



Patient- and area-level predictors of prostate cancer among South Carolina veterans: a spatial analysis

Peter Georgantopoulos^{1,2,3,4} · Jan M. Eberth^{1,2} · Bo Cai² · Christopher Emrich⁵ · Gowtham Rao² · Charles L. Bennett^{2,3,4} · Kathlyn S. Haddock⁴ · James R. Hébert^{1,2}

Received: 1 November 2018 / Accepted: 21 December 2019 / Published online: 23 January 2020
© Springer Nature Switzerland AG 2020

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 and distance traveled to receive care, affect PrCA risk. The purpose of this research was to use the Veterans Administration Medical System, which provides a unique means for studying PrCA epidemiology among diverse individuals with ostensibly equal access to healthcare, to determine whether area-level characteristics influence PrCA incidence while accounting for individual-level risk factors.

Methods From the US Veteran's Health Administration (VHA) electronic medical records (EMR) database from January 1999 to December 2015, we identified 3,736 PrCA patients and 104,017 cancer-free controls from South Carolina (SC). The VHA EMRs were linked to the US census which provided area-level factors. US census data were used to construct the Social Vulnerability Index which is a continuous composite measure of area-level vulnerability and was divided into tertiles for modeling purposes. Data were analyzed using a Bayesian multivariate conditional autoregressive model (CAR) which accounted for individual-level factors, area-level factors, spatial random effects, and autocorrelation, which were used to identify areas of higher- or lower-than-expected PrCA incidence after controlling for risk factors.

Results As expected, after accounting for age (sixfold and 13-fold increases in men 40–50 years and > 50 years, respectively), race was an important risk factor, with threefold higher odds among Blacks in the fully adjusted model [OR_{adj} 2.98 (2.77, 3.20)]. After accounting for all other factors, residing in a ZIP code tabulated areas (ZCTA) with the greatest level social vulnerability versus the lowest, least vulnerable ZCTA's, increased PrCA risk by 39% [OR_{adj} 1.39 (1.11, 1.75)].

Conclusions While accounting for known risk factors for PrCA, 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.

Keywords Prostate cancer · Epidemiology · Incidence · Spatial epidemiology · Veteran · Veteran's Health Administration

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10552-019-01263-2>) contains supplementary material, which is available to authorized users.

✉ Peter Georgantopoulos
georgant@email.sc.edu

¹ South Carolina Statewide Cancer Prevention and Control Program, University of South Carolina, Columbia, SC, USA

² Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, SC, USA

Introduction

Prostate cancer (PrCA) is the most common cancer in men, accounting for 22% of all new cancer cases in American men [1]. Identifying novel risk factors associated with PrCA risk beyond established risk factors has been met

³ Southern Network On Adverse Reactions (SONAR), South Carolina Center of Economic Excellence for Medication Safety, College of Pharmacy, University of South Carolina, Columbia, SC, USA

⁴ Columbia VA Health Care System, Columbia, SC, USA

⁵ College of Health and Public Affairs, University of Central Florida, Orlando, USA

with little success. Established and actively researched risk factors include race, education, marital status, family history, diet, and age [1–6].

In the general United States (US) population, PrCA incidence displays a non-uniform geographical distribution, where in 2015 the national average was 99.1 with new PrCA cases per 100,000 [7]. The variation in PrCA incidence by state can be clearly seen by comparing the three highest states for PrCA incidence [i.e., New Jersey (127.4.7 per 100,000), Mississippi (126.7 per 100,000), and Louisiana (125.7 per 100,000)] to the three states with the lowest PrCA incidence [i.e., Alaska (61.0 per 100,000), Nevada (69.5 per 100,000), and New Mexico (72.6 per 100,000)] [7]. These differences across states may be potentially associated with macro-level factors such as area-level income, educational attainment, rurality, and environmental factors [8–10]. 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. Social determinants of health are defined as factors measured at the community level, but impact the health outcomes of individuals [11]. Identifying areas with higher- or lower-than-expected numbers of PrCA cases using spatial techniques [12–16], employing spatial nested models [8–10] or autoregressive models [17, 18], and using ecological factors of the geographical unit of investigation [8–10] can provide a more thorough understanding of PrCA by identifying the expanded universe of risk factors associated with PrCA [17–20]. 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 [21].

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 PrCA incidence that is higher than the national average [22]. This is a small geographical area as compared to other states, and the veteran population is broadly distributed throughout the state. 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.

Methods

Study population

This study employed a retrospective nested case–control 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 ZIP codes developed by the US Postal Service [23].

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 [24, 25]. 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) [24, 25]. 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 [26–28]. 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., ecology-level information), or directly obtained from the 2015 ZCTA shapefile from the US Census Bureau (i.e., individual ZCTA boundaries) [28].

All data elements were screened to exclude females. Subjects were excluded based on the following criteria (see Fig. 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 [24], 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 1 January 1999, date of birth differs by more than 365 days between VA MedSAS[®] datasets [24] and VA Vital Status datasets [25], 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 [24]. There were 235,782 unique PrCA

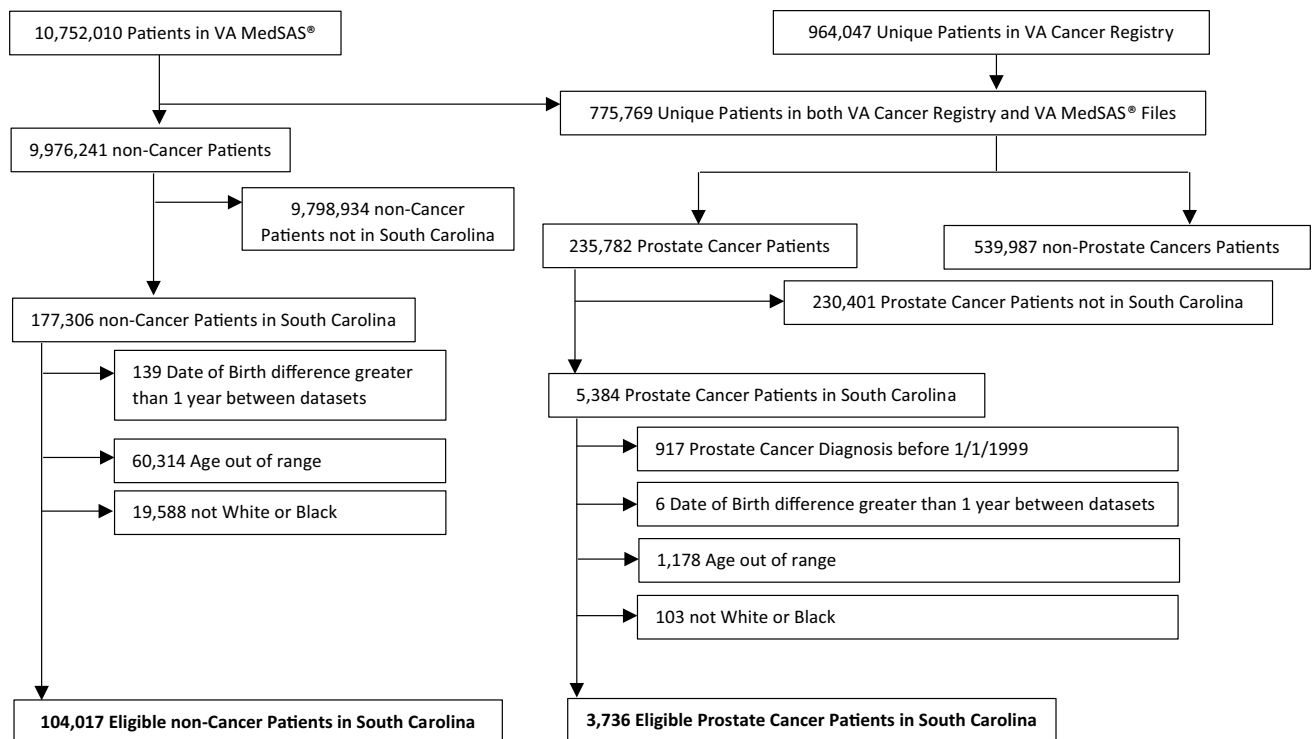


Fig. 1 STROBE diagram for the generation of the analytical cohort comparing prostate cancer patients to non-cancer patients among the South Carolina veteran population who seek healthcare at Vet-

eran Health Administration facilities between 1 January 1999 and 31 December 2015 and meet the study inclusion criteria

cases in the Primary Oncology Dataset that could be linked to the VA MedSAS® Files [24]. There were 230,401 (97.72%) PrCA cases that were excluded 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 1 January 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 [24]. These subjects were classified as not having cancer. There were 9,798,934 (98.22%) non-cancer cases that 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 (Fig. 1). South Carolina veterans receiving care at VHA facilities resided in 413 of South Carolina's 424 ZIP codes (97.41%). There were 67 (16.2%) ZIP codes without PrCA cases among the 413 South Carolina ZIP codes with veterans receiving care at VHA facilities in South Carolina.

Primary outcome variables

Primary PrCA patients between 1 January 1999 and 31 December 2015 were identified from the VA Primary Oncology Dataset [24, 25]. Non-cancer controls were identified by removing all cancer patients in the VA Primary Oncology Dataset from the VA MedSAS® Files [24, 25]. Non-cancer controls were then limited to those patients whose first VA visit occurred between 1 January 1999 and 31 December 2015.

Individual-level 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 [24, 25]. Four age strata variables were created: age < 40, age from 40 to 50, age from > 50 to 60, age from > 60 to 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 [25]. Subjects other than White or Black racial category were excluded because they accounted for 0.68% of the PrCA and non-cancer patients. 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 [24]. Marital status was categorized as married, previously married (divorced, widowed, and separated), and never married/unknown. The unique VA facility numerical codes in the MedSAS[®] datasets were used to identify which facility a subject visited [24]. The number of different VA facilities, the most frequented VA facility, and the number of times each subject visited their most visited VA facility were determined.

ZIP code-related Information

The ZIP code listed the most number of times for each subject's total visits in the MedSAS[®] datasets was used for all ZIP code-related information for each subject [24]. 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 ($n = 6,784$) [24].

There were 131 unique VA facilities (i.e., VA hospitals and VA Community-Based Outpatient Clinics) determined from VA MedSAS[®] datasets [24]. The distance between the most frequented VA facility and ZCTA-linked ZIP code listed the most times it was represented as a continuous variable and obtained using the MedSAS[®] datasets [24]. This ZIP code information was linked to the 2015 ZCTA shapefile from the US Census Bureau [28]. Using the shapefile, a contiguity straight line origin-to-destination distance matrix between VA facilities and ZCTA code centroids was used to determine the distance traveled by a veteran to his most frequented VA facility. The distance between the most-frequented VA facility and the ZCTA-linked ZIP code listed the most times it was categorized into: 0–25 miles, > 25–55 miles, and > 55 miles [24, 28].

Area-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 [29, 30].

The SoVI[®] uses 10 ZCTA measurements obtained from the 2010 US Census and 17 ZCTA measurements obtained from 2007 to 2011 5-Year American Community Survey [26, 27]. Selection of measurements was based on community-level factors that have shown broad associations across multiple health-related outcomes. The development of the SoVI[®] scores proceeded in accordance with previously

generated SoVI[®] scores for research purposes within the HVRI [29, 30].

The SoVI[®] is a relative measure represented on a continuous scale from negative infinity to positive infinity. A greater SoVI[®] score indicates the less able a geographical unit is prepared for, can respond to, and can recover from a disaster compared to a geographical unit with a lower SoVI[®] score. A SoVI[®] score was assigned for each subject's most frequently listed ZIP code. The SoVI[®] measurements were stratified into three categories using cutoffs of 1 standard deviations: low ($-\infty, -2.04$), medium ($-2.04, 2.79$), and high ($2.79, \infty$).

Other ZIP code-level factors include the following: 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%] [27].

Statistical analyses

Descriptive statistics were calculated for the analytic cohort (i.e., *t*-test for all continuous independent variables and chi-square tests for all categorical

independent variables). Frequentist multivariate general linear models using SAS[®] were first developed to inform key steps in the development and evaluation of the final models [31]. These models were developed for all individual-level risk factors and nested models incorporating both individual- and ecological-level risk factors with patient ZCTA-linked ZIP code residence. Nested models were evaluated as both fixed effect and 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- and 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 (i.e., extent of clustering between neighboring ZCTA-linked ZIP codes in the dataset).

All categorical variables were evaluated using created dummy variables. Any independent individual-level variables with undefined 95% confidence intervals or extremely large point estimates with extremely wide 95% confidence

intervals were excluded. 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 ecological-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 [32]. Subsequently, the ecological factor with the next-lowest AIC was added to the initial nested model. If any risk factor estimates changed by more than 10% then the additional ecological factor was retained; this was repeated for all ecological-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 calculated 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) [33]. 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 [21]. 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 non-significant, 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 general linear models were developed. 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 (Supplemental Fig. 1). The convergence of the sample chains was evaluated using the Geweke diagnostics, which compares the first 10% of kept iterations to the last 50% of kept iterations [34]. 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) [35]. The model with the lowest DIC was selected as the best-fitted model for that comparison in that analytical cohort.

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 [31]. All Bayesian modeling was conducted using WinBUGS[®] version 1.4.3 and R version 3.3.2 [35, 36]. All evaluation of Bayesian modeling was conducted using R [36]. 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 WJB Dorn VA (10,404).

Results

There were 3,736 PrCA cases in 377 ZCTA's with a median number of PrCA cases of eight and a range of 0 to 153 across those ZCTA's. These PrCA cases were distributed by year, as shown in Supplementary Table 1. There were 104,017 non-cancer cases in 413 ZIP codes with a median of 130 and a range of 1 to 2,044 across those ZIP codes. PrCA patients accounted for 3.5% of the final analytical cohort. The plurality of patients were between > 60 and 70 years of age (37.7%). Most PrCA patients were White (65.5%), married (62.3%), and traveled more than 55 miles to a VA facility for care (52.0%) (Table 1). The median distance traveled to a VA facility was 60.0 miles.

Univariate statistics

PrCA patients were more likely than non-PrCA patients to be older for their first VA visit (age from > 50 to 60 years: 45.1% vs. 34.0%; age from > 60 to 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 2).

Multivariable Bayesian modeling

Seven models of increasing complexity using a Bayesian framework were fit (Table 3). All independent variables were significant for the first model (Model 1, Table 3). The second model was the nested null model which accounted for the subject's ZCTA-linked ZIP code. This model had a MOR of 1.37 (1.30, 1.44) as well as the highest DIC of all seven

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

Categorical variables	<i>n</i> = 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)

models. The third model accounted for the individual-level factors as well as nesting the subjects in their ZCTA-linked ZIP codes (Model 3, Table 3). The fourth model included area-level factors including percent of the population living in poverty and categorized SoVI® scores (Model 4, Table 3). 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 (Model 6, Table 3). 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 (Model 7, Table 3).

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 (OR_{adj} 5.63 (3.94,

Table 2 Univariate χ^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 between 1 January 1999 and 31 December 2015 and meet the study inclusion criteria

Risk factors	All PrCA patients (<i>n</i> = 3,736)	Non-cancer patients (<i>n</i> = 104,017)	<i>p</i> -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%)	

8.39), >50–60 years old [OR_{adj} 13.11 (9.29, 19.53)], and >60–70 years old [OR_{adj} 13.31 (9.44, 19.81)]. PrCA patients were more likely to be Black [OR_{adj} 2.98 (2.77, 3.20)]. PrCA patients were more likely to have been married previously as compared to being currently married [OR_{adj} 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 [OR_{adj} 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 [OR_{adj} 1.06 (0.92, 1.21)]. This stratum for this variable was also statistically insignificant in Model 4, while it was 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 [OR_{adj} 0.86 (0.75, 0.98)]. This stratum for this variable was also statistically significant in Model 4, while it was 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 [OR_{adj} 1.35 (1.11, 1.65)] and 3 standard deviations from the ZCTA-linked ZIP codes with the lowest SoVI® score

Table 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 1 January 1999 and 31 December 2015 and meet the study inclusion criteria

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

Model 1: individual-level risk factors; Model 2: nested null model (median odds ratio: 1.37 95% credible interval (1.15, 1.26); DIC: 32,079); Model 3: nested model with individual-level risk factors; Model 4: full nested model; Model 5: spatial null model (median odds ratio: 1.23 95% credible interval (1.16, 1.31); DIC: 32,067); 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

[OR_{adj} 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 [OR_{adj} 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 [OR_{adj} 1.13 (0.93, 1.35)].

Discussion

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

The distance traveled to receive care evaluates the geographic accessibility among the veteran population and demonstrates a unique dynamic in attempting to quantify the impact of travel experienced by veterans given a travel stipend [37, 38]. 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-min to 1-h 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, for veterans traveling > 55 miles (i.e., a VA healthcare facility more than an approximate 1-h commute), it 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 [39]. The primary VA healthcare facilities in South Carolina are located in urban areas, which are limited in 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 PrCA resulting in a fewer PrCA cases being associated with that area [39]. The potential influence of a rural–urban disparity warrants the evaluation of measurements characterizing such a disparity, such as the rural–urban commuting area codes (RUCAs), in future studies based on PrCA incidence [40].

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 Clinic (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 provide. This analysis did not differentiate between VA hospitals and CBOCs. However, based on the plausible inference from 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 [41, 42]. 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 VA, these results indicate that veterans living within an approximate 30-min to 1-h commute to a VA healthcare facility maybe at increased risk for being diagnosed with an incident case of PrCA, 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-h 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 [39].

We also showed that established PrCA risk factors, such as marital status and being Black, were statistically significant factors associated with PrCA [1, 43–45]. 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 PrCA diagnosis for veterans. Veterans diagnosed with incident PrCA were more likely to begin their VA healthcare later in life than those veterans without incident PrCA or who did not develop incident PrCA during the 16-year study timeframe. Furthermore, the risk of incident PrCA more than doubled from the 40–50 to the > 50 –60-year-old strata, at which point there was slight differences in incident PrCA 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 PrCA, but may cause a bias when used as part of a larger comparison group to evaluate those chronic diseases at this point in time.

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 [46]. The data files used have been extensively utilized and queried in other VA-based peer-reviewed research and are of high reliability and validity [47–50]. 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 PrCA 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 [51]. 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.

Another consequence of using VA electronic medical records is that patient addresses could not be obtained and therefore the centroid of the ZCTA that each patient resided was used to determine the straight-line distance traveled by patients to their most frequented VA facility. This issue could not be avoided given the patient confidentiality protections [52]. However, given this limitation, any bias will be non-directional given the relatively small geographical area of each ZIP code, which would result in minor differences between the centroid distance to the VA facility and the actual distance. Furthermore, there would be equal probability that any random individual veteran can live closer or farther to their most frequented VA facility than the distance from the centroid. Therefore, the analytical cohort would have a narrow, normally distributed curve in the difference between their actual straight line distance to their most frequented VA facility and the centroid distance.

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 closest 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 PrCA epidemiology for two reasons. The first is that it was the smallest of the two geographical units available (the other being county). Therefore, while census tracts are more commonly used in spatial models in evaluating areas smaller than counties, such geographical units were not available in the VA records. The second reason is a consequence of the first. The more homogenous the population was, the more likely it is to differentiate between ZCTAs in the models and therefore to identify ZCTAs with higher- or 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 directional bias.

No South Carolina ZCTA 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 [53]. 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 who 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.

The use of ZCTA-level measurements from one time-frame is a limitation. However, the measurements from the 5-Year American Community Survey that were used individually and as components of the SoVI[®] and the SoVI[®] have been shown to remain stable over several consecutive years [54]. However, this analytical approach demonstrated the possibility of successfully modeling individual-level risk factors, area-level risk factors, and spatial autocorrelation within one modeling framework. This approach opens up the possibility to expand this research to include the incorporation of a temporal component into the modeling framework.

Additionally, "age at PrCA diagnosis" could not be used in the multivariate general linear models because the non-cancer comparison group does not have those values to calculate that measure. For this reason, the creation and use of "age at first VA visit" was calculated as novel approach that allowed age to be assessed between the PrCA population

and the non-cancer population for multivariate general linear modeling purposes. While this measures the age at which a veteran began accessing VHA care, any bias between the PrCA group and the non-cancer comparison group would most likely be non-directional; i.e., not biased toward the null. Chronic diseases other than PrCA (e.g., cardiovascular disease, type 2 diabetes mellitus) also are likely to occur in middle to late adulthood. Developing such chronic diseases is an equal probable in both the PrCA group and the non-cancer comparison group [55, 56]. It also is important to note that such non-cancer chronic diseases were not exclusion factors in this study.

Finally, in addition to the 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 ZCTA-linked 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.

Author contributions PG developed the study idea under the mentorship of JME, BO, GR, CLB, and JRH. PG obtained, managed, cleaned, and analyzed the data. PG drafted the manuscript and was responsible for incorporating and resolving all manuscript edits and comments. JRH and JME guided PG through all epidemiological aspects of the study. BO guided PG through all biostatistical aspects of the study. GR guided PG in acquiring and cleaning the data. CLB guided PG through all clinical considerations of the study. CE guided PG in developing the Social Vulnerability Index for ZIP code tabulated areas. KSH guided PG through all required Department of Veteran's Affairs requirements necessary to conduct a study using Veteran Health Administration data. JRH provided PG guidance in overseeing and managing all aspects of this study. JME, BO, GR, CLB, CE, KSH, and JRH reviewed, edited, and commented on all drafts of this manuscript.

Funding None.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Siegel RL, Miller KD (2018) Jemal A (2018) Cancer statistics. *CA Cancer J Clin* 68(1):7–30. <https://doi.org/10.3322/caac.21442>
2. Adami H-O, Hunter DJ, Trichopoulos D (2008) Textbook of cancer epidemiology. Monographs in epidemiology and biostatistics, 2nd ed. vol 36. Oxford University Press, Oxford, New York
3. Clegg LX, Reichman ME, Miller BA, Hankey BF, Singh GK, Lin YD, Goodman MT, Lynch CF, Schwartz SM, Chen VW, Bernstein L, Gomez SL, Graff JJ, Lin CC, Johnson NJ, Edwards BK (2009) 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* 20(4):417–435. <https://doi.org/10.1007/s10552-008-9256-0>
4. Hebert JR, Hurley TG, Olendzki BC, Teas J, Ma Y, Hampel JS (1998) Nutritional and socioeconomic factors in relation to prostate cancer mortality: a cross-national study. *J Natl Cancer Inst* 90(21):1637–1647. <https://doi.org/10.1093/jnci/90.21.1637>
5. McMahon DM, Burch JB, Hebert JR, Hardin JW, Zhang J, Wirth MD, Youngstedt SD, Shivappa N, Jacobsen SJ, Caan B, Van Den Eeden SK (2019) Diet-related inflammation and risk of prostate cancer in the California Men's Health Study. *Ann Epidemiol* 29:30–38. <https://doi.org/10.1016/j.annepidem.2018.10.008>
6. Mohseni R, Abbasi S, Mohseni F, Rahimi F, Alizadeh S (2019) Association between dietary inflammatory index and the risk of prostate cancer: a meta-analysis. *Nutr Cancer* 71(3):359–366. <https://doi.org/10.1080/01635581.2018.1516787>
7. Prostate Cancer in the United States in (2015) by State. Interactive Cancer Atlas, CDC
8. Klassen AC, Curriero FC, Hong JH, Williams C, Kulldorff M, Meissner HI, Alberg A, Ensminger M (2004) The role of area-level influences on prostate cancer grade and stage at diagnosis. *Prev Med* 39(3):441–448. <https://doi.org/10.1016/j.ypmed.2004.04.031>
9. Oliver MN, Smith E, Siadaty M, Hauck FR, Pickle LW (2006) Spatial analysis of prostate cancer incidence and race in Virginia, 1990–1999. *Am J Prev Med* 30(2 Suppl):S67–S76. <https://doi.org/10.1016/j.amepre.2005.09.008>
10. Xiao H, Gwede CK, Kiros G, Milla K (2007) Analysis of prostate cancer incidence using geographic information system and multi-level modeling. *J Natl Med Assoc* 99(3):218–225
11. National Center for Health Statistics (U.S.) (2012) Healthy people 2010: final review. DHHS publication, vol no (PHS) 2012–1038. U.S. Dept. of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville
12. DeChello LM, Gregorio DI, Samociuk H (2006) Race-specific geography of prostate cancer incidence. *Int J Health Geogr* 5:59. <https://doi.org/10.1186/1476-072X-5-59>
13. Hebert JR, Daguise VG, Hurley DM, Wilkerson RC, Mosley CM, Adams SA, Puett R, Burch JB, Steck SE, Bolick-Aldrich SW (2009) Mapping cancer mortality-to-incidence ratios to illustrate racial and sex disparities in a high-risk population. *Cancer* 115(11):2539–2552. <https://doi.org/10.1002/cncr.24270>
14. Hsu CE, Mas FS, Miller JA, Nkhoma ET (2007) A spatial-temporal approach to surveillance of prostate cancer disparities in population subgroups. *J Natl Med Assoc* 99(1):72–80
15. Wagner SE, Bauer SE, Bayakly AR, Vena JE (2013) Prostate cancer incidence and tumor severity in Georgia: descriptive epidemiology, racial disparity, and geographic trends. *Cancer Causes Control* 24(1):153–166. <https://doi.org/10.1007/s10552-012-0101-0>
16. Wagner SE, Hurley DM, Hebert JR, McNamara C, Bayakly AR, Vena JE (2012) Cancer mortality-to-incidence ratios in Georgia: describing racial cancer disparities and potential geographic determinants. *Cancer* 118(16):4032–4045. <https://doi.org/10.1002/cncr.26728>
17. Goovaerts P, Xiao H (2011) Geographical, temporal and racial disparities in late-stage prostate cancer incidence across Florida: a multiscale joinpoint regression analysis. *Int J Health Geogr* 10:63. <https://doi.org/10.1186/1476-072X-10-63>
18. Goovaerts P, Xiao H (2012) The impact of place and time on the proportion of late-stage diagnosis: the case of prostate cancer in Florida, 1981–2007. *Spat Spatio-temporal Epidemiol* 3(3):243–253. <https://doi.org/10.1016/j.sste.2012.03.001>

19. Zhou H, Lawson AB, Hebert JR, Slate EH, Hill EG (2008) A Bayesian hierarchical modeling approach for studying the factors affecting the stage at diagnosis of prostate cancer. *Stat Med* 27(9):1468–1489. <https://doi.org/10.1002/sim.3024>
20. Zhou H, Lawson AB, Hebert JR, Slate EH, Hill EG (2008) Joint spatial survival modeling for the age at diagnosis and the vital outcome of prostate cancer. *Stat Med* 27(18):3612–3628. <https://doi.org/10.1002/sim.3232>
21. Waller LA, Gotway CA (2004) Applied spatial statistics for public health data. Wiley series in probability and statistics, Wiley, Hoboken
22. Prostate Cancer in the United States in 2011–2015 by State. Interactive Cancer Atlas. CDC
23. ZIP Code™ Tabulation Areas (ZCTAs™). US Department of Commerce. <https://www.census.gov/geo/reference/zctas.html>
24. Department of Veterans Affairs. VA Medical SAS® Files (121VA10P2). VA Corporate Data Center Operations, Austin Information Technology Center, Austin
25. Department of Veterans Affairs. VA Vital Status File (121VA10P2). VA Corporate Data Center Operations, Austin Information Technology Center, Austin
26. United States Census Bureau. 2010 Census. U.S. Census Bureau. 2010. https://www2.census.gov/census_2010/. Accessed Aug 2014
27. United States Census Bureau. 2011 Five Year American Community Survey. U.S. Census Bureau. 2011. https://www2.census.gov/acs2011_5yr/. Accessed Aug 2014
28. United States Census Bureau. TIGER/Line® Shapefiles. <https://www.census.gov/cgi-bin/geo/shapefiles/index.php>. Accessed Aug 2014
29. Cutter SL, Boruff BJ, Shirley WL (2003) Social vulnerability to environmental hazards. *Soc Sci Q.* <https://doi.org/10.1111/1540-6237.8402002>
30. Schmidtlein MC, Deutsch RC, Piegorsch WW, Cutter SL (2008) A sensitivity analysis of the social vulnerability index. *Risk Anal* 28(4):1099–1114. <https://doi.org/10.1111/j.1539-6924.2008.01072.x>
31. SAS 9.4. SAS Institute Inc., Cary
32. Kleinbaum DG, Kupper LL, Nizam A, Rosenberg ES (2013) Applied regression analysis and other multivariable methods, 5th edn. Cengage Learning, Boston
33. Merlo J, Chaix B, Ohlsson H, Beckman A, Johnell K, Hjerpe P, Rastam L, Larsen K (2006) 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* 60(4):290–297. <https://doi.org/10.1136/jech.2004.029454>
34. Geweke J (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In Paper presented at the Bayesian Statistics, vol 4
35. Lunn D (2013) The BUGS book: a practical introduction to Bayesian analysis. texts in statistical science. CRC Press/Taylor & Francis Group, Boca Raton
36. Development Core Team R (2010) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
37. Penchansky R, Thomas JW (1981) The concept of access: definition and relationship to consumer satisfaction. *Med Care* 19(2):127–140
38. Delamater PL, Messina JP, Shortridge AM, Grady SC (2012) Measuring geographic access to health care: raster and network-based methods. *Int J Health Geogr* 11(1):15. <https://doi.org/10.1186/1476-072X-11-15>
39. Obertova Z, Brown C, Holmes M, Lawrenson R (2012) Prostate cancer incidence and mortality in rural men—a systematic review of the literature. *Rural Remote Health* 12(2):2039
40. WWAMI Rural Health Research Center. <https://depts.washington.edu/uwruca/>
41. Cheng I, Witte JS, McClure LA, Shema SJ, Cockburn MG, John EM, Clarke CA (2009) Socioeconomic status and prostate cancer incidence and mortality rates among the diverse population of California. *Cancer Causes Control* 20(8):1431–1440. <https://doi.org/10.1007/s10552-009-9369-0>
42. Gilligan T (2005) Social disparities and prostate cancer: mapping the gaps in our knowledge. *Cancer Causes Control* 16(1):45–53. <https://doi.org/10.1007/s10552-004-1291-x>
43. Coker AL, Sanderson M, Ellison GL, Fadden MK (2006) Stress, coping, social support, and prostate cancer risk among older African American and Caucasian men. *Ethn Dis* 16(4):978–987
44. Newell GR, Pollack ES, Spitz MR, Sider JG, Fueger JJ (1987) Incidence of prostate cancer and marital status. *J Natl Cancer Inst* 79(2):259–262
45. Tyson MD, Andrews PE, Etzioni DA, Ferrigni RG, Humphreys MR, Swanson SK, Castle EK (2013) Marital status and prostate cancer outcomes. *Can J Urol* 20(2):6702–6706
46. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstein EV, Lehmann HP, Hripscak G, Hartzog TH, Cimino JJ, Saltz JH (2013) Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care* 51(8 Suppl 3):S30–S37. <https://doi.org/10.1097/MLR.0b013e31829b1dbd>
47. Rao GA, Mann JR, Bottai M, Uemura H, Burch JB, Bennett CL, Haddock KS, Hebert JR (2013) Angiotensin receptor blockers and risk of prostate cancer among United States veterans. *J Clin Pharmacol* 53(7):773–778. <https://doi.org/10.1002/jcph.98>
48. Rao GA, Mann JR, Shoaibi A, Bennett CL, Nahhas G, Sutton SS, Jacob S, Strayer SM (2014) Azithromycin and levofloxacin use and increased risk of cardiac arrhythmia and death. *Ann Fam Med* 12(2):121–127. <https://doi.org/10.1370/afm.1601>
49. Rao GA, Mann JR, Shoaibi A, Pai SG, Bottai M, Sutton SS, Haddock KS, Bennett CL, Hebert JR (2013) Angiotensin receptor blockers: are they related to lung cancer? *J Hypertens* 31(8):1669–1675. <https://doi.org/10.1097/HJH.0b013e3283621ea3>
50. Shoaibi A, Rao GA, Cai B, Rawl J, Haddock KS, Hebert JR (2017) Prostate specific antigen-growth curve model to predict high-risk prostate cancer. *Prostate* 77(2):173–184. <https://doi.org/10.1002/pros.23258>
51. Grubesic TH, Matisziw TC (2006) On the use of ZIP codes and ZIP code tabulation areas (ZCTAs) for the spatial analysis of epidemiological data. *Int J Health Geogr* 5:58. <https://doi.org/10.1186/1476-072X-5-58>
52. United States. Congress. Senate. Committee on Labor and Human Resources. (1998) Health Insurance Portability and Accountability Act of 1996, first-year implementation concerns: hearing before the Committee on Labor and Human Resources, United States Senate, One Hundred Fifth Congress, second session ... 19 March 1998. S hrg, vol 105–490. U.S. G.P.O.: For sale by the U.S. G.P.O., Supt. of Docs., Congressional Sales Office, Washington
53. United States. Congress (104th 2nd session: 1996) (1996) Health Insurance Portability and Accountability Act of 1996: conference report (to accompany H.R. 3103). Report/104th Congress, 2nd session, House of Representatives 104-736. U.S. G.P.O., Washington

54. Cutter SL, Finch C (2008) Temporal and spatial changes in social vulnerability to natural hazards. *Proc Natl Acad Sci USA* 105(7):2301–2306. <https://doi.org/10.1073/pnas.0710375105>
55. Fryar CD, Herrick K, Afful J, Ogden CL (2016) Cardiovascular disease risk factors among male veterans, U.S., 2009–2012. *Am J Prev Med* 50 (1):101–105. <https://doi.org/10.1016/j.amepre.2015.06.011>
56. Liu Y, Sayam S, Shao X, Wang K, Zheng S, Li Y, Wang L (2017) Prevalence of and trends in diabetes among veterans, United States, 2005–2014. *Prev Chronic Dis* 14:E135. <https://doi.org/10.5888/pcd14.170230>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.