CrI: 3.14 – 6.92]) as compared to all other TNM staging. There was a protective association for PrCA patients receiving radiation therapy (HR: 0.66 [95%CrI: 0.54 – 0.81]); however, PrCA patients were at increased risk of death if they received hormone therapy (HR: 1.31 [95%CrI: 1.07 – 1.60]). Not being married increased the risk of death among PrCA patients: for previously married (HR: 1.39 [95%CrI: 1.16 – 1.67]), and for PrCA patients who have never been married or their marital status in unknown (HR: 1.57 [95%CrI: 1.13 – 2.14]). PrCA patients who traveled more than 55 miles to their most frequented VA facility of care were at increased risk (HR: 1.25 [95%CrI: 1.04 – 1.50]) as compared to PrCA patients who traveled at most 55 miles. Each additional comorbidity in the modified Elixhauser comorbidity index increased the risk of death (HR: 1.11 [95%CrI: 1.07 – 1.15]). The MOR for the unstructured variance of this model was 1.09 (95%CrI: 1.02 – 1.19) (Table 4.3). These results remained consistent when TNM tumor stage was removed from the model (Table 4.4).

4.5 Discussion

This study used an innovative approach to determine the potential impact of individual-level risk factors, area-level risk factors, and spatial autocorrelation on prostate cancer mortality within one modeling framework. The best fitting multivariate Bayesian model was Model 3 (Table 4.3), which controlled for individual level risk factors and accounted for ZCTA-linked ZIP code residence as a random effect. Our study identified an association between where a veteran lives and prostate cancer mortality.

We found that the ZCTA-linked ZIP code in which a veteran with PrCA resides had an effect on his mortality in two distinct ways. The first was that a veteran was at a 25% increased risk of death if the distance traveled to his most frequented VA facility was more than 55 miles from his residence (Model 3 Table 4.3). This approximates to roughly a one-hour commute one way. This distance variable may also be considered as a component of the rural-urban divide that starkly characterizes SC, where all VA medical centers are located in the few urbanized localities, within the state while the SC veteran population who receive VA healthcare reside in nearly all the ZIP codes of the state.

The second was that the ZCTA-linked ZIP code in which a veteran resides may increase the risk of mortality by approximately 9% (Model 3 Table 4.3). This is reflected in the median odds ratio (MOR) for the model. This MOR indicates that there are ZCTA-linked ZIP code-level factors that are not accounted for in the model which results in a cumulative median increased risk between the ZCTA-linked ZIP code with the highest mortality as compared to the ZCTA-linked ZIP code with lowest mortality for veterans with identical individual level factors that were measured. This result is of unique interest since several ZCTA-linked ZIP code level factors as well as the SoVI® (which is a composite measure of 30 ZIP code level factors) were evaluated during the model building phase of this study and were found to not be associated with PrCA mortality when the individual level factors were accounted for. This may indicate the potential for at least one unmeasured ZIP code-level factor that could influence PrCA mortality.

Additionally, each increase of the modified Elixhauser Comorbidity Index score increases the risk of mortality for veterans with PrCA by 11% (Model 3, Table 4.3). This is expected because the modified Elixhauser Comorbidity Index score reflects how sick a patient is with comorbid chronic diseases that also might affect mortality.

Our best-fitting model confirms established risk factors for PrCA mortality. These include increasing mortality risk with increased age[87], a Gleason score of at least 7[87], a TMN stage of distant[2], and a protective effect against mortality of being married[88] (Model 3 Table 4.3). Under further investigation, while veterans with PrCA were 4.71 times more likely to die with a TNM stage of distant; removal of TNM stage from the model did not alter: 1) the best-fitting model, 2) the remaining risk factors that comprise the best-fitting model, and 3) the risk estimates and 95% credible intervals of the remaining risk factors by a significant amount (Tables 4.3 – 4.4).

Furthermore, receiving radiation therapy was a protective effect while receiving hormone therapy was a risk factor of PrCA mortality. (Model 3 Table 4.3) These two results should be considered as confounding by indication; i.e. these therapies reflect the aggressiveness of the PrCA tumor which directly impacts patient mortality rather than the therapies themselves "increasing" or "decreasing" risk of mortality. Clinically, a more aggressive PrCA tumor will be treated by hormone therapy while a less aggressive tumor will be monitored and/or begun on radiation therapy. It should also be noted that these estimates did not differ when TMN staging was removed from the models. (Model 3 Table 4.4) This occurrence may possibly reflect the clinical practice that for PrCA patients TNM staging is not as critical in dictating treatment.

Being Black is a known risk factor for PrCA mortality[2,45]; however, in all the multivariate Bayesian models, being black was not a risk factor in assessing PrCA mortality among South Carolina veterans seeking care at VA facilities. (Tables 4.3 – 4.4). This was initially seen in the Kaplan-Meier curves were there was not a statistically significant difference between White and Black. This was reinforced in the univariate analyses (Table 4.2). One possible explanation is that while being Black is a risk factor for developing PrCA[2,45]; after diagnosis, whatever latent risk factors in racial categorization that impact PrCA mortality are 1) societal in nature and not biological and 2) those factors are minimized in SC veterans with PrCA due to equalizing access to healthcare in the VHA population. This observation should further be replicated again in veteran populations who receive VA care in other states.

4.5.1 Strengths and Limitations

Limitations exist in the use of electronic medical records for retrospective observational studies.[74] However, the VA electronic medical records have been extensively used for such studies.[75-78]

In regards to this study, it is most problematic when matching ZCTA's and ZIP codes for areas without populations.[79] The VHA electronic medical records provides patient ZIP code information while the US Census provides shapefiles using ZCTA's. While ZCTA's approximate the shape of ZIP codes; ZCTA's are not a perfect match for ZIP codes and ZCTA's are not updated when ZIP codes change shape or new ZIP codes are added. However, limiting the population under investigation to South Carolina, which is a small, stably-populated state as compared to the rest of the nation should mitigate

the boundary issue (i.e. changing ZIP code shapes) as well as the addition of new ZIP codes. Therefore, the only areas impacted by boundary issues are those ZCTAs bordering the states of Georgia and North Carolina, which is always present in all spatial analyses where there is an adjacent area not included in the analysis, which is known as the edge effect.[89]

The 2015 ZCTA shapefile for South Carolina was chosen because it was created during the last year of the study timeframe.[64] However, the study timeframe was wide, ranging from 1999 to 2015. ZCTA shapefiles for other years were excluded because utilizing the most recent shapefile possible would also account for additions in SC ZCTA's but not deletions, which are rarer then additions. Ultimately, conducting a spatial analysis over multiple years using geographical units that have the potential to change will continue to pose such issues. What needs to be considered is how those changes impact the study. For this reason, as was previously mentioned, SC has a small and stable population, which would minimize any changes (both major and minor) in the shapes of ZCTA's across the study timeframe.

South Carolina veterans who did not meet the exclusion criteria and seek healthcare at a VA facility not in South Carolina were still included in the analyses. Receiving care at a South Carolina VA facility was not an exclusion criterion for that reason. The closet VA healthcare facility may be located across the state line for many border areas of the South Carolina. This occurrence does not impact the analyses in determining if residence-specific characteristics are associated with PrCA.

Interestingly, the use of ZCTA. which spread the distribution of 3,073 PrCA cases among 346 ZIP codes, is a possible explanation of why accounting for the spatial autocorrelation results in large DIC's because the spatial autocorrelation component of the model is only accounting for background variation.

More importantly, the ZCTA-linked ZIP code was selected as the geographical unit of analysis because it was the most appropriate choice to assess the spatial impact of prostate cancer epidemiology for two reasons. The first is that it was the smallest of the two geographical units available (the other being county). The second is a consequence of the first reason. The more homogenous the population was, the more likely to differentiate between ZCTA-linked ZIP codes in the models and therefore identify ZCTA-linked ZIP codes with higher and lower than expected risks for PrCA. As the geographical unit area increases, the population becomes more heterogeneous in characteristics that could not be assessed. This increases the likelihood of introducing an unaccounted-for directional bias.

Finally, all-cause mortality was selected instead of cause-specific (i.e. PrCA-specific) mortality because this was an initial evaluation in accounting for spatial factors in multivariate survival models for PrCA. The next phase will be to evaluate the data using cause-specific mortality as the outcome of primary interest.

In closing, we were able to 1) confirm established risk factors (e.g. age at diagnosis, marital status, Gleason Score, and TNM stage), 2) identify that modeling radiation and hormone therapy is probably confounding by indication, and 3)

demonstrate that where a veteran resides does influence PrCA mortality with the South Carolina veteran population.

Table 4.1. Descriptive statistics of South Carolina veteran population who were treated for prostate cancer at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria

Categorical Variables	N = 3,073 (%)
Mortality Status	
Alive	2,539 (82.62)
Dead	534 (17.38)
Prostate Cancer Gleason Score	
High Risk (Gleason Score ≥ 7)	1,495 (48.65)
Low Risk (Gleason Score < 7)	1,578 (51.35)
Prostate Cancer Stage Category 1	
Local	2,803 (91.21)
Regional	170 (5.53)
Distant	53 (1.72)
Unknown	47 (1.53)
Prostate Cancer Stage Category 2	
Distant	53 (1.72)
Non-Distant	3,020 (98.28)
Age at First VA Visit	
< 40	28 (0.91)
40 – 50	448 (14.58)
>50 – 60	1,451 (47.22)
>60 – 70	1,146 (37.29)
Race	
White	1,337 (43.51)
Black	1,736 (56.49)
Marital Status	
Married	1,767 (57.50)
Ever Married	1,035 (33.68)
Never Married /Unknown	271 (8.82)
Radical Prostatectomy	
Yes	5 (0.16)
No	3,068 (99.84)

Radiation Therapy		
Yes	1,370 (44.58)	
No	1,703 (55.43)	
Chemotherapy		
Yes	6 (0.20)	
No	3,067 (99.80)	
Hormone Therapy		
Yes	805 (26.20)	
No	2,268 (73.80)	
Distance Category 1 (miles)		
≤ 25	860 (27.99)	
> 25 – 55	728 (23.69)	
> 55	1,485 (48.32)	
Distance Category 2 (miles)		
≤ 55	1,588 (51.68)	
> 55	1,485 (48.32)	
Age at Prostate Cancer Diagnosis Category 1 (years)		
≤ 50	76 (2.47)	
> 50 – 60	833 (27.11)	
> 60 – 70	1,742 (56.69)	
> 70	422 (13.73)	
Age at Prostate Cancer Diagnosis Category 2 (years)		
≤ 60	909 (29.58)	
> 60 – 70	1,742 (56.69)	
> 70	422 (13.73)	
Continuous Variable	Mean (Standard Deviation)	Median (Minimum, Maximum)
Time in Study (years)	5.73 (3.55)	5.17 (<0.01, 15.88)
Modified Elixhauser Comorbidity Index	3.92 (2.34)	4 (0, 14)

Table 4.2. Univariate comparisons between censored status and risk factors among South Carolina veteran population who were treated for prostate cancer at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria

Categorical Variables	Died; n = 534 (%)	Alive; n = 2,539 (%)	p-value
Prostate Cancer Gleason Score			
High Risk (Gleason Score ≥ 7)	321 (60.11)	1,174 (46.27)	<0.01
Low Risk (Gleason Score < 7)	213 (39.89)	1,365 (53.76)	
Prostate Cancer Stage Category 1			
Local	456 (85.39)	2,347 (92.44)	<0.01
Regional	39 (7.30)	131 (5.16)	
Distant	31 (5.81)	22 (0.87)	
Unknown	8 (1.50)	39 (1.54)	
Prostate Cancer Stage Category 2			
Distant	31 (5.81)	22 (0.87)	<0.01
Non-Distant	503 (94.19)	2,517 (99.13)	
Age at First VA Visit			
< 40	1 (0.19)	27 (1.06)	<0.01
40 – 50	38 (7.12)	410 (16.15)	
>50 – 60	210 (39.33)	1,241 (48.88)	
>60 – 70	285 (53.37)	861 (33.91)	
Race			
White	267 (50.00)	1,070 (42.14)	<0.01
Black	267 (50.00)	1,469 (57.86)	
Marital Status			
Married	277 (51.87)	1,490 (58.68)	0.01
Ever Married	207 (38.76)	828 (32.61)	
Never Married /Unknown	50 (9.36)	221 (8.70)	
Radical Prostatectomy			
Yes	0 (0)	5 (0.20)	0.38
No	534 (100)	2,534 (99.80)	
Radiation Therapy			
Yes	216 (40.45)	1,154 (45.45)	0.03

No	318 (59.55)	1,385 (54.55)	
Chemotherapy			
Yes	2 (0.37)	4 (0.16)	0.21
No	532 (99.63)	2,535 (99.84)	
Hormone Therapy			
Yes	206 (38.58)	599 (23.59)	<0.01
No	328 (61.42)	1,940 (76.41)	
Distance Category 1 (miles)			
≤ 25	130 (24.34)	730 (28.75)	<0.01
> 25 – 55	110 (20.60)	618 (24.34)	
> 55	294 (55.06)	1,191 (46.91)	
Distance Category 2 (miles)			
≤ 55	240 (44.94)	1,348 (53.09)	<0.01
> 55	294 (55.06)	1,191 (46.91)	
Age at Prostate Cancer Diagnosis Category 1 (years)			
≤ 50	9 (1.69)	67 (2.64)	<0.01
> 50 – 60	115 (21.54)	718 (28.28)	
> 60 – 70	295 (55.24)	1,447 (56.99)	
> 70	115 (21.54)	307 (12.09)	
Age at Prostate Cancer Diagnosis Category 2 (years)			
≤ 60	124 (23.22)	785 (30.92)	<0.01
> 60 – 70	295 (55.24)	1,447 (56.99)	
> 70	115 (21.54)	307 (12.09)	
Continuous Variables	Mean (95% Confidence Interval)	Mean (95% Confidence Interval)	
Time in Study (years)	4.96 (4.67, 5.23)	5.89 (5.75, 6.03)	<0.01
Modified Elixhauser Comorbidity Index	4.67 (4.45, 4.89)	3.76 (3.67, 3.84)	<0.01

Table 4.3. Hazard ratios and 95% credible intervals for multivariate models that included prostate cancer tumor classification among South Carolina veteran population who were treated for prostate cancer at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria

Risk Factors	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Age at Diagnosis							
50 – 60	Reference		Reference		Reference		Reference
>60 – 70	1.67 (1.34 – 2.08)		1.65 (1.35 – 2.04)		1.35 (1.25 – 1.47)		1.36 (1.25 – 1.48)
>70	2.36 (1.82 – 3.09)		2.34 (1.80 – 3.02)		1.31 (1.16 – 1.47)		1.32 (1.17 – 1.48)
Gleason Score ≥7	1.51 (1.24 – 1.82)		1.51 (1.26 – 1.82)		1.17 (1.08 – 1.26)		1.17 (1.08 – 1.26)
Distant Stage	4.72 (3.12 – 6.83)		4.71 (3.14 – 6.92)		1.86 (1.38 – 2.47)		1.87 (1.40 – 2.48)
Radiation Therapy	0.66 (0.55 – 0.80)		0.66 (0.54 – 0.81)		0.86 (0.79 – 0.93)		0.86 (0.79 – 0.93)
Hormone Therapy	1.31 (1.07 – 1.60)		1.31 (1.07 – 1.60)		0.94 (0.86 – 1.03)		0.94 (0.86 – 1.03)
Marital Status							
Married	Reference		Reference		Reference		Reference
Previously Married	1.40 (1.17 – 1.68)		1.39 (1.16 – 1.67)		1.14 (1.06 – 1.23)		1.14 (1.06 – 1.23)
Never Married/Unknown	1.58 (1.15 – 2.14)		1.57 (1.13 – 2.14)		1.24 (1.08 – 1.40)		1.24 (1.09 – 1.41)
Travelled > 55 Miles	1.25 (1.06 – 1.46)		1.25 (1.04 – 1.50)		0.99 (0.91 – 1.09)		1.00 (0.91 – 1.11)
Modified Elixhauser Comorbidity Index	1.11 (1.08 – 1.15)		1.11 (1.07 – 1.15)		0.97 (0.96 – 0.99)		0.97 (0.96 – 0.99)
Median Odds Ratio (unstructured variance)	NA	1.08 (1.02 – 1.18)	1.09 (1.02 – 1.19)	NA	NA	1.02 (1.01 – 1.03)	1.00 (1.00 – 1.01)
Median Odds Ratio (structured variance)	NA	NA	NA	1.03 (1.01 – 1.05)	1.02 (1.00 – 1.04)	1.03 (1.01 – 1.05)	1.02 (1.00 – 1.04)
Deviance	10,821.60	11,037.88	10,818.82	52,408.41	52,277.65	52,391.46	52,271.10
Bayesian							
Iterations	40,000	40,000	20,000	20,000	40,000	40,000	20,000
Burn–in	15,000	15,000	5,000	5,000	15,000	15,000	5,000
Thinning	10	1	1	1	1	1	5

Model 1: Individual Level Risk Factors Only; Model 2: Nested Null Model (Unstructured Variance; Random Effects); Model 3: Nested Model with Individual Level Risk Factors (Unstructured Variance; Random Effects); Model 4: Nested Null Model (Structured Variance; CAR Model); Model 5: Nested Model with Individual Level Risk Factors (Structured Variance; CAR Model); Model 6: Nested Null Model (Unstructured and Structured Variance; Random Effects and CAR Models); Model 7: Nested Model with Individual Level Risk Factors (Unstructured and Structured Variance; Random Effects and CAR Models)

Table 4.4. Hazard ratios and 95% credible intervals for multivariate models that did not include prostate cancer tumor classification among South Carolina veteran population who were treated for prostate cancer at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria

Risk Factors	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Age at Diagnosis							
50 – 60	Reference		Reference		Reference		Reference
>60 – 70	1.66 (1.35 – 2.04)		1.65 (1.34 – 2.06)		1.35 (1.25 – 1.47)		1.35 (1.25 – 1.47)
>70	2.78 (1.77 – 2.94)		2.26 (1.74 – 2.93)		1.30 (1.16 – 1.47)		1.30 (1.16 – 1.46)
Gleason Score ≥7	1.55 (1.29 – 1.86)		1.55 (1.30 – 1.85)		1.18 (1.09 – 1.27)		1.18 (1.09 – 1.27)
Distant Stage	NA		NA		NA		NA
Radiation Therapy	0.60 (0.50 – 0.71)		0.60 (0.50 – 0.72)		0.84 (0.76 – 0.91)		0.84 (0.78 – 0.91)
Hormone Therapy	1.49 (1.23 – 1.81)		1.48 (1.22 – 1.79)		0.97 (0.88 – 1.06)		0.97 (0.87 – 1.06)
Marital Status							
Married	Reference		Reference		Reference		Reference
Previously Married	1.37 (1.15 – 1.64)		1.38 (1.15 – 1.66)		1.14 (1.05 – 1.23)		1.14 (1.06 – 1.24)
Never Married/Unknown	1.53 (1.12 – 2.04)		1.54 (1.14 – 2.09)		1.23 (1.08 – 1.40)		1.24 (1.08 – 1.40)
Travelled > 55 Miles	1.28 (1.07 – 1.53)		1.26 (1.07 – 1.48)		0.99 (0.90 – 1.09)		0.99 (0.89 – 1.09)
Modified Elixhauser Comorbidity Index	1.10 (1.06 – 1.13)		1.09 (1.06 – 1.14)		0.97 (0.96 – 0.99)		0.97 (0.96 – 0.99)
Median Odds Ratio (unstructured variance)	NA	1.08 (1.02 – 1.18)	1.07 (1.02 – 1.18)	NA	NA	1.02 (1.01 – 1.03)	1.00 (1.00 – 1.01)
Median Odds Ratio (structured variance)	NA	NA		1.03 (1.01 – 1.05)	1.02 (1.00 – 1.04)	1.03 (1.01 – 1.05)	1.01 (1.00 – 1.04)
Deviance	10,863.44	11,037.88	10,861.33	52,408.41	52,293.97	52,391.46	52,290.60
Bayesian							
Iterations	60,000	40,000	60,000	20,000	20,000	40,000	40,000
Burn–in	30,000	15,000	30,000	5,000	5,000	15,000	15,000
Thinning	10	1	5	1	5	1	1

Model 1: Individual Level Risk Factors Only; Model 2: Nested Null Model (Unstructured Variance; Random Effects); Model 3: Nested Model with Individual Level Risk Factors (Unstructured Variance; Random Effects); Model 4: Nested Null Model (Structured Variance; CAR Model); Model 5: Nested Model with Individual Level Risk Factors (Structured Variance; CAR Model); Model 6: Nested Null Model (Unstructured and Structured Variance; Random Effects and CAR Models); Model 7: Nested Model with Individual Level Risk Factors (Unstructured and Structured Variance; Random Effects and CAR Models)

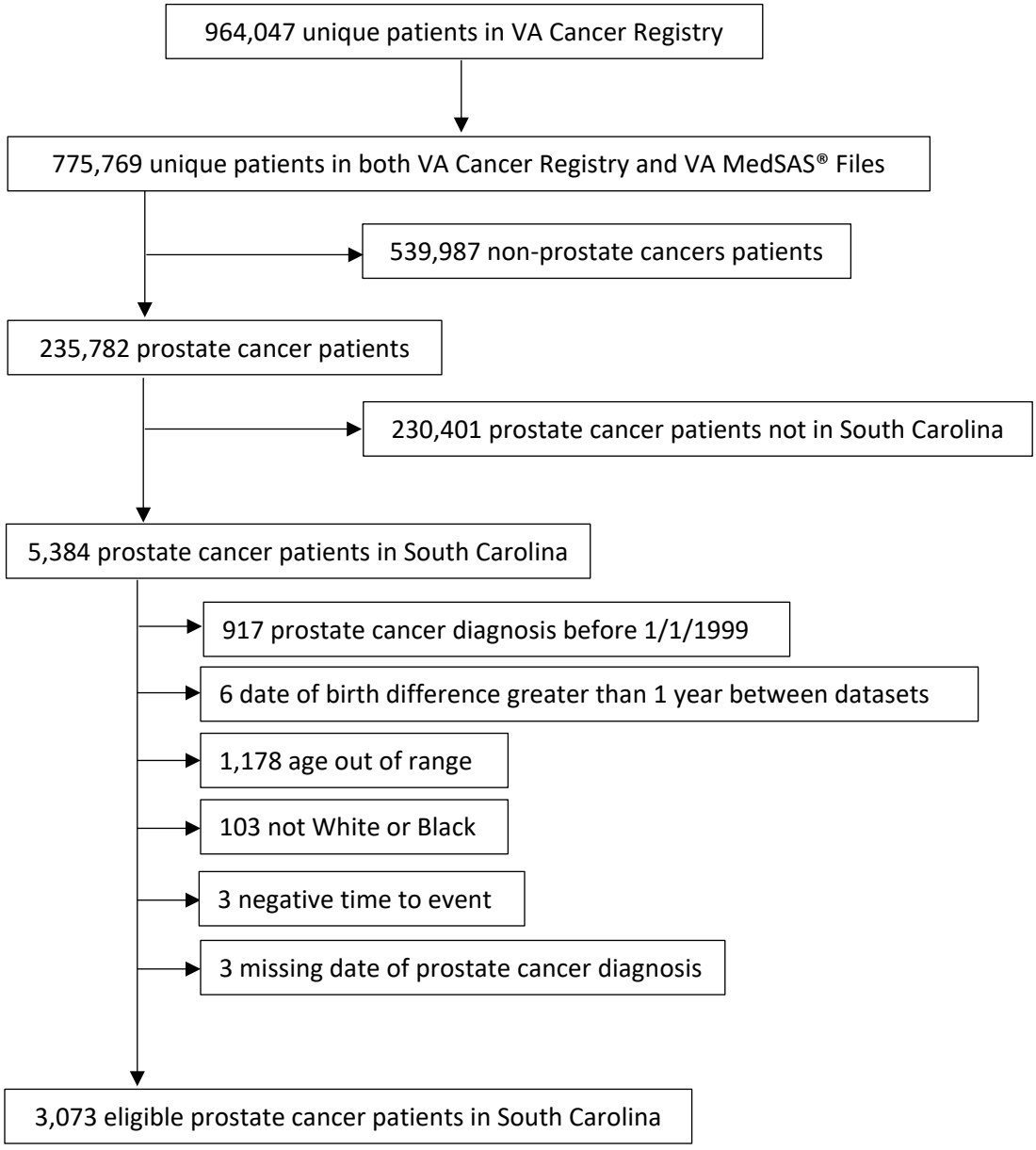

Figure 4.1. STROBE Diagram for the generation of the analytical cohort for survival modeling among the South Carolina veteran population with prostate cancer who seek healthcare at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria

CHAPTER 5

Prostate Cancer Mortality-to-Incidence Ratios Among South Carolina Veterans: 1999 – 2015

5.1 Abstract

Background: The mortality-to-incidence ratio (MIR) is an established epidemiological measure that approximates one minus survival and therefore is a valid approximation of frailty. Use of MIRs in recent years has helped to describe virulence of cancer in specific populations. A key feature of the MIR is its ability to be used to compare cancers across different populations. This study describes prostate cancer MIRs in the South Carolina (SC) veteran population from 1999 to 2015.

Methods: US Veterans Health Administration electronic medical records from January 1999 to December 2015, were used to identify 3,073 PrCA patients residing in 346 ZIP codes within SC. An overall MIR was calculated along with MIRs stratified by key prostate cancer risk factors and neighborhood risk factors. PrCA MIRs were also computed for ZIP code tabulation areas (ZCTA) in which veterans resided. Global and local Moran I statistics were computed to examine spatial clustering of MIRs across SC ZCTAs.

Results: The overall PrCA MIR in this population was 0.17. The MIR was 0.15 among Blacks, 0.20 among Whites, 0.16 among men who were married at PrCA diagnosis, 0.18 in men residing in a metropolitan ZCTA, and 0.16 in men residing in a

non-metropolitan ZCTA. Collectively, MIRs did not vary spatially by ZCTAs. Two ZCTA clusters of higher than expected MIRs were found in the upstate region.

Conclusions: The overall MIR was lower than for the population as a whole. This is the first known report showing that the MIR was lower among Blacks than Whites (25% lower as opposed to 59% higher in the US as a whole). The PrCA MIR among veterans during the study timeframe varied across key risk factor strata for PrCA. Also, geographically, statistically significant clusters MIRs were found. Application of the MIR in this context revealed that veterans, on average, have lower mortality given incidence than in the population as a whole and that Blacks have a MIR which is much lower (i.e., by 38%) than for the country as a whole. Additional work is needed to understand the reason for these differences.

5.2 Introduction

The mortality-to-incidence ratio (MIR) for cancers has been established to be a valid approximation of the 5-year frailty measure, which is the equivalent of 1-survival estimate.[90,91] Use of the MIR to describe cancer virulence is becoming more popular among health scientists due to the relative ease of computing MIRs, as opposed to obtaining the requisite data to compute survival or frailty.[44,45,90-101] A key feature of the MIR is that cancer cases do not have to be followed individually. The only information required is the number of new cancer cases and the number of individuals with that same cancer who died in a specific timeframe. The cancer MIR is an important measure that can be used for both descriptive and inferential comparisons.[44,45,92-101] This utility of the MIR for epidemiological cancer research allows scientists to describe and

understand how a specific cancers affect the general population and how these factors may vary in key subgroups or segments of population.

The United States (US) Veterans Health Administration (VHA) has an electronic health records database that affords researchers a robust source of important clinical data for a unique segment of the population. The information in the VHA electronic medical records is standardized for the entire US making it the largest such repository in the US. Another important and unique feature of the VHA is that it is not a fee-for-service healthcare enterprise. Therefore, financial considerations associated with healthcare utilization are minimized for veterans who qualify to receive care at the VHA.

Prostate cancer (PrCA) is an important cancer to investigate for two reasons. The first is that PrCA is the most common epithelial cancer among men in the US, accounting for 19% of all new cancers in 2018 while only accounting for 9% of all cancer mortality during 2018.[2] This difference between PrCA incidence and mortality has led to the colloquial but accurate statement that an "individual with PrCA is more likely to die with it than due to it". The second reason is that large differences exist in both PrCA incidence and mortality rates between Whites and Blacks and across different age groups.[2] By better understanding the epidemiology of PrCA in various groups can begin to minimize if not eliminate these disparities. The study of PrCA among veterans becomes more important since as of 2015, nationally males comprise 86.89% of the veteran population and 89.98% in South Carolina (SC).[102]

This study was designed to describe and evaluate PrCA MIRs among SC veterans who receive care at VHA facilities. SC was chosen because of its small geographical size

relative to other states, high representation by Black Americans, and its stable veteran population. Overall PrCA MIRs for all veterans and by key subgroups was determined. Additionally, we sought to evaluate geospatial differences in veteran MIRs by ZIP code tabulation areas (ZCTAs).

5.3 Methods

5.3.1 Study Design

This study employed a retrospective cohort design. The timeframe was from 01 January 1999 to 31 December 2015. The geographical unit of investigation was the ZCTA, which the US Census Bureau develops every ten years by aggregating Census blocks whose addresses have a common Zone Improvement Plan (ZIP) code.[64]

5.3.2 Data Sources

Data were obtained from the US Department of Veteran Affairs and the U.S Census Bureau. From the US Department of Veteran Affairs the following datasets were used: all MedSAS® datasets, Master Vital Status dataset, Mini Vital Status dataset, and Primary Oncology dataset from the VA Central Cancer Registry (VACCR).[52,53] The VA Master Vital Status and VA Mini Vital Status datasets provided patient-level information (i.e., each unique VA patient is listed only once in these files) while the VA MedSAS® datasets provided visit-level information (i.e., with ≥1 record/subject).[52,53] From the US Census Bureau the 2015 ZCTA shapefile for the US was obtained.[64]

5.3.3 South Carolina Geographical Area

Eligible South Carolina veterans with PrCA resided in 346 of South Carolina's 424 ZIP codes (81.6%). The 2015 ZCTA shapefile from the US Census Bureau was used. South

Carolina ZCTAs were in turn used to link the selected ZCTAs to patient's ZIP codes with identical designations in the VA MedSAS® Files.[52,64]

5.3.4 Exclusion Criteria

Subjects were excluded based on the following criteria (see Figure 4.1 for the STROBE Diagram establishing the analytic cohort): data in the Primary Oncology Dataset from the VA Cancer Registry could not be linked to MedSAS® datasets[52], a non-PrCA diagnosis in the Primary Oncology Dataset from the VA Cancer Registry, PrCA diagnosis in the Primary Oncology Dataset from the VA Cancer Registry prior to January 1, 1999, date of birth differs by more than 365 days between VA MedSAS® datasets[52] and VA Vital Status datasets[53], age during the timeframe did not wholly or partially fall between 40 and 70 years old, missing ZIP code information, racial classification other than White or Black because those racial classifications accounted for 2.53% of the PrCA patients, not residing in a SC ZIP code, missing Gleason Score listed in the VACCR, missing PrCA diagnosis date, and negative time to event due to data entry error (e.g., diagnosis date occurred after date of death).

Overall, 964,047 unique subjects from a total of 1,159,188 (83.17%) unique cancer diagnoses were included in the Primary Oncology Dataset. Of these subjects, 80.47% could be linked to unique patients in the MedSAS® datasets.[52]

There were 235,782 unique PrCA cases in the Primary Oncology Dataset that could be linked to the VA MedSAS® Files. There were 230,401 (97.72%) PrCA cases excluded due to not residing in a SC ZIP code or have a missing ZIP code. Of the 5,384 PrCA cases residing in a SC ZIP code; 1,648 (30.61%) were ineligible due to diagnosis

prior to 01/01/1999, date of birth discrepancy, age at first VA visit not satisfied, and/or not being White or Black, 659 (12.2%) did not have a Gleason Score, three (0.06%) did not have a diagnosis date, and three (0.06%) had a date of death prior to their date of PrCA diagnosis. The final analytical cohort consisted of 3,073 eligible SC veterans with PrCA (Figure 5.1).

5.3.5 Mortality-to-Incidence Ratio Calculations

The overall MIR for PrCA in SC among Black and White veterans who sought healthcare at VHA facilities was calculated as the number of those cases who died due to PrCA according to the VACCR divided by the number of PrCA cases in the VACCR.[45] MIRs have a minimum value of 0 (i.e. no death for the disease of interest among everyone with the disease during the timeframe of interest) to 1 (i.e. everyone died for the disease of interest among everyone with the disease during the timeframe of interest). Stratified PrCA MIRs (e.g., by socio-economic status) for this cohort were calculated using the same calculation.

5.3.6 Stratified Mortality-to-Incidence Ratios for Individual-Level Variables

Age at First VA Visit: The date of the first VA visit for each subject was obtained from the MedSAS® datasets and the date of birth (DOB) was obtained from the VA Mini Vital Status dataset.[52] Four age strata variables were created: age <40 years, age from 40 – 50 years, age from >50 – 60 years, age from >60 – 70 years. The number of PrCA cases and the number of deaths among these PrCA cases were aggregated by each stratum to calculate the MIR for each age at first VA visit strata.

Race-Related Variable: The information located within the VA Master Vital Status file for race and ethnicity were classified into White, Black, Hispanic, Asian, Hawaiian/Pacific Islander, and Native/Alaskan America, Other, and Unknown.[53] The numbers of PrCA cases and deaths among these PrCA cases were aggregated by each stratum to calculate the MIR for each racial strata.

Marital Status-Related Variable: The marital status at the time of diagnosis (obtained from the VACCR) was categorized as Married, Previously Married (Divorced, Widowed, Separated), and Never Married/Unknown. The number of PrCA cases and the number of deaths among these PrCA cases were aggregated to calculate the MIR for the marital status strata.

5.3.7 Stratified Mortality-to-Incidence Ratios for Neighborhood -Level Variables

The Social Vulnerability Index (SoVI®): The SoVI®, initially developed by the Hazards and Vulnerability Research Institute (HVRI) at the University of South Carolina, is a composite measure of neighborhood-level factors obtained from publicly available population based datasets of factors associated with the health of individuals within those neighborhoods.[58] The SoVI® uses ZCTA measurements obtained from the 2010 US Census and the 2007 – 2011 5-Year American Community Survey.[4,5] The SoVI® is a relative measure represented on a continuous scale from negative infinity to positive infinity in which the greater a geographical unit's SoVI® score is the less prepared for, able to respond to, and able to recover from a disaster in that area compared to a geographical unit with a lower SoVI® score.[66] A SoVI® score was assigned for each patient's ZIP code listed at the time of PrCA diagnosis. The SoVI® measurements were

stratified into 3 categories using cutoffs of 1 standard deviations: Low (-∞, -2.04], Medium (-2.04, 2.79], and High (2.79, ∞). The number of PrCA cases and the number of deaths among these PrCA cases were aggregated to calculate the MIR for the SoVI® strata.

Rural-Urban Community Area (RUCA) classification: The RUCA is a measure of how rural an area is. Originally developed by the Economic Research Service, which is part of the US Department of Agriculture.[103] The RUCA was initially determined for states, counties, and census tracts.[103] The WWAMI Rural Health Research Center at the University of Washington extended the RUCA to include ZIP codes.[104] The SC ZIP code RUCA from the University of Washington was merged with ZCTA measurements obtained from the US Census Bureau.[64,104] Every ZCTA-linked ZIP codes with PrCA cases were able to link to the corresponding RUCA ZIP code. The ZIP code-level RUCA was stratified into metropolitan and non-metropolitan classifications for each ZCTA. The number of PrCA cases and the number of deaths among these PrCA cases were aggregated by each RUCA stratum to calculate the MIR for each RUCA strata.

5.3.8 Descriptive Spatial Analyses

MIRs for each ZCTA were plotted on the SC map obtained from the US Census Bureau using R to create a choropleth map of SC PrCA MIRs by ZCTA among SC veterans (Figure 5.2).[61,64] A global Moran's I was calculated to determine whether there is overall spatial randomness among MIRs across SC ZCTAs.[37] A local Moran's I was calculated and mapped to determine if specific ZCTA regions in SC display spatial clustering.[37]

5.4 Results

There were 3,073 PrCA cases in the VACCR in SC from January 1, 1999 to December 31, 2015. Of those, 534 died during that time where the cause of death was directly attributed to PrCA. This resulted in an overall MIR of 0.17. There were 1,337 Whites with PrCA and 267 deaths among White SC veterans due to PrCA, for a MIR of 0.20 (Figure 5.2). There were 1,736 Blacks with PrCA were 267 deaths among Black SC veterans due to PrCA for an MIR of 0.15. Among veterans who had their first VA visit prior to 40 years of age, there were 28 PrCA cases with 1 death due to PrCA for a MIR of 0.04. Among veterans who had their first VA visit between 40 and 50 years of age, there were 448 PrCA cases with 38 deaths due to PrCA for a MIR of 0.08. Among veterans who had their first VA visit between after 50 and 60 years of age, there were 1,451 PrCA cases with 210 deaths due to PrCA for a MIR of 0.14. Among veterans who had their first VA visit between after 60 and 70 years of age, there were 1,146 PrCA cases with 285 deaths due to PrCA for a MIR of 0.25. There were 1,767 PrCA cases who were married at the time of diagnosis of which 277 died due to PrCA for a MIR among married SC veterans of 0.16. Among ZCTAs whose had the lowest SoVI® values, there were 168 PrCA cases with 22 deaths due to PrCA for a MIR among the ZCTAs with the lowest SoVI® values of 0.13. Among ZCTAs whose had the medium SoVI® values, there were 2,342 PrCA cases with 426 deaths due to PrCA for a MIR among the ZCTAs with medium SoVI® values of 0.18. Among ZCTAs whose had the highest SoVI® values, there were 563 PrCA cases with 86 deaths due to PrCA for a MIR among the ZCTAs with the lowest

SoVI® values of 0.15. There were 1,974 PrCA cases among veterans who lived in metropolitan ZCTA based of

RUCA values with 362 deaths due to PrCA, for a MIR of 0.18. There were 1,098 PrCA cases among veterans who lived in non-metropolitan ZCTA based of RUCA values with 172 deaths due to PrCA for a MIR of 0.16 (Figure 5.3).

The global Moran's I for PrCA MIRs by ZCTA was statistically non-significant (p-value = 0.08). From this we can conclude that PrCA MIRs among SC veterans during the study timeframe among all SC ZCTAs did not exhibit spatial clustering. However, based on the local Moran's I test, there are two statistically significant clusters of ZCTA (Figure 5.4).

5.5 Discussion

The overall MIR for PrCA among veterans receiving care at VHA facilities from January 1, 1999 to December 31, 2015 of 0.17 is higher than the 2011 – 2015 national PrCA MIR of 0.15.[38]

A distinct steep linear trend is observed in the MIR strata for the categories of age at first VA visit. Possible explanations for this trend include that veterans who start their VHA healthcare at younger ages are more likely to have preventive care visits. Another possibility is that veterans who wait to start their VHA care are doing so because of pre-existing conditions and are starting to go to VHA facilities to reduce the cost of healthcare from non-VHA sources. It needs to be noted that age of PrCA diagnosis was not used because there were several cases (not a majority) who did not have that information recorded in the VACCR.

As expected, those veterans who live in the ZCTAs deemed least vulnerable to the impact of natural disasters recovery efforts have lower MIRs than those who live in more vulnerable ZCTAs as well as the overall MIR. Interestingly, veterans living in ZCTAs that are deemed being at medium level of social vulnerbility have the highest MIRs among the three SoVI® strata. One possible explanation for this occurrence is that veterans who seek care at VHA facilities are less likely to reside in either the most or least affluent ZCTAs in SC; therefore, they are concentrated in ZCTAs that look more like the mean/median with respect to affluence.

Veterans who are married also have a slightly lower PrCA MIR. It should be noted that married veterans comprise slightly less than 58% of the entire cohort. This confirms the established protective impact of marriage in PrCA survival.[88]

There also was a rural/urban disparity observed in PrCA MIRs in this cohort. Veterans residing in metropolitan ZCTAs had a lower MIR than those observed for the overall cohort while veterans residing in non- metropolitan ZCTAs had higher a MIR than those observed for the overall cohort. This confirms that veterans in rural areas (like the general population) are at increased risk of not undergoing routine clinical care and therefore increased mortality for PrCA.[8]

Interesting and counter to expectations, Blacks had a lower MIR than whites in this veteran population. Hebert et al. found that from 2001 to 2005 the statewide PrCA MIR for Blacks was 0.26 while for Whites it was 0.16.[45] Comparing these MIR values raises many questions. Black veterans in SC comprise the majority of VHA PrCA patients in SC, which is not the situation among PrCA cases in the entire state of SC (where it is

closer to 40%) .[39] One explanation for this observation is that whatever socio-economic latent factors influence the disparity in either the mortality, incidence, or both among Blacks in SC as compared to Whites are minimized by the care received at SC VHA facilities as compared to the rest of SC. Another, unique observation is that Whites have a higher PrCA MIR as compared to the general population in the state. Is the care provided for PrCA at SC VHA facilities better for Blacks but worse for Whites? This observation needs to be investigated further by determining MIRs for other cancers in veteran populations in SC and nationwide.

It also was determined that clusters of higher than expected MIRs in ZCTAs exist in two portions of the Upstate Region of SC. What makes these areas at increased risk for high PrCA MIRs in the veteran population is unknown. However, the utility of applying a standard geospatial statistic to spatial MIR distributions resulted in this observation. This descriptive approach is the first step in determining the reasons for those clusters, and to test if those clusters persist in other cancers among veterans.

In conclusion, the use of MIRs to describe PrCA epidemiology in South Carolina among veterans has served as an instrumental first step in understanding the epidemiology of PrCA in this population. This in turn has facilitated the development of several potential research hypotheses that will hopefully be evaluated. The first is to investigate why the PrCA MIR among SC Black veterans is lower than the statewide PrCA MIR for Blacks and/or conversely, why the PrCA MIR among SC White veterans is higher that the statewide PrCA MIR among SC Whites. The second is to evaluate in more detail the two statistically significant clusters isolated in one region of the state. How are the

MIR values with the clusters similar to each other and different from the surrounding ZCTAs that are not part of those clusters.

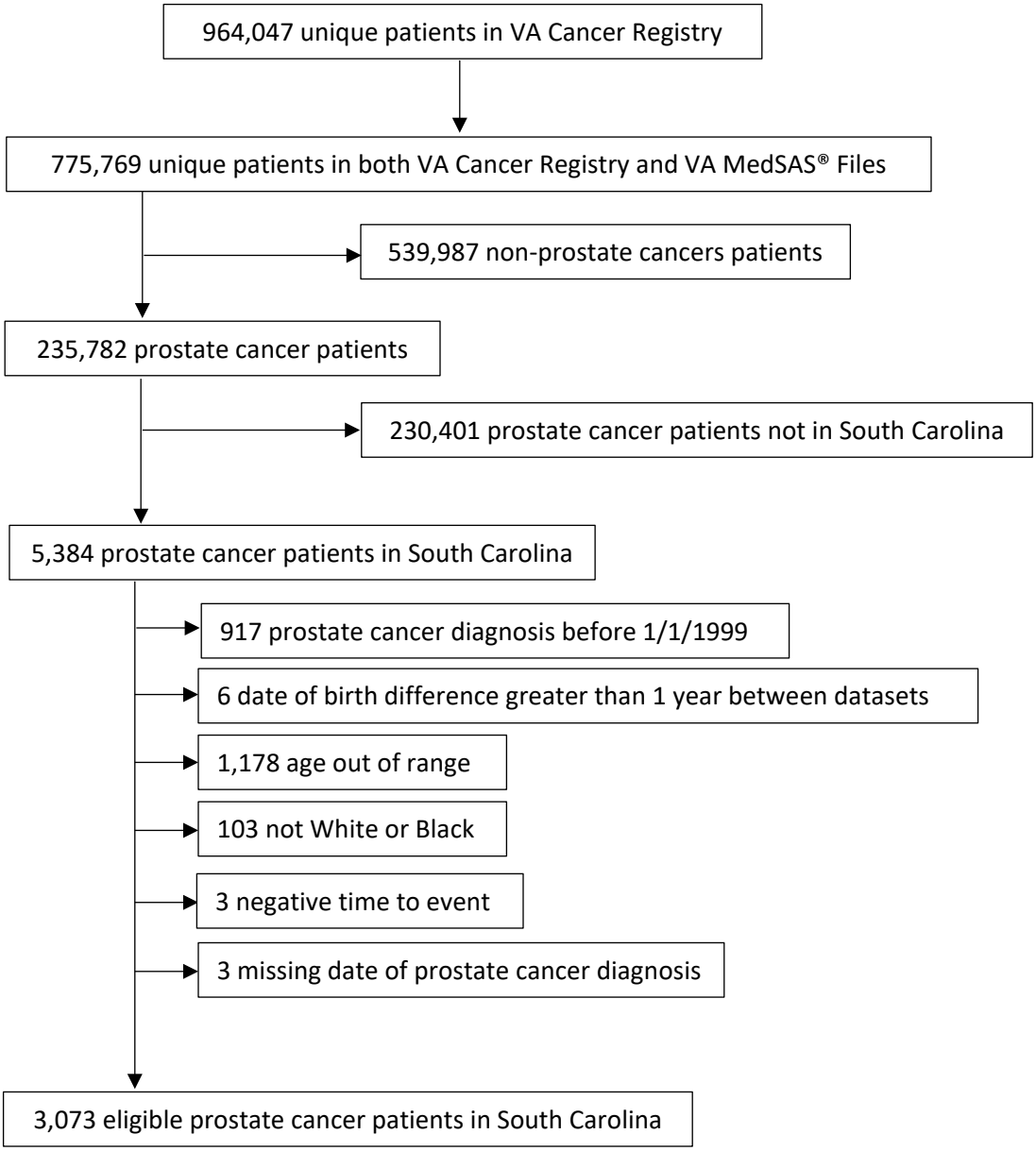

Figure 5.1. STROBE Diagram for the generation of the analytical cohort for survival modeling among the South Carolina veteran population with prostate cancer who seek healthcare at Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria

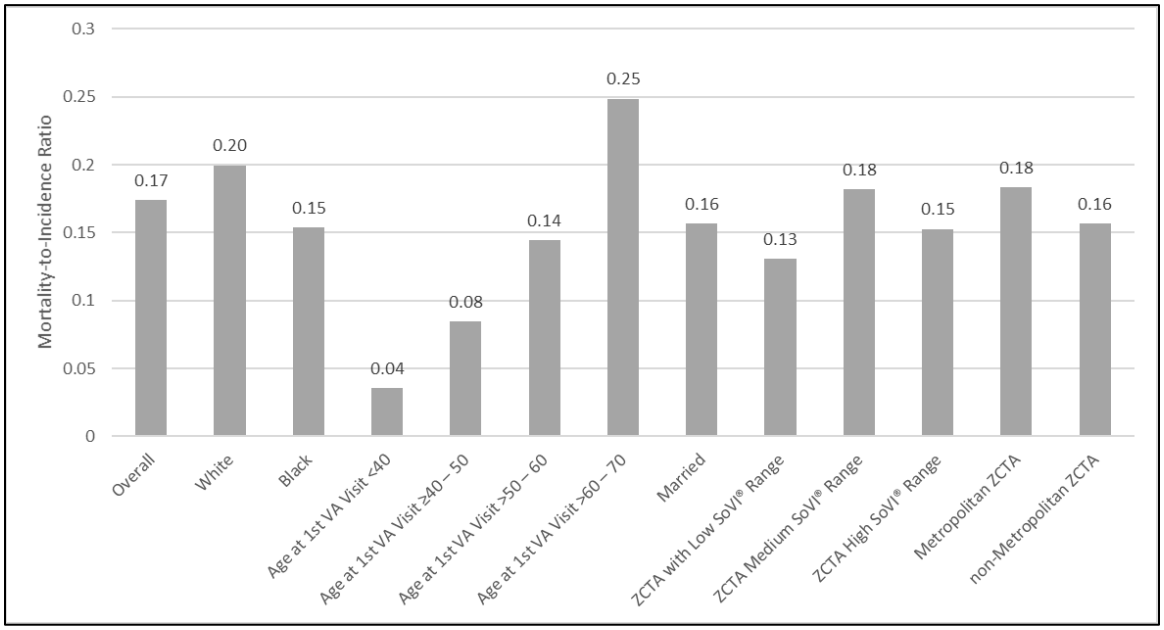

Figure 5.2. Overall and categorical mortality-to-incidence ratios among South Carolina veterans diagnosed with prostate cancer who receive care at US Veteran Health Administration facilities from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria

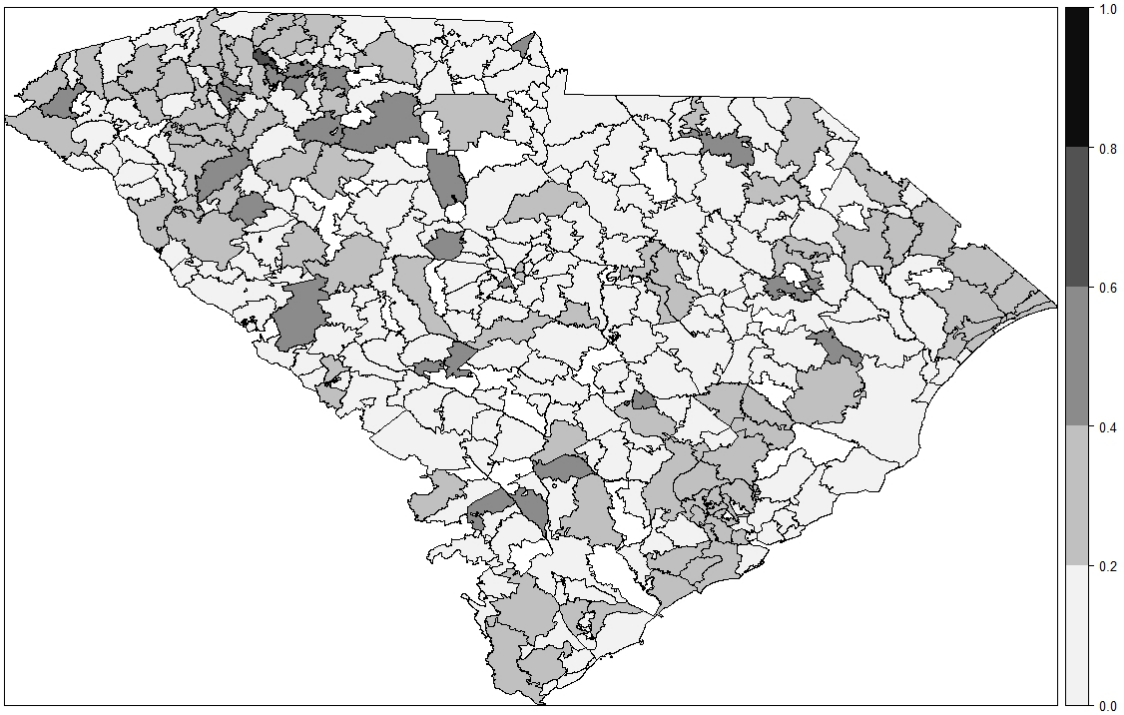

Figure 5.3. Prostate cancer mortality-to-incidence ratios among South Carolina veterans who receive care at US Veteran Health Administration facilities by ZIP code tabulation areas from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria

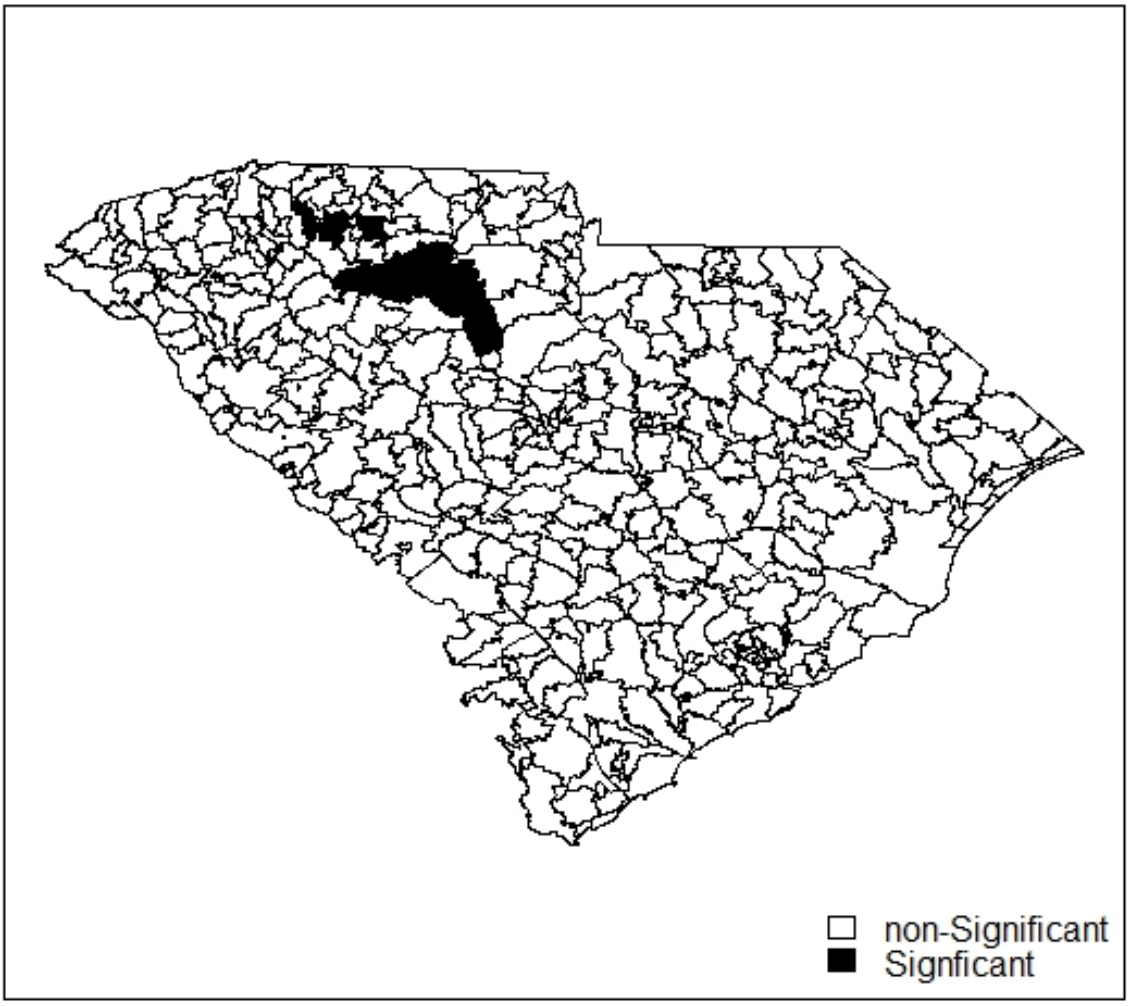

Figure 5.4. Prostate cancer mortality-to-incidence ratios results of local Moran's I among South Carolina veterans who receive care at US Veteran Health Administration facilities by ZIP code tabulation areas to identify MIR ZIP code tabulation area clusters from January 1, 1999 to December 31, 2015 and meet the study inclusion criteria

CHAPTER 6

CONCLUSIONS

6.1 Research Accomplishments

Are there spatial variations that influence prostate cancer (PrCA) epidemiology among South Carolina (SC) veterans who receive their healthcare from the Veteran Health Administration (VHA)? This was the overarching hypothesis of this research project. Each aim of this research sought to evaluate the possibility of spatial variation in PrCA epidemiology in this cohort.

Aim 1 (i.e. Chapter 3 of this dissertation) evaluated the influence of geospatial factors on the incidence of PrCA among SC veterans. The research undertaken for Aim 1 determined that if a veteran resides 25 to 55 miles from his most frequently visited VA facility, he is approximately 20% more likely to develop PrCA when accounting for only individual-level factors and individual-level factors nested within ZIP code residence were the ZIP code is evaluated as random effect (Models 1 and 3 Table 3.3). However, distance traveled becomes a statistically non-significant predictor once ZIP code characteristics were evaluated; for these analyses, increasing social vulnerability using the Social Vulnerability Index (SoVI®) in the ZIP code increases the likelihood of developing PrCA by 27% and residing in a ZIP code with at least 15.86% percent of the population in poverty increases the risk of PrCA by 24% (Model 5 Table 3.3). Interestingly, the most appropriate model in predicting PrCA incidence within the

veteran population in South Carolina (SC) is the most complex model in which individual- and neighborhood -level factors, and the random effects and spatial autocorrelation of ZIP codes are accounted for (Model 7 Table 3.3). In this model, while distance traveled and percent of poverty in the ZIP code are statistically non-significant, increasing SoVI® in the ZIP remains statistically significant and increases the risk of PrCA by at least 35% (Model 7 Table 3.3).

Based on these results we can conclude that no matter how complex the epidemiological model used for predicting PrCA, at least one geospatial factor predicts PrCA in veterans in SC, even after accounting for known individual-level factors. Furthermore, given how these factors vary between statistical significance and statistical non-significance, they may be accounting for a latent characteristic within the ZIP code that can be accounted for by several of these geospatial factors. These, in turn, may result in one geospatial factor remaining statistically significant while another geospatial factor becomes statistically insignificant. This phenomenon indicates possible research potential into expanding how to assess geospatial characteristics of ZIP codes to include environmental factors as well as evaluate if specific latent factors that comprise the SoVI® can serve as better predictors of PrCA incidence than the over SoVI®. It also highlights the fact that when we are able to move to larger geographical units, such as regions within the US or even the entire country, it is conceivable and indeed likely, that results that were statistically non-significant at a state level will emerge as significant in larger geographical units.

Aim 2 (i.e. Chapter 4 of this dissertation) evaluated the influence of geospatial factors on the mortality of PrCA among SC veterans. The research that was conducted for Aim 2 resulted in determining three things related to geospatial factors and PrCA mortality (Model 3 Table 4.3). The first is that a veteran traveling more than 55 miles was 26% more likely to die. The second is that the best-fitted model accounted for ZIP codes as random effects, which indicate the unique nature of ZIP code residence to this population regarding PrCA mortality. The third is that while it is best to account for the random effect of ZIP codes in this population; the ZIP code factors evaluated, however, were not appropriate in predicting PrCA mortality. Therefore, further research needs to be conducted to determine other geospatial factors that influence PrCA mortality in this population.

Aim 3 (i.e. Chapter 5 of this dissertation) evaluated whether the mortality-to-incidence ratio (MIR) of PrCA among SC veterans varied spatially across ZIP codes. While the overall distribution of PrCA MIRs across ZIP codes was found not to vary spatially; there were two distinct clusters of ZIP codes in the Upstate region of SC that had higher than expected PrCA MIRs as compared to their neighboring ZIP codes.

Additionally, from a modeling perspective, this research was able to incorporate and evaluated simultaneously three distinct levels of spatial factors. This accomplishment is of central importance to Aims 1 and 2 (Chapters 4 and 5 of this dissertation). The first geospatial factor considered was the distance a SC veteran traveled from the ZIP code listed most often on in his VHA electronic medical records to the VHA facility that SC veteran traveled to most often. This geospatial factor was used

as a potential predictor at the individual-level unit of analyses. Second, we were able to: 1) nest the veterans within each veteran's ZIP code of residence and analyze that ZIP code as a random effect; and 2) evaluate ZIP code-level characteristics as possible predictors to consider at the second-level hierarchical unit of analyses. The third geospatial factor that was considered was the spatial autocorrelation between ZIP codes. The spatial autocorrelation was evaluated for the individual-level only predictors for PrCA and for both the individual- and nested-level factors for PrCA concurrently. Evaluation of the ZIP code spatial autocorrelation is important in 1) identifying ZIP codes with lower- and higher than expected occurrences of the outcome of interest (i.e. PrCA incidence in aim 1 and PrCA mortality in aim 2), and 2) it determines a risk estimates and corresponding 95% credible intervals for the geographical units of measurement (i.e. ZIP codes for this research).

6.2 Research Strengths

There are three core strengths of this dissertation research. The first is the dissemination of a novel analysis with value to epidemiologists, clinicians, and policy makers. Epidemiologists can use this approach to more thoroughly evaluate diseases. Clinicians can assess the work of epidemiologists to assess how much value to place on where their patients reside for specific disease. Policy makers can use this approach to assist in determining how best to allocate resources based on the risk for diseases on where individuals reside. As modelling approaches continue to evolve so, too, must the approach that epidemiologists use to identify and investigative health issues. This will facilitate our ability to evaluate the causes of disease and point to methods for

intervening to prevent and control PrCA. Understanding how to use and interpret the results of more advanced modeling equations that can account for nested and spatial variations will benefit the health delivery and scientific communities.

Targeted research in specific, higher-than-normal-risk communities will allow researchers to investigate specific risk factors that are driving increased disease risk, especially risk of aggressive disease. In addition, this information will allow more community-specific interventions and prevention programs to be developed and evaluated.

The second core strength is the use the use of a national cancer registry for an entire population; i.e., veterans who seek care in the VHA system. The VA Central Cancer Registry (VACCR) meets the highest cancer registry standards and is uniform in its data collection procedures for the entire US In addition, because this uses VHA data, conclusions from this dissertation may lead to further evaluations of the services received by veterans seeking care within the VHA system and/or improve the current services those veterans receive.

The third core strength of this dissertation is that by utilizing a single state to establish the research approach, this approach can easily be applied to evaluate the epidemiology of PrCA incidence and mortality among veterans in larger geographical units, including on a national level. By better understanding the individual-level risk factors, neighborhood level risk factors, and spatial considerations associated with the incidence and mortality of PrCA; this will provide a template to evaluate other cancers

within the VHA population as well as provide a resource in shaping VHA policy nationwide as it relates to PrCA.

CHAPTER 7

REFLECTIONS

My journey in obtaining my doctoral degree in epidemiology has been a long, interesting, challenging, and rewarding experience. Earning my doctorate is rewarding on my levels. It is obvious on an intellectual and academic level, demonstrating that one can undertake complex and thorough research, and understand the details of your field of study.

For me, earning my doctorate has a unique personal quality. As a first-generation American, my parents passed up their academic and intellectual goals for the opportunity to immigrate to the United States. They sacrificed to earn a life in the United States so that their children may not have to make such difficult decisions in their lifetimes. They instilled in me and my siblings the importance of knowledge and wisdom that comes from intellectual endeavors. They reminded us that power and money are fleeting things that can be taken from anyone by force or lost by one's follies. However, no one can take away one's knowledge.

My initial studies led to my first passion, biology. More specifically, the cell and how magnificently complex it is. I thought I would have a career as a bench scientist after I completed my undergraduate degree in biology. My first job after college was testing human samples for diseases using state-of-the-art techniques at Quest Diagnostics. I found the work rewarding. However, as I gained more experience and

learned more laboratory techniques; I found the work to border on mundane. There was no challenge in it for me.

Serendipitously, Quest provided seminars for employees to better understand all aspects of the diseases we worked on. It was during these seminars that I found out about “epidemiology”. That was the first time that I heard of such a field. I found it fascinating that such a field requires so much cross-disciplinary skills and knowledge. I began to learn more about epidemiology with each seminar and my own study of it. I knew that I would eventually return to school to pursue a graduate degree. After more than two years of working for Quest, I decided to return to school; however, instead of cellular biology, I decided to pursue a Master’s in Public Health (MPH) degree.

Of the school’s that accepted me, I chose Saint Louis University because it offered the degree with dual concentrations in epidemiology and biostatistics. I found the additional workload to meet the requirements for both concentrations fun. I enjoyed learning the core principals of epidemiology and the underlying statistics that enable epidemiologists to care out their work.

One of the great features of the MPH program at Saint Louis University is that is requires 360 hours to be completed as an internship. I was fortunate to find my internship at the Infectious Disease Division at Washington University in Saint Louis. That internship turned into a job for me. I began by first assisting with research studies and then after a little over of year, I became a research coordinator for research studies. One of the great things about that job was the variety. I was part of case-control, retrospective cohort, prospective cohort, and randomized control trials. Each study was

unique. I was never bored. However, I learned the lesson that the freedom to conduct such a wide variety of studies in an academic environment comes with a price. More precisely, grant money. When grant funding runs out, so does your time for that type of research.

Thankfully, my experience in epidemiological research led to me being hired by the St. Louis County Department of Health as an Epidemiologist. Specifically, I focused on bio-preparedness; bioterrorism, pandemics, unique outbreaks. Practically, while I needed to be well trained and knowledgeable in bio-preparedness, I was used as a general county epidemiology. Interestingly, the bio-preparedness that I was my primary focus was pandemic influenza. I was hired a year before the 2009 influenza season in which the major strain of the virus mutated more than was expected. Therefore that season's influenza vaccine was mostly ineffective at the start of the influenza season while updated vaccines were being rushed into production for distribution. The pandemic influenza plan for St. Louis County that I had recently overhauled with major revisions was implemented. After the influenza season was over and based in part to my involvement in the response; the Director of Communicable Diseases strongly encouraged me to pursue my doctorate in epidemiology. He told me that he felt that I had the skill sets to one day sit in his position. I found the prospective of that potential future for myself rewarding, and so I applied to doctorate programs.

Of the schools that accepted my application, I chose the University of South Carolina because of all the cross-disciplinary centers that the School of Public Health offered. Due to my experience as an epidemiologist, I was able to test out of all

introductory and mid-level epidemiology courses. This created a fortunate opportunity for me to befriend individuals who have been in the program for a year or two. Because of the resulting experiences, I was able to better navigate the course work requirements of the program and to learn about faculty whose classes I had not taken yet.

One of those students at that time was Gowtham Rao who was nearly finished with his dissertation research at the VA. Sharing many of the interests in research, outlooks in conducting research, and personal characteristics; we became friends. It was Gowtham who introduced me to Dr. Charles Bennett in the College of Pharmacy and to Dr. Sue Haddock at the WJB Dorn Veteran's Affairs (VA). It was Gowtham who showed me the research potential of using VA data, which he was using in his dissertation work. He was instrumental in opening up the VA to me to conduct research.

After completing my coursework and passing my progression exam, I needed to develop a dissertation research project. I knew that I wanted to use VA data. What I did not know was the topic of my dissertation. As with many things, inspiration comes from many places. The first place was when I took Dr. Susan Steck's Cancer Epidemiology course and my research project for the class was a literature review of prostate cancer temporal and spatial epidemiology. The second place was from concurrently taking my PhD Seminar course, which was being proctored by Dr. Swann Adams. She required the class to draft a mock NIH grant. I used the project for Dr. Steck's course to serve as the Background Section of the mock grant for Dr. Adams' course. Drawing inspiration from the literature search, I developed a potential study. I quickly realized that this mock grant can be used as a dissertation project. Next, I needed to find my committee.

Thankfully, I was not naive to the importance of choosing my dissertation committee and my dissertation research topic. This was in no small part to all my friends that I met when I started the program who were ahead of me in the dissertation process. I knew my committee could not and should not, should not and would not be a rubber stamp. I knew I needed to thoroughly assess my own characteristics to find individuals that would complement my strengths, improve my weaknesses, and provide examples to make me a better researcher, keep me focused, and critically assess my work.

I knew that Dr. James Hébert would be ideal as my committee chairman. However, he was and is an extremely busy man. Would he have the time to serve on another dissertation committee? Would he have the interest in my dissertation project? Would he have the interest in me conducting that dissertation project? As it turned out, the answer to all those questions was "yes".

To Dr. James Hébert,

Thank you for so much! It is because of you, more than most, that I am able to reach this pinnacle. You were the fire under my ass making sure I never rested. You were the beacon in the distance making sure I stayed focused. You shared my exasperation when the Cancer Registry was giving me the bureaucratic run around for almost a year. You shared my frustration when I spent almost a year trying to make aim 1 analysis work for larger cohorts and the nation. You shared my excitement with each new set of results. Your example epitomizes what it means to train the next generation of researchers so as to continuously carry

forward the need for better scientists to conduct better science. As wonderful as this experience has been for me, I'm looking just as forward to the next phase of my academic research career knowing that I have your guidance in helping me navigate what is unexplored territory for me.

The next step after Dr. Hébert agreed to serve as my committee chairman was to find the other members. His first suggestion was Dr. Jan Eberth, who was at the time a new faculty member here. Dr. Hébert informed me that her research interests and experience overlaps with my proposed dissertation research. After meeting with Dr. Eberth and discussion my proposed dissertation research, she agreed to serve on my committee.

To Dr. Jan Eberth,

Thank you for taking a chance on me! We had never met prior to me asking you to serve on my committee. You will never know how much our meetings meant to me. You were a rock during this process. Constantly reminding me either implicitly or explicitly that "I am an Epidemiologist", not a data manager. As someone who has a habit of getting mired in the details, having a constant such as yourself made it possible to know what specifically needed to get done and why. (the "why" sometimes took a little longer)

The committee member that I absolutely knew that I would like to serve on my committee before I even began developing my dissertation proposal was Dr. Bo Cai. Having taken some courses from him, I knew that I could work well Dr. Cai.

To Dr. Bo Cai,

Thank you for making sure I did not mess up my analyses. I know that I straddle the line between confidence and arrogance when it comes to my statistically ability. You made sure my dissertation research would not suffer from my arrogance. Without you, the results we found would have remained a mystery. You made sure you guided me to properly begin to learn and apply Bayesian statistics. You made sure the results were valid and that I understood those results. Because of you, we can replicate this analytical approach in future research projects.

With the department faculty in place on my committee, I proceeded to select who would serve as outside members. Thankfully, the first choice made complete sense. Dr. Charles Bennett at the University of South Carolina College of Pharmacy is an oncologist at the VA who only cares for PrCA patients. In addition, he has extensive research experience with students and fellows doing PrCA research. Finally, I served as one of his graduate assistants during my academic studies.

To Dr. Charles Bennett,

It has been an amazing experience working for you. I have learned so such working on your studies that has expanded my knowledge in conducting epidemiological research. I have never met a scientist with as much passion in his/her area of expertise as you. You make research fun. It is only now, towards the end of my formal studies that understand why research can by fun. For that, I'll forever be indebted to you.

I was must concerned about not being allowed to have my second outside committee member. Yet, it made logical sense to have him on my committee. Dr. Gowtham Rao, was a recent graduate with whom I had taken classes with. He was appointed as an adjunct faculty at the School of Public Health upon his graduation. His most important expertise as it related to my dissertation was that he was a beta-tester for analyzing VA electronic medical records in the VA's, at the time, new virtual computing environment. In other words, he was part of a vanguard of VA health scientists in the nation using VA electronic medical records using 21st century technology. Thankfully for my dissertation research, the School of Public Health agreed with my reasons.

To Dr. Gowtham Rao

I'd like to believe that meeting you on my first day at the University of South Carolina was fated. You introduced me to a treasure trove of data that would take multiple lifetimes to address all possible research questions. You did not "hold my hand" but rather pointed me in the correct way and warned me if I was going down the wrong way. Because of the that, there are few epidemiologists in the nation who have had as much experience analyzing VA data using the benefits of the VA's virtual environment. On behalf of my career, thank you for opening this world to me!

Undertaken a dissertation research is like the journeys in literature's greatest stories were the protagonist meets individuals that help him/her in his/her quest as he/she faces obstacles to overcome. Some are obvious; like family, friends, dissertation

committee members. Other's are unexpected welcome additions. I had three such individuals.

The first was Dr. Sue Haddock, who is the Associate Chief of Staff for Research at the WJB Dorn VA. Dr. Haddock has been instrumental in developing my career as a VA scientist while giving me latitude to finish my dissertation.

To Dr. Sue Haddock,

Thank you for making sure I didn't lose myself in the government bureaucracy that can be overwhelming for new scientists working in the government. You made sure I didn't burden myself with too many studies. You promoted my research and my expertise. Because of you, my research has begun to establish me as someone capable of conducting VA research. I am grateful to have met you; your genuine interest in both myself, my dissertation, and my other research has always been a welcoming encourage to keep progressing forward.

The second was Dr. Linda Hazlett, who is the Graduate Director for Epidemiology at the School of Public Health. Students could ask for no better faculty advocate than Dr. Hazlett.

To Dr. Linda Hazlett,

Thank you for your sincere interest in my progress. Whenever I completed a milestone, whether major or minor in my dissertation progress, I knew that telling you in person would make it more "real". Your positive encouragement was always appreciated and, at times, needed (even if it was unbeknownst to you). Even when you would tell me to "hurry it alone", you did it with positive

reinforcement. You also made sure I didn't forget some logistical issue that would prevent me from completing my dissertation. I don't know how many student's say this to you; "thank you".

The third person was by far the most unexpected, Dr. Christopher Emrich. Dr. Emrich was at the University of South Carolina College of Geography when I met him. I was interested in using the Social Vulnerability Index (SoVI®) that he oversees in my dissertation research. What resulted was a positive year of working and learning from him. Because of that experience, I was able to write a first draft grant proposal for a Public Health inspired SoVI® that I used as my comprehensive exam.

To Dr. Chris Emrich,

It may have been by happenstance that we met but working with you opened a new avenue for future research that I hope I get to do. I want to let you know that your assistance with my dissertation research improved how impactful the results are. My hope is that we both find the time to find the next project we can work on together instead of merely exchanging emails messages.

In closing, this chapter was written for three reasons. The first to look back personally and take stock of what this experience has meant to me. The second to express how grateful I am to everyone that I mentioned directly and indirectly in the chapter. The third is in the hope that perhaps some doctoral student who is beginning his/her dissertation reads this and gleans some small amount of wisdom in how and what he/she is about to undertake.

REFERENCES

1. Murphy SL, Xu J, Kochanek KD, Curtin SC, Arias E. Deaths: Final Data for 2015. *Natl Vital Stat Rep.* 2017;66(6):1-75.
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. *CA: a cancer journal for clinicians.* 2018;68(1):7-30.
3. United States Cancer Statistics: Data Visualizations. Centers for Disease Control and Prevention. https://gis.cdc.gov/Cancer/USCS/DataViz.html. Accessed: 7 Sept 2018.
4. United States Census Bureau. 2010 Census.U.S. Census Bureau. 2010. http://www2.census.gov/census_2010/ . Accessed: August 2014. In.
5. United States Census Bureau. 2011 Five Year American Community Survey.U.S. Census Bureau. 2011. http://www2.census.gov/acs2011_5yr/ . Accessed: August 2014. In.
6. Cheng I, Witte JS, McClure LA, et al. Socioeconomic status and prostate cancer incidence and mortality rates among the diverse population of California. *Cancer causes & control : CCC.* 2009;20(8):1431-1440.
7. Clegg LX, Reichman ME, Miller BA, et al. Impact of socioeconomic status on cancer incidence and stage at diagnosis: selected findings from the surveillance, epidemiology, and end results: National Longitudinal Mortality Study. *Cancer causes & control : CCC.* 2009;20(4):417-435.
8. Obertova Z, Brown C, Holmes M, Lawrenson R. Prostate cancer incidence and mortality in rural men--a systematic review of the literature. *Rural and remote health.* 2012;12(2):2039.
9. Steenland K, Rodriguez C, Mondul A, Calle EE, Thun M. Prostate cancer incidence and survival in relation to education (United States). *Cancer causes & control : CCC.* 2004;15(9):939-945.
10. Yin D, Morris C, Allen M, Cress R, Bates J, Liu L. Does socioeconomic disparity in cancer incidence vary across racial/ethnic groups? *Cancer causes & control : CCC.* 2010;21(10):1721-1730.
11. Major JM, Norman Oliver M, Doubeni CA, Hollenbeck AR, Graubard BI, Sinha R. Socioeconomic status, healthcare density, and risk of prostate cancer among African American and Caucasian men in a large prospective study. *Cancer causes & control : CCC.* 2012;23(7):1185-1191.
12. Bigler SA, Pound CR, Zhou X. A retrospective study on pathologic features and racial disparities in prostate cancer. *Prostate cancer.* 2011;2011:239460.
13. Prostate Gland. In. *Encyclopaedia Britannica*.
14. Adami H-O, Hunter DJ, Trichopoulos D. *Textbook of cancer epidemiology.* 2nd ed. Oxford ; New York: Oxford University Press; 2008.

15. Understanding Task Force Recommendations. In: U.S.D.H.H.S., ed. *Screening for Prostate Cancer*2012.
16. H. Ballentine Carter PCA, Michael J. Barry, Ruth Etzioni, Stephen J. Freedland, Kirsten Lynn Greene, Lars Holmberg, Philip Kantoff, Badrinath R. Konety, Mohammad Hassan Murad, David F. Penson, and Anthony L. Zietman. *Early Detection of Prostate Cancer: AUA Guideline.* American Urological Association;2013.
17. PSA. *Medline Plus* http://www.nlm.nih.gov/medlineplus/ency/article/003346.htm.
18. Draisma G, Etzioni R, Tsodikov A, et al. Lead time and overdiagnosis in prostate-specific antigen screening: importance of methods and context. *Journal of the National Cancer Institute.* 2009;101(6):374-383.
19. Wever EM, Draisma G, Heijnsdijk EA, et al. Prostate-specific antigen screening in the United States vs in the European Randomized Study of Screening for Prostate Cancer-Rotterdam. *Journal of the National Cancer Institute.* 2010;102(5):352-355.
20. Schroder FH, Hugosson J, Roobol MJ, et al. Prostate-cancer mortality at 11 years of follow-up. *The New England journal of medicine.* 2012;366(11):981-990.
21. Andriole GL, Crawford ED, Grubb RL, 3rd, et al. Prostate cancer screening in the randomized Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial: mortality results after 13 years of follow-up. *Journal of the National Cancer Institute.* 2012;104(2):125-132.
22. Odom BD, Mir MC, Hughes S, et al. Active surveillance for low-risk prostate cancer in African American men: a multi-institutional experience. *Urology.* 2014;83(2):364-368.
23. Risbridger GP, Shibata A, Ferguson KL, Stamey TA, McNeal JE, Peehl DM. Elevated expression of inhibin alpha in prostate cancer. *The Journal of urology.* 2004;171(1):192-196.
24. Rodrigues G, Lukka H, Warde P, et al. The prostate cancer risk stratification (ProCaRS) project: recursive partitioning risk stratification analysis. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology.* 2013;109(2):204-210.
25. Stamey TA, Caldwell MC, Fan Z, et al. Genetic profiling of Gleason grade 4/5 prostate cancer: which is the best prostatic control tissue? *The Journal of urology.* 2003;170(6 Pt 1):2263-2268.
26. Zhou X, Lawrence TJ, He Z, Pound CR, Mao J, Bigler SA. The expression level of lysophosphatidylcholine acyltransferase 1 (LPCAT1) correlates to the progression of prostate cancer. *Experimental and molecular pathology.* 2012;92(1):105-110.
27. *Prostate Cancer.* National Comprehensive Cancer Network;2013.
28. Noone A, Howlader N, Krapcho M, et al.: SEER Cancer Statistics Review, 1975-2015, National Cancer Institute. Bethesda, MD, https://seer.cancer.gov/csr/1975_2015/, based on November 2017 SEER data submission, posted to the SEER web site, April 2018.

29. Goovaerts P. Medical Geography: a Promising Field of Application for Geostatistics. *Mathematical geology.* 2009;41:243-264.
30. Goovaerts P. How do multiple testing correction and spatial autocorrelation affect areal boundary analysis? *Spatial and spatio-temporal epidemiology.* 2010;1(4):219-229.
31. Goovaerts P. Geostatistical analysis of health data with different levels of spatial aggregation. *Spatial and spatio-temporal epidemiology.* 2012;3(1):83-92.
32. Jacquez GM, Goovaerts P. The emerging role and benefits of boundary analysis in spatio-temporal epidemiology and public health. *Spatial and spatio-temporal epidemiology.* 2010;1(4):197-200.
33. Banerjee S, Carlin BP, Gelfand AE. *Hierarchical modeling and analysis for spatial data.* Boca Raton, Fla.: Chapman & Hall/CRC; 2004.
34. Bivand R, Pebesma EJ, Gómez-Rubio V. *Applied spatial data analysis with R.* New York: Springer; 2008.
35. Haining RP. *Spatial data analysis : theory and practice.* Cambridge, UK ; New York: Cambridge University Press; 2003.
36. Lawson A. *Statistical methods in spatial epidemiology.* 2nd ed. Chichester, England ; Hoboken, NJ: Wiley; 2006.
37. Waller LA, Gotway CA. *Applied spatial statistics for public health data.* Hoboken, N.J.: John Wiley & Sons; 2004.
38. Prostate Cancer in the United States in 2011 - 2015 by State. In. *Interactive Cancer Atlas*: CDC.
39. Prostate Cancer in the United States in 2011 - 2015 by Race by State. In. *Interactive Cancer Atlas*: CDC.
40. Goovaerts P, Xiao H. Geographical, temporal and racial disparities in late-stage prostate cancer incidence across Florida: a multiscale joinpoint regression analysis. *International journal of health geographics.* 2011;10:63.
41. Goovaerts P, Xiao H. The impact of place and time on the proportion of late-stage diagnosis: the case of prostate cancer in Florida, 1981-2007. *Spatial and spatio-temporal epidemiology.* 2012;3(3):243-253.
42. Xiao H, Gwede CK, Kiros G, Milla K. Analysis of prostate cancer incidence using geographic information system and multilevel modeling. *Journal of the National Medical Association.* 2007;99(3):218-225.
43. Wagner SE, Bauer SE, Bayakly AR, Vena JE. Prostate cancer incidence and tumor severity in Georgia: descriptive epidemiology, racial disparity, and geographic trends. *Cancer causes & control : CCC.* 2013;24(1):153-166.
44. Wagner SE, Hurley DM, Hebert JR, McNamara C, Bayakly AR, Vena JE. Cancer mortality-to-incidence ratios in Georgia: describing racial cancer disparities and potential geographic determinants. *Cancer.* 2012;118(16):4032-4045.
45. Hebert JR, Daguise VG, Hurley DM, et al. Mapping cancer mortality-to-incidence ratios to illustrate racial and sex disparities in a high-risk population. *Cancer.* 2009;115(11):2539-2552.

46. Zhou H, Lawson AB, Hebert JR, Slate EH, Hill EG. A Bayesian hierarchical modeling approach for studying the factors affecting the stage at diagnosis of prostate cancer. *Statistics in medicine.* 2008;27(9):1468-1489.
47. Zhou H, Lawson AB, Hebert JR, Slate EH, Hill EG. Joint spatial survival modeling for the age at diagnosis and the vital outcome of prostate cancer. *Statistics in medicine.* 2008;27(18):3612-3628.
48. Hsu CE, Mas FS, Miller JA, Nkhoma ET. A spatial-temporal approach to surveillance of prostate cancer disparities in population subgroups. *Journal of the National Medical Association.* 2007;99(1):72-80, 85-77.
49. Oliver MN, Smith E, Siadaty M, Hauck FR, Pickle LW. Spatial analysis of prostate cancer incidence and race in Virginia, 1990-1999. *American journal of preventive medicine.* 2006;30(2 Suppl):S67-76.
50. Klassen AC, Curriero FC, Hong JH, et al. The role of area-level influences on prostate cancer grade and stage at diagnosis. *Preventive medicine.* 2004;39(3):441-448.
51. DeChello LM, Gregorio DI, Samociuk H. Race-specific geography of prostate cancer incidence. *International journal of health geographics.* 2006;5:59.
52. Department of Veterans Affairs. VA Medical SAS® Files (121VA10P2). In. Austin, TX: VA Corporate Data Center Operations, Austin Information Technology Center.
53. Department of Veterans Affairs. VA Vital Status File (121VA10P2). In. Austin, TX: VA Corporate Data Center Operations, Austin Information Technology Center.
54. United States Census 2010. http://www.census.gov/2010census/.
55. American Community Survey. http://www.census.gov/acs/www/.
56. Social Vulnerability Index | Selected Applications of the Usage of SoVI®. *Hazards and Vulnerability Research Institute* http://webra.cas.sc.edu/hvri/products/SoVIapplications.aspx.
57. Social Vulnerability Index for the United States - 2006-10. *Hazards and Vulnerability Research Institute* http://webra.cas.sc.edu/hvri/products/sovi.aspx.
58. Cutter SL, Boruff BJ, Shirley WL. Social Vulnerability to Environmental Hazards. *Social Science Quarterly.* 2003;84(2).
59. *SAS 9.4. SAS Institute Inc., Cary, NC.* [computer program].
60. Lunn D. *The BUGS book : a practical introduction to Bayesian analysis.* Boca Raton, FL: CRC Press, Taylor & Francis Group; 2013.
61. *R: A language and environment for statistical computing.* [computer program]. Vienna, Austria: R Foundation for Statistical Computing; 2010.
62. Prostate Cancer in the United States in 2015 by State. In. *Interactive Cancer Atlas*: CDC.
63. ZIP Code™ Tabulation Areas (ZCTAs™). http://www.census.gov/geo/reference/zctas.html.
64. United States Census Bureau. TIGER/Line® Shapefiles. https://www.census.gov/cgi-bin/geo/shapefiles/index.php . Accessed: August 2014.

65. STROBE Statement. https://www.strobe-statement.org/index.php?id=strobe-home. Accessed March 2015.
66. Schmidtlein MC, Deutsch RC, Piegorsch WW, Cutter SL. A sensitivity analysis of the social vulnerability index. *Risk Anal.* 2008;28(4):1099-1114.
67. Kleinbaum DG, Kupper LL, Nizam A, Rosenberg ES. *Applied regression analysis and other multivariable methods.* Fifth edition. ed. Boston, MA: Cengage Learning; 2013.
68. Merlo J, Chaix B, Ohlsson H, et al. A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *J Epidemiol Community Health.* 2006;60(4):290-297.
69. Geweke J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Bayesian Statistics 4; 1992.
70. Gilligan T. Social disparities and prostate cancer: mapping the gaps in our knowledge. *Cancer causes & control : CCC.* 2005;16(1):45-53.
71. Coker AL, Sanderson M, Ellison GL, Fadden MK. Stress, coping, social support, and prostate cancer risk among older African American and Caucasian men. *Ethn Dis.* 2006;16(4):978-987.
72. Newell GR, Pollack ES, Spitz MR, Sider JG, Fueger JJ. Incidence of prostate cancer and marital status. *Journal of the National Cancer Institute.* 1987;79(2):259-262.
73. Tyson MD, Andrews PE, Etzioni DA, et al. Marital status and prostate cancer outcomes. *Can J Urol.* 2013;20(2):6702-6706.
74. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care.* 2013;51(8 Suppl 3):S30-37.
75. Rao GA, Mann JR, Bottai M, et al. Angiotensin receptor blockers and risk of prostate cancer among United States veterans. *J Clin Pharmacol.* 2013;53(7):773-778.
76. Rao GA, Mann JR, Shoaibi A, et al. Azithromycin and levofloxacin use and increased risk of cardiac arrhythmia and death. *Ann Fam Med.* 2014;12(2):121-127.
77. Rao GA, Mann JR, Shoaibi A, et al. Angiotensin receptor blockers: are they related to lung cancer? *J Hypertens.* 2013;31(8):1669-1675.
78. Shoaibi A, Rao GA, Cai B, Rawl J, Haddock KS, Hebert JR. Prostate Specific Antigen-Growth Curve Model to Predict High-Risk Prostate Cancer. *Prostate.* 2017;77(2):173-184.
79. Grubesic TH, Matisziw TC. On the use of ZIP codes and ZIP code tabulation areas (ZCTAs) for the spatial analysis of epidemiological data. *International journal of health geographics.* 2006;5:58.
80. United States. Congress (104th 2nd session : 1996). *Health Insurance Portability and Accountability Act of 1996 : conference report (to accompany H.R. 3103).* Washington, D.C.: U.S. G.P.O.; 1996.
81. Rapiti E, Fioretta G, Schaffar R, et al. Impact of socioeconomic status on prostate cancer diagnosis, treatment, and prognosis. *Cancer.* 2009;115(23):5556-5565.

82. Prostate Cancer in the United States in 2014 by State. In. *Interactive Cancer Atlas*: CDC.
83. VHA Eligibility Criteria. https://www.va.gov/healthbenefits/resources/publications/hbco/hbco_basic_eligibility.asp. Accessed 10/8/2018.
84. Sohn MW, Zhang H, Arnold N, et al. Transition to the new race/ethnicity data collection standards in the Department of Veterans Affairs. *Popul Health Metr.* 2006;4:7.
85. South Carolina Statistical Abstract 2011: Health and Demographics. South Carolina Budget and Control Board, Office of Research and Statistics, 2011. (Accessed 26 May 2011, 2011, at http://ors.sc.gov/hd/default.php.
86. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care.* 2005;43(11):1130-1139.
87. Druskin SC, Mamawala M, Tosoian JJ, et al. Older Age Predicts Biopsy and Radical Prostatectomy Grade Reclassification to Aggressive Prostate Cancer in Men on Active Surveillance. *The Journal of urology.* 2018.
88. Xiao H, Tan F, Goovaerts P, et al. Impact of Comorbidities on Prostate Cancer Stage at Diagnosis in Florida. *Am J Mens Health.* 2016;10(4):285-295.
89. Fotheringham AS, Brunsdon C, Charlton M. *Geographically weighted regression : the analysis of spatially varying relationships.* Chichester, England ; Hoboken, NJ, USA: Wiley; 2002.
90. Asadzadeh Vostakolaei F, Karim-Kos HE, Janssen-Heijnen ML, Visser O, Verbeek AL, Kiemeney LA. The validity of the mortality to incidence ratio as a proxy for site-specific cancer survival. *European journal of public health.* 2011;21(5):573-577.
91. Parkin DM, Bray F. Evaluation of data quality in the cancer registry: principles and methods Part II. Completeness. *European journal of cancer.* 2009;45(5):756-764.
92. Babatunde OA, Adams SA, Eberth JM, Wirth MD, Choi SK, Hebert JR. Racial disparities in endometrial cancer mortality-to-incidence ratios among Blacks and Whites in South Carolina. *Cancer causes & control : CCC.* 2016;27(4):503-511.
93. Choi E, Lee S, Nhung BC, et al. Cancer mortality-to-incidence ratio as an indicator of cancer management outcomes in Organization for Economic Cooperation and Development countries. *Epidemiol Health.* 2017;39.
94. Liu S, Yang L, Yuan Y, et al. Cancer incidence in Beijing, 2014. *Chin J Cancer Res.* 2018;30(1):13-20.
95. Mak D, Sengayi M, Chen WC, Babb de Villiers C, Singh E, Kramvis A. Liver cancer mortality trends in South Africa: 1999-2015. *BMC Cancer.* 2018;18(1):798.
96. McDaniel JT, Nuhu K, Ruiz J, Alorbi G. Social determinants of cancer incidence and mortality around the world: an ecological study. *Glob Health Promot.* 2017:1757975916686913.

97. Pan R, Zhu M, Yu C, et al. Cancer incidence and mortality: A cohort study in China, 2008-2013. *International journal of cancer Journal international du cancer.* 2017;141(7):1315-1323.
98. Sung WW, Wang SC, Hsieh TY, et al. Favorable mortality-to-incidence ratios of kidney Cancer are associated with advanced health care systems. *BMC Cancer.* 2018;18(1):792.
99. Sunkara V, Hebert JR. The Colorectal Cancer Mortality-to-Incidence Ratio as an Indicator of Global Cancer Screening and Care. *Cancer.* 2015;121(10):1563-1569.
100. Tsai MC, Wang CC, Lee HL, et al. Health disparities are associated with gastric cancer mortality-to-incidence ratios in 57 countries. *World J Gastroenterol.* 2017;23(44):7881-7887.
101. Wang SC, Sung WW, Kao YL, et al. The gender difference and mortality-to-incidence ratio relate to health care disparities in bladder cancer: National estimates from 33 countries. *Sci Rep.* 2017;7(1):4360.
102. United States Department of Veterans Affairs. National Center for Veterans Analysis and Statistics. https://www.va.gov/vetdata/veteran_population.asp. Accessed June 2018.
103. Rural-Urban Commuting Area Codes. United States Department of Agriculture. Economic Research Service. https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/ . Accessed: September 2018.
104. WWAMI Rural Health Research Center. http://depts.washington.edu/uwruca/.