The Relative Importance of Race Compared to Healthcare and Social Factors in Predicting Prostate Cancer Mortality: A Random Forest Approach

Heidi A. Hanson, PhD[1,2], Claire L. Leiser, MSPH[3]; Christopher Martin, MD[1]; Brock O'Neil, MD[1]; William T. Lowrance, MD, MPH[1]; Ken R. Smith, PhD[3,4]

[1] Division of Urology, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT; [2]Department of Surgery and Population Sciences, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT [3]Department of Population Sciences, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT; [4]Department of Family and Consumer Studies, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT

Word Count: Abstract 297 (300 max), main text 2858 (3000 max)

Corresponding Author:
Heidi A. Hanson
University of Utah
2000 Circle of Hope, Rm 1501
Salt Lake City, UT 84112
Heidi.hanson@hci.utah.edu

Key Points

Question: What is the relative importance of African-American race in comparison to healthcare and social factors in predicting prostate cancer mortality?

Findings: In this retrospective cohort study of 514,878 men with a PCa diagnosis from 2004–2012, we found that race had a significant impact in many age groups and disease stages. However, healthcare access and quality, and social factors had greater or similarly important effects across all ages and stages.

Meaning: Bridging the gap in racial disparities for prostate cancer will require extending beyond a focus of biological differences and considering social disparities including healthcare access and quality.

Abstract

Importance: African-American men have increased risks for prostate cancer (PCa) mortality with both biological and socio-environmental mechanisms theorized. With personalized medicine, focus has increasingly shifted to underlying genetic differences. However, the non-biologic factors may play a larger role in the observed disparities.

Objective: To measure the relative importance of African-American race in comparison to healthcare and social factors in predicting PCa specific mortality.

Design: Using the Surveillance, Epidemiology, and End Results (SEER) Database, we identified 514,878 men with PCa diagnosed at age 40 or older between 2004-2012. We selected a subset of patients matching African-American men to white men by birth year, stage at diagnosis, and age at diagnosis. We stratified patients by age group (40-54, 55-69, 70+) and disease stage resulting in 18 groups. Applying random forest methods with variable importance measures, we analyzed four broad categories of factors (tumor characteristics, race, healthcare factors, and social factors) and their relative importance for PCa specific mortality for the matched subset and the cohort overall.

Setting: SEER registry sites

Participants: Men with PCa in SEER database

Interventions: None

Main Outcome and Measure: PCa-specific mortality

Results: Tumor characteristics at time of diagnosis were overwhelmingly the most important factors associated with risk of mortality. Across all groups, race was less than 5% as important as tumor characteristics, and only more important than healthcare and social factors for 2/18 groups in our analyses. Although race had a significant impact in many age groups and disease stages, healthcare and social factors, known to also be associated with racial disparities, had greater or similarly important effects across all ages and stages.

Conclusions: Eradicating disparities in prostate cancer survival will take a multi-pronged approach including advances in precision medicine. However, disparities will persist unless healthcare access and social equality is achieved among all populations.

Keywords: prostate cancer, SEER. Racial disparities, survival, access-to-care

Introduction

African-American men, have a 1.6 times higher incidence of prostate cancer (PCa) and more than twice the rate of PCa mortality compared to white men in the United States.[1] PCa death rates have steadily declined since the 1990's, a phenomenon commonly attributed to improved diagnostic and treatment regimens.[2,3] However, the African-American/white PCa mortality differential has slightly increased during this time, in contrast to a decrease in the differential for overall cancer mortality.[4] Genetic differences in PCa susceptibility are not likely to contribute to increasing disparities. It is more likely a combination of genetic and environmental factors, including underlying social and healthcare factors.

Genetic differences in PCa susceptibility and tumor behavior have been associated with increased PCa mortality for African-American men. African-American men show differences in TGFβ signaling[5] and an increased rate of susceptibility loci associated with PCa.[6] Additionally, African-American individuals have higher prostate-specific antigen (PSA) values for equivalently staged cancer.[7] Those with very low-risk disease have been shown to have an increased risk of clinical upstaging or progression to treatment on active surveillance,[8] as well as adverse pathology after prostatectomy.[9]

Racial disparities for PCa due to exogenous factors, such as healthcare access and quality have also been well documented.[10] Compared to White men, African-American men have been shown to have higher stage disease at the time of diagnosis,[11] particularly among younger age groups.[12] While some studies find that racial differences in survival can be explained by access to care and sociodemographic variables,[13,14] other find these factors are only partial mediators.[15,16] Fully understanding the interplay between biologic and social factors and their respective contribution to the observed differences in mortality is essential for reducing racial disparities in prostate cancer mortality.

Minimizing racial disparities due to genetic differences requires individualized race-specific and genetic approaches to care, while decreasing the social and healthcare factors requires more equitable distribution of resources. A firmer understanding of the impact and complex interplay of disease-specific characteristics, race, healthcare access and quality, and other social factors are necessary to identify the determining factors for prostate cancer mortality and to improve outcomes for African-American men. Identifying the most important factors for predicting prostate cancer mortality requires methodological approaches that take into account multiple variables (and their interactions) simultaneously, as each factor may be highly correlated and create heretofore unappreciated synergies.

Using patient cohorts from the Surveillance Epidemiology and End-Results (SEER) database, we take advantage of random forest regression, a supervised machine learning method, to identify the factors that are strong predictors of prostate cancer mortality. These methods allow measurement of the relative importance of each factor while exploring all possible interactions and selection of the most predictive factors. We hypothesize that tumor characteristics (Gleason score and PSA) will be the strongest predictors of PCa mortality among all stages of disease. We also hypothesize that race will have a measurable impact on lethality in many age groups and disease stages, although it will have a relatively smaller impact compared to tumor characteristics, and healthcare factors (access and quality) and social factors (median family income, population density, and social vulnerability), will be similarly impactful compared to race.

Methods

Data source

Data for this study are captured by the 18 registries comprising the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI).[16] This study includes all males over age 40 (at time of diagnosis) with a PCa diagnosis from 2004–2012 (N = 514,878). County-level healthcare data were obtained from the Department of Health and Human Services Area Health Resource File (AHRF), Center for Disease Control (CDC), and Agency for Healthcare and Research Quality (AHRQ). County-level measurements are based on an individual's county of residence at time of diagnosis.

Sample selection

Figure 1 displays the sample selection tree. We excluded individuals with a diagnosis at time of death or autopsy. We also required individuals to link to the Area Health Resource File (AHRF), have a known stage at time of diagnosis, and link to county-level measures of socioeconomic status.

FIGURE 1 HERE

Measures

SEER Data for individual level measures: Time was measured as months from PCa diagnosis to PCa death and observations were treated as censored at time of death from all other causes or end of the follow-up period. Race/ethnicity was grouped into the following categories; African-American, non-Hispanic white, Hispanic, other race, or unknown race/ethnicity. Staging was coded using the SEER Derived American Joint Committee on Cancer (AJCC) 6th edition clinical stage from 2004 to the present.[16] AJCC Stage I diagnoses were relatively rare, accounting for 0.1% of all diagnoses, and were combined with Stage II diagnoses. Tumor grade was based on the International Society of Urological Pathology grade group system.[17] PSA values were grouped into <4ng/mL, 4-10ng/mL, 10-20ng/mL, and >20ng/mL.

Measures of access to healthcare at the county level were drawn from the AHRF database and included number of physicians, radiation oncologists, urologists, and number of chemotherapy treatment centers. We compared counties with one or more urologists to those with none.[14, 15] We expected a non-linear relationship between physician density and PCa mortality; therefore, we compared counties in the 25th and 75th percentiles to the middle 50% of the distribution. The AHRQ Prevention Quality Indicator (PQI) was used to measure the quality of ambulatory care in the county.

Social factors that may affect prostate cancer mortality were derived from multiple sources. Rural/Urban designation was identified from the AHRF. County-level median family income collected from the 2000 U.S. Census was also included in the models. We also considered the proportion of the Medicare population in each county that is also eligible for Medicaid enrollment ("dual enrollees").[18] The CDC Social Vulnerability Index (SVI) was included in the models.[18] SVI measures the resilience of communities when confronted by external stresses on human health, stresses such as natural or human-caused disasters, or disease outbreaks.[18]

Statistical Methods

Random forests (RF) are an extension of Classification and Regression Tree (CART) modeling. In this method, *n* trees are grown using bootstrapped samples from the learning sample.[19,20] As these trees are grown each branch point, called a node, occurs to separate the groups based on the factor that best explains the variability in mortality within the group. For example if Gleason score of >8 explains more of the variance in mortality identified within the group compared to PSA values >20 then at the next node the single group will separate into two groups based on Gleason score, regardless of PSA values and other factors. Unlike CART, there is no trimming or stopping criterion, the trees are fully grown. Additionally, a random subset of variables is selected for inclusion at each node. This method of random subspace selection is done to avoid correlation between trees in the forest and decreases the risk of multicollinearity. It also allows for the identification of the most relevant variables when multicollinearity is present and therefore reduces the variables of interest to those with the most explanatory value.[21] One benefit of the RF method is the ability to quantify the variable importance.

Since each tree is a random 70% subset of the original dataset, the remaining 30% of the data not selected (e.g., out of bag observations (oob)) can be used to calculate the variable importance. The oob data is used to create permutation accuracy variable importance measure (VIMP) by predicting class membership in the oob sample and then randomly permuting the values and calculating the decrease in predictive accuracy with permuted variables (it is also known as the Mean Decrease in Accuracy (MDA) measure). The average difference in accuracy of the oob versus permuted oob observations over all trees is the VIMP, with a VIMP close to zero implying that the variable has no predictive power.

Models were stratified by age at diagnosis (40 - 54, 55 - 69, and 70+) and stage at diagnosis (Stage I/II, Stage III, and Stage IV). This allowed us to look for differences within stage of diagnosis, therefore netting out racial differences in prostate cancer survival due to stage at diagnosis. All analyses were run in R using the *randomForestSRC* and *survival* packages. We performed our analysis with a subset of the sample in which all African-American individuals were matched to a non-Hispanic white counterpart by birth year, stage at diagnosis, and age at diagnosis. This approach allowed us to compare the importance of the variables without it being biased by the unequal distribution of race in the total population as well as directly compare non-Hispanic whites to African-Americans.

We included a total of 15 variables in each model. For descriptive purposes, variables were binned into four broad categories; tumor characteristics, social factors, healthcare factors, and race. To create a measure of relative importance, VIMPs were summed across categories and then compared to the sum of tumor characteristic VIMPs in order to standardize the measure across models stratified by age and stage. We then repeated these analyses with the total patient cohort to measure the impact of race more broadly.

Results

Descriptive statistics of all 15 variables are displayed in Table 1 by age at diagnosis. PCa was diagnosed in 55,493 men ages of 40–54, 282,358 men ages 55-69, and 174,213 ages 70+. There was more racial heterogeneity in the youngest age groups relative to men 70+. Men in the 40-54 age category were also more likely to have a lower stage and grade at diagnosis, and reside in higher income, metropolitan counties.

TABLE 1 HERE

Tumor characteristics (Gleason and PSA) were the most important predictors of survival across all ages and stages. The range of importance for each individual factor is included in Supplementary Tables 1 - 6. Figure 2 shows variable importance measures (VIMP) for social factors, healthcare factors, and race relative to the VIMP for tumor characteristics by age and stage (1/2, 3, and 4). Panel A includes the results for the matched subset analysis of African-Americans and non-Hispanic whites, while Panel B includes results for the population more broadly.

Across both analyses and all stages and age groups, race was never greater than 1/20th as important as tumor characteristics. In fact, in many subgroups, particularly among younger patients and stage 4 disease, race was not a measurably important factor for mortality. Race was only more important than healthcare and social factors for 2 of the 18 groups from our analyses (stage 3, age 55-69 and 70+ in the matched subset analysis). Comparatively, healthcare factors were important predictors of prostate cancer mortality across all ages and stages and social factors were important predictors of mortality among every group except the stage 3, 40-54 age group in the matched subset analysis.

There are interesting differences in the pattern of variable importance in the overall cohort compared to the matched subset. For the matched subset, social factors are more important in lower stages for men age 55-69 and in general, factors other than tumor characteristics are less important as stage increases. Comparatively, in the full cohort, healthcare and social factors are less important relative to tumor characteristics at lower stages of the disease and in general these factors increase in importance as stage increases, particularly among the younger age groups.

FIGURE 2 HERE

Discussion

Tumor characteristics were the most important factors for predicting prostate cancer mortality for all ages and stages. All other factors combined ranged from 1/4th to 1/10th as important when directly compared to tumor characteristics. Healthcare factors, social factors and race were important predictors of prostate cancer mortality, but the relative importance varied between the different age groups and disease stages. While race was a measurably important factor in predicting prostate cancer mortality, the healthcare and social factors (factors also known to associated with racial disparities) were more important for all but 2 of the 18 groups in our analysis.

The low importance of race in this analysis brings to question the benefit of differential treatment for African-American men regarding prostate cancer. Many experts have included African-American race as an independent risk factor with prostate cancer management, potentially exposing them to variations in screening or treatment based on these risks.[22-24] Personalized medicine has largely focused on the biological variables that can be used to individualize health outcomes, but personalized medicine that fails to incorporate the framework of socially based health disparities will likely fail to improve outcomes.[25] The overall effect of race as a determinant of prostate cancer outcomes is not solely biological, but instead may be tied to the social milieu that is often tied to health outcomes in general. Patient genetics and tumor biology among African-American patients are inextricably linked to a population with poorer health access and less resources.[26]

Leveraging random forests, we were able to disentangle the complex interplay of tumor, race, healthcare and social factors. Traditional regression approaches, which require all interactions be specified a priori are not well suited for the exploratory analyses needed to identify variables involved in the complex interplay between the individual and broader social context. When disentangling biology from social and healthcare factors using traditional regression methods one needs to specify interactions to test for differences in the effect of individual and neighborhood characteristics on prostate cancer mortality and predetermine the number of interactions. Fundamentally interactions involving multiple variables are difficult to interpret and introduce multicollinearity.

Random forests allow us to explore the complex associations between multiple variables and measure the importance of each variable in predicting prostate cancer mortality. This highly predictive modeling approach considers all possible combinations of variable interactions and addresses multicollinearity issues. This data driven, “reverse-engineering” approach allows us to disentangle the role of individual and neighborhood factors that are correlated with race/ethnicity and identify the strength of their contribution to driving prostate cancer outcomes. These models are known to have better predictive power than traditional regression methods and can identify complex hidden relationships in the data, especially for questions such as this where there is high collinearity between measures and decomposition of the importance of the effects is difficult.[27]

There are significant challenges to measuring the impact of race on prostate cancer mortality and it is important to note that the methods used in this analysis are designed to optimize prediction and not assess causation. The predictability of our models, which ranged from 0.75 – 0.85 for the different age groups and disease stages, suggests that the factors we considered explain a large amount of variation in prostate cancer mortality within this cohort. These models suggest that while race is often an important predictor in prostate cancer mortality, risk is also tied up in social forces that influence these outcomes. Additionally, the slight variation between the matched subset analysis and the cohort overall suggests that the pattern of racial disparities among other groups may be different than those found for African-American individuals. Whatever these specific patterns it is clear that one of the fundamental methods used to reduce racial disparities in prostate cancer outcomes has to involve an elimination of racial disparities for social and healthcare factors as well.

There are important considerations with the interpretation of our analyses. While it is likely that fundamental differences exist in regards to prostate cancer incidence and disease behavior, it is currently difficult for clinicians to incorporate race as a factor when clinically managing patients with a PCa diagnosis, particularly when considering confounding factors such as access and quality of care and the paucity of clinical research devoted to African-American patients. Our analyses do not suggest that race is unimportant with regards to prostate cancer mortality in general. As our cohort consisted of patients already diagnosed with prostate cancer of a given stage, our analysis does not consider differences in incidence of prostate cancer or variation of stage at diagnosis between races, both of which are known disparities among African-Americans.[1,11,12] However, for a patient with equivalent disease specific factors it appears that healthcare access and quality and other social factors known to be associated with race are as important (if not more important) when considering risks for PCa-specific mortality.

This novel approach is an important step towards disentangling the relationship of biology and other factors with regards to race. While we cannot determine if each of the confounding factors associated with race (independent of biology or genetics) have been isolated in this analysis, the low impact of race in many of the patient age groups and stages suggests that many of factors associated with race and prostate cancer mortality have been identified and delineated. Our next step will be incorporation of incidence and stage at diagnosis in similar modeling and characterizing the importance healthcare and social factors associated with those racial disparities.

Conclusions

Attempting to reduce racial disparities through a personalized medicine driven approach without addressing racial disparities in social and healthcare factors may close the racial gap for the economically advantaged, while reinforcing disparities for the disadvantaged. Although race had a significant impact in many age groups and disease stages, healthcare factors and social factors had greater or similarly important effects across most stages and age groups. It is extremely difficult to completely disentangle these factors, however, a reduction in the disparities in social factors and access to healthcare may provide the largest gains in improving prostate cancer mortality and minimizing racial disparities.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians.* 2018;68(1):7-30.
2. Hankey BF, Feuer EJ, Clegg LX, et al. Cancer Surveillance Series: Interpreting Trends in Prostate Cancer—Part I: Evidence of the Effects of Screening in Recent Prostate Cancer Incidence, Mortality, and Survival Rates. *JNCI: Journal of the National Cancer Institute.* 1999;91(12):1017-1024.
3. C. CK, E. TR, P. FH. Trends in prostate cancer mortality among black men and white men in the United States. *Cancer.* 2003;97(6):1507-1516.
4. E. DC, L. SR, Goding SA, et al. Cancer statistics for African Americans, 2016: Progress and opportunities in reducing racial disparities. *CA: A Cancer Journal for Clinicians.* 2016;66(4):290-308.
5. Elliott B, Zackery DL, Eaton VA, et al. Ethnic differences in TGFbeta-signaling pathway may contribute to prostate cancer health disparity. *Carcinogenesis.* 2018;39(4):546-555.
6. Lachance J, Berens AJ, Hansen MEB, Teng AK, Tishkoff SA, Rebbeck TR. Genetic hitchhiking and population bottlenecks contribute to prostate cancer disparities in men of African descent. *Cancer research.* 2018.
7. Vijayakumar S, Winter K, Sause W, et al. Prostate-Specific Antigen Levels are Higher in African-American Than in White Patients in a Multicenter Registration Study: Results of RTOG 94-12. *International Journal of Radiation Oncology*Biology*Physics.* 1998;40(1):17-25.
8. Abern MR, Bassett MR, Tsivian M, et al. Race is associated with discontinuation of active surveillance of low-risk prostate cancer: results from the Duke Prostate Center. *Prostate cancer and prostatic diseases.* 2013;16(1):85-90.
9. Sundi D, Ross AE, Humphreys EB, et al. African American Men With Very Low–Risk Prostate Cancer Exhibit Adverse Oncologic Outcomes After Radical Prostatectomy: Should Active Surveillance Still Be an Option for Them? *Journal of Clinical Oncology.* 2013;31(24):2991-2997.
10. McClelland S, Page BR, Jaboin JJ, Chapman CH, Deville C, Thomas CR. The pervasive crisis of diminishing radiation therapy access for vulnerable populations in the United States, part 1: African-American patients. *Advances in Radiation Oncology.* 2017;2(4):523-531.
11. Powell IJ, Bock CH, Ruterbusch JJ, Sakr W. Evidence supports a faster growth rate and/or earlier transformation to clinically significant prostate cancer in black than in white American men, and influences racial progression and mortality disparity. *J Urol.* 2010;183(5):1792-1796.
12. He T, Mullins CD. Age-Related Racial Disparities in Prostate Cancer Patients: A Systematic Review. *Ethnicity & health.* 2017;22(2):184-195.
13. Optenberg SA, Thompson IM, Friedrichs P, Wojcik B, Stein CR, Kramer B. Race, treatment, and long-term survival from prostate cancer in an equal-access medical care delivery system. *Jama.* 1995;274(20):1599-1605.
14. Schwartz K, Powell IJ, Underwood W, 3rd, George J, Yee C, Banerjee M. Interplay of race, socioeconomic status, and treatment on survival of patients with prostate cancer. *Urology.* 2009;74(6):1296-1302.
15. Bernard B, Muralidhar V, Chen YH, et al. Impact of ethnicity on the outcome of men with metastatic, hormone-sensitive prostate cancer. *Cancer.* 2017;123(9):1536-1544.
16. Bach PB, Schrag D, Brawley OW, Galaznik A, Yakren S, Begg CB. Survival of blacks and whites after a cancer diagnosis. *Jama.* 2002;287(16):2106-2113.
17. Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic

Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. *The American journal of surgical pathology.* 2016;40(2):244-252.

18. Barry E. Flanagan EWG, Elaine J. Hallisey, Janet L. Heitgerd, and, Lewis B. A Social Vulnerability Index for Disaster Management. *Journal of Homeland Security and Emergency Management.* 2011;8(1).
19. Breiman L. Random Forests. *Machine Learning.* 2001;45(1):5-32.
20. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat.* 2008;2(3):841-860.
21. Archer KJ, Kimes RV. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis.* 2008;52(4):2249-2260.
22. Powell IJ, Banerjee M, Sakr W, et al. Should African-American men be tested for prostate carcinoma at an earlier age than white men? *Cancer.* 1999;85(2):472-477.
23. Saltzman AF, Luo S, Scherrer JF, Carson KD, Grubb RL, Hudson MLA. Earlier prostate-specific antigen testing in African American men—Clinical support for the recommendation. *Urologic Oncology: Seminars and Original Investigations.* 2015;33(7):330.e339-330.e317.
24. Shenoy D, Packianathan S, Chen AM, Vijayakumar S. Do African-American men need separate prostate cancer screening guidelines? *BMC Urology.* 2016;16:19.
25. Brothers KB, Rothstein MA. Ethical, legal and social implications of incorporating personalized medicine into healthcare. *Personalized medicine.* 2015;12(1):43-51.
26. Kirby JB, Kaneda T. Unhealthy and Uninsured: Exploring Racial Differences in Health and Health Insurance Coverage Using a Life Table Approach. *Demography.* 2010;47(4):1035-1051.
27. Chen X, Ishwaran H. Random forests for genomic data analysis. *Genomics.* 2012;99(6):323-329.

Table 1. Descriptive Statistics for Full Sample

	Age 40 - 54		**Age 55 - 69**		**Age 70+**	
	n= 55,493		n =282,358		n = 174,213	
	n	%	n	%	n	%
Individual level variables						
Race/ethnicity						
White	34,472	62.12%	194,353	68.83%	124,499	71.46%
African-American	12,794	23.06%	45,310	16.05%	19,336	11.10%
Hispanic	5,045	9.09%	23,403	8.29%	14,665	8.42%
Other	1,827	3.29%	13,151	4.66%	10,725	6.16%
Unknown	1,355	2.44%	6,141	2.17%	4,988	2.86%
Stage						
I/II	46,812	84.36%	239,335	84.76%	148,742	85.38%
II	5,252	9.46%	26,170	9.27%	7,752	4.45%
IV	3,429	6.18%	16,853	5.97%	17,719	10.17%
Prostate Cancer Mortality						
Alive or dead of other cause of death	53,837	97.02%	272,755	96.60%	157,912	90.64%
Prostate cancer death	1,656	2.98%	9,603	3.40%	16,301	9.36%
Tumor Characteristics						
International Society of Urological Pathology Grade Group; Gleason Equivalent						
1; <=6	28,475	51.31%	130,482	46.21%	62,554	35.91%
2; 7 (3,4)	15,719	28.33%	77,089	27.30%	40,041	22.98%
3; 7 (4,3)	4,578	8.25%	29,407	10.41%	21,293	12.22%
4; 8	2,608	4.70%	20,253	7.17%	20,555	11.80%
5; 9 & 10	2,416	4.35%	16,276	5.76%	18,129	10.41%
Unknown	1,697	3.06%	8,851	3.13%	11,641	6.68%
PSA Category						

<4.0	40,214	72.47%	197,207	69.84%	90,652	52.04%
4.0 – 9.9	5,413	9.75%	32,914	11.66%	29,393	16.87%
10.0 – 19.9	4,306	7.76%	23,209	8.22%	27,101	15.56%
20.0	5,560	10.02%	29,028	10.28%	27,067	15.54%
Social Factors						
Metro						
Rural	2,055	3.70%	13,476	4.77%	9,416	5.40%
Metro	50,385	90.80%	249,820	88.48%	151,630	87.04%
Micro	3,053	5.50%	19,062	6.75%	13,167	7.56%
Median Family Income						
Mean (std dev)	54920 (13,185)		55,000 (13,278)		54,390 (13,255)	
Dual Enrollment						
Bottom 90%	50,470	90.95%	256,344	90.79%	157,489	90.40%
Top 10%	5,023	9.05%	26,014	9.21%	16,724	9.60%
Social Vulnerability						
Low social vulnerability (<0.75)	36,433	65.65%	185,035	65.53%	112,477	64.56%
High social vulnerability (≥0.75)	19,060	34.35%	97,323	34.47%	61,736	35.44%
Access to Healthcare						
Urologists (Urologists per 100,000 people)						
Zero urologists	5,447	9.82%	30,047	10.64%	18,890	10.84%
1+ Urologist	50,046	90.18%	252,311	89.36%	155,323	89.16%
Number of Chemotherapy Treatment Centers						
1+ chemotherapy treatment centers	48,870	88.07%	245,253	86.86%	150,976	86.66%
Zero chemotherapy treatment centers	6,623	11.93%	37,105	13.14%	23,237	13.34%
Number of doctors						
Bottom 25%	12,828	23.12%	70,428	24.94%	45,946	26.37%
Middle 50%	27,267	49.14%	139,472	49.40%	85,738	49.21%
Top 25%	15,398	27.75%	72,458	25.66%	42,529	24.41%
Number of radiation/Oncologists						

Bottom 25%	13,527	24.38%	71,247	25.23%	44,926	25.79%
Middle 50%	26,378	47.53%	136,326	48.28%	85,874	49.29%
Top 25%	15,588	28.09%	74,785	26.49%	43,413	24.92%
Prevention Quality Index						
Top 75% (County Level)	47,772	86.09%	244,806	86.70%	150,583	86.44%
Bottom 25% (County Level)	7,721	13.91%	37,552	13.30%	23,630	13.56%
Prevention Quality Index - African-American						
Top 75% (County Level)	37,145	66.94%	194,487	68.88%	119,849	68.79%
Bottom 25% (County Level)	18,348	33.06%	87,871	31.12%	54,364	31.21%

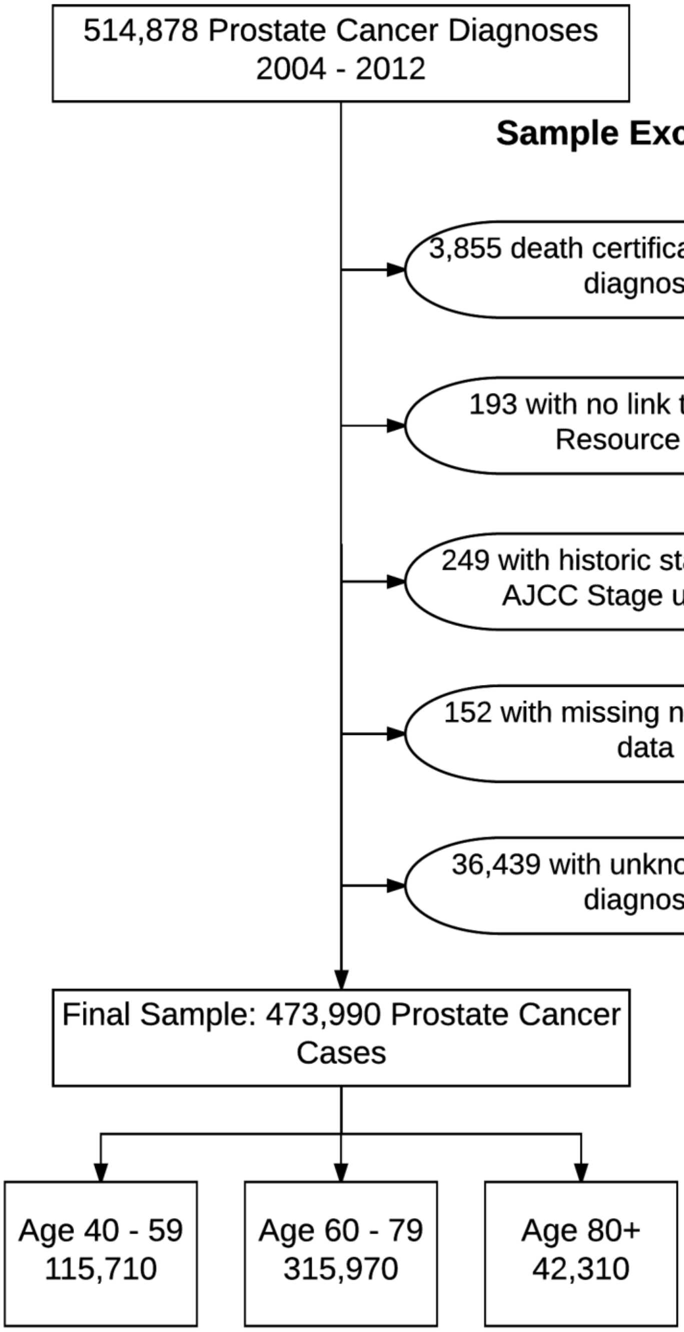

514,878 Prostate Cancer Diagnoses
2004 - 2012
Sample Exclusions
3,855 death certificate or autopsy diagnosis
193 with no link to the Area Resource File
249 with historic stage in situ or AJCC Stage unknown
152 with missing neighborhood data
36,439 with unknown stage at diagnosis
Final Sample: 473,990 Prostate Cancer Cases
Age 40 - 59
115,710
Age 60 - 79
315,970
Age 80+
42,310

Figure 1. Sample Selection Tree

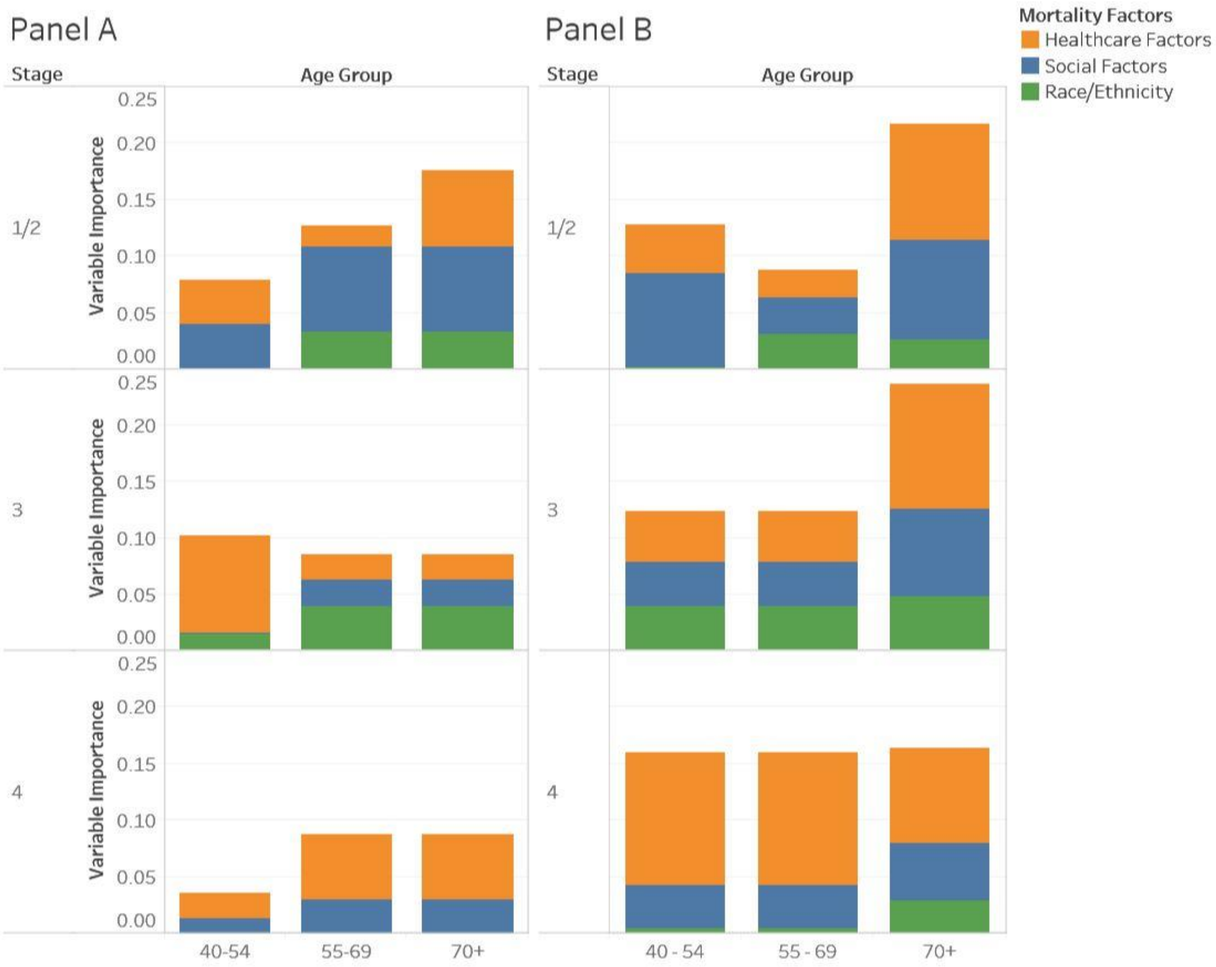


Figure 2. Summary of relative importance of Variable Importance Measures (VIMP) compared to tumor characteristic VIMPs. Panel A shows the results for the balanced subset African-American/White men with prostate cancer. Panel B shows the results for the full cohort.

Supplement

Data source

Data for this study are captured by the 18 registries comprising the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI).[16] The SEER data contain diagnostic information on all tumors diagnosed within the catchment areas of Connecticut, New Mexico, Utah, Hawaii, Iowa, New Jersey, Kentucky, Louisiana, Georgia, California, Atlanta, Detroit, San Francisco-Oakland, San Jose-Monterey, Los Angeles, Seattle-Puget Sound, and among Arizona Native Americans, and Alaska natives. The 18 SEER registries, which cover approximately 28% of the US population and link to the National Center for Health Statistics.[17] This study includes all males over age 40 (at time of diagnosis) with a PCa diagnosis from 2004–2012 (N = 514,878).

County-level health care data were obtained from the Department of Health and Human Services Area Health Resource File (AHRF), a database that contains county-specific health care and economic measures (http://ahrf.hrsa.gov). These data include the number of physicians by subspecialty within a county obtained from the 2005 American Medical Association Physician Master files and the number of hospitals from the American Health Association Hospital Facilities Database. Males were linked to county-level data in AHRF using their county of residence at time of diagnosis.

Measures

Individual-level variables were derived using the SEER data. Time was measured as months from PCa diagnosis to PCa death and observations were treated as censored at time of death from all other causes or end of the follow-up period. Race/ethnicity was grouped into the following categories; African-American, non-Hispanic white, Hispanic, other race, or unknown race/ethnicity. Staging was coded using the SEER Derived American Joint Committee on Cancer 6th edition clinical stage from 2004 to the present. Stage is based on information collected under the Collaborative Stage Data Collection

System (CS) and coded using the CS algorithm. AJCC Stage I diagnoses were relatively rare, accounting for 0.1% of all diagnoses, and were combined with Stage II diagnoses. Individuals with unknown stage at diagnosis were excluded from the analysis, as the factors leading to an unknown stage were heterogeneous and would yield no prognostic factors to affect clinical decision making (Figure 1). Tumor grade, established based on the SEER histologic grading system was measured using the following categories; <= 6, 7 (combination of 3,4), 7 (combination of 4,3), 8, 9 & 10, and unknown. Prostate screening antigen at time of diagnosis was binned into the categories as well based on the American Urologic Association criteria (<4, 4 – 9.9, 10 – 19.9, and 20+).

Selected county-level measures from the AHRF database included number of physicians and surgeons, radiation oncologists, urologists, and short-term general hospitals per 100,000 population. We expected a non-linear relationship between physician density and PCa mortality; therefore, we compared counties in the 25th and 75th percentiles to the middle 50% of the distribution. Based on findings from other studies [14, 15], we compared counties with one or more urologists to those with none. The AHRF also includes a code for Metropolitan/Micropolitan Statistical Areas, with the following categories; 1) rural, 2) Metropolitan Statistical Areas having at least one urbanized area of 50,000 or more population, and 3) Micropolitan Statistical Areas having at least one urban cluster with a population of 10,000–50,000.

The Agency for Heathcare and Research Quality (AHRQ) Prevention Quality Indicator (PQI) was used to measure the quality of care for “ambulatory care sensitive conditions”. This set of measures can be used with hospital inpatient discharge data to identify quality of care for "ambulatory care sensitive conditions." The PQIs are population based and adjusted for covariates. Even though these indicators are based on hospital inpatient data, they provide insight into the community health care system or services outside the hospital setting. For example, patients with diabetes may be hospitalized for diabetic complications if their conditions are not adequately monitored or if they do not receive the patient education needed for appropriate self-management.

County-level demographics from the 2000 U.S. Census were also included in the models. Median–family-income (MFI) and the percent of families below the poverty line were also drawn from the Census data. As this population is largely over age 65, we considered the proportion of the Medicare population in each county that is also eligible for Medicaid enrollment ("dual enrollees").[18] The Center for Disease Control (CDC) Social Vulnerability Index (SVI) was included in the models. The SVI uses 15 variables from the 2000 US Census data to measure four domains of social vulnerability; 1) Socioeconomic status (income, poverty, employment, and education), 2) Household composition and disability (age, single parents, and disability), 3) Minority status and language profile (race, ethnicity, and English language), and 4) Housing and transportation profile (housing structure, crowding, and vehicle access). Counties are then given an overall vulnerability rank.

Statistical Methods

Classification and Regression Trees

Traditional regression approaches specify interactions to test for differences in the effect of individual, family history, and neighborhood characteristics on prostate cancer mortality. However, there are some limitations to using interactions to bin individuals into groups with similar prognostic outcomes. All interactions need to be specified a priori and the number of interactions and interactions involving multiple variables are difficult to interpret. A classification and regression tree (CART) is an alternative method that allows us to explore the structure of the data with the goal of predicting survival outcomes based on individual and county level characteristics that may affect prostate cancer mortality.

A regression tree is a hierarchal structure that has a top node (or root) and observations are passed down the tree. Each decision point, which is selected by the algorithm to explain the most deviance, is labeled a node (sometimes called daughter nodes) until it reaches a terminal node (or leaf). CART uses a binary splitting process to identify the best model for classifying individuals into distinct groups. The central aim of the regression tree approach is to form meaningful classes that are determined by the data (not a priori assumptions). The results from CART likely do not represent additive functions that consist only of main effects, but complex interactions between variables in the data. Essentially, we are asking how a compilation of variables come together to define distinct sub-classes of individuals.

Random Forest

Random forests (RF) are an extension of CART. In this method, *n* trees are grown using a bootstrapped sample from the learning sample.[2, 3] The number of trees grown is specified by the user, with a default of 1000 in the R statistical package *randomForestSRC* with a logrank splitting rule. We chose to have the algorithm fit 200 trees and constrained the model to have a terminal node size of 50. Unlike CART, there is no trimming or stopping criterion, the trees are fully grown (the user can modify this criterion, however the standard practice is to fully grow the tree). Additionally, a subset of variables are randomly selected for inclusion at each node. This method of random subspace selection is done to avoid correlation between trees in the forest and decreases the error. It also allows for the selection of the most relevant variables when there are multicollinearity issues and therefore reduces the variables of interest to those with the most explanatory value.[4] All models had error rates that ranged between 17% and 29%, with the highest stage of disease models having the highest error rates.

One benefit of the RF method is the ability to quantify the variable importance. We used the Breiman-Cutler measure of importance (or permutation) measure, the most frequently used measure for random forests. Since each tree is a random subset of the original dataset, the remaining 30% of the data not selected (e.g., out of bag observations (oob)) can be used to calculate the variable importance. The oob data is used to create permutation accuracy variable importance measure (VIMP) by predicting class membership in the oob sample and then permuting the variables and calculating the predictive accuracy with permuted variables. The average difference in accuracy of the oob versus permuted oob observations over the trees is the VIMP, with a VIMP close to zero implying that the variable has no predictive power. Correlation between variables was assessed to assure that none of the variables were highly (r>0.75) predictive of the variables in the model.

Cox Proportional Hazard Regression

Cox regression with all variables were also run for each age and stage model. These models only included main effects and, while the interpretation between RF and Cox PH models differs, were used to assess the benefit of using an RF approach. We found that there were substantive differences in interpretation between the RF and Cox PH mod

Supplementary Table 1. Variable Importance Measures for Age 50 – 54 at Time of Diagnosis

Stage 1/2		
Variable	VIMP	Category
State	0.0064	Macro
Diagnosis Year	0.0053	Macro
Gleason Score	0.1099	Tumor
PSA	0.0575	Tumor
Median Family Income	0.0109	Social
Number of Radiation/Oncologists	0.0032	Access to health
Rural/Urban	0.0019	Social
Number of Chemotherapy Treatment Centers	0.0018	Access to Health
Number of Doctors	0.0013	Access to Health
Social Vulnerability	0.0008	Social
Number of Urologists	0.0008	Access to Health
Medicare/Medicaid Dual Enrollment	0.0004	Social
Race/Ethnicity	0.0002	Race/Ethnicity
Stage 3		
Diagnosis Year	0.0021	Macro
Gleason Score	0.1335	Tumor
PSA	0.0073	Tumor
Race/Ethnicity	0.0055	Race/Ethnicity
Social Vulnerability	0.0038	Social
Number of Doctors	0.0037	Access to Health
Number of Chemotherapy Treatment Centers	0.0018	Access to Health
Median Family Income	0.0018	Social
Number of Radiation/Oncologists	0.0005	Access to health
Prevention Quality Index	0.0004	Access to Health
Stage 4		
Gleason Score	0.0886	Tumor
PSA	0.0582	Tumor
Number of Radiation/Oncologists	0.0053	Access to health
Number of Doctors	0.0045	Access to health
Median Family Income	0.0038	Social
Number of Chemotherapy Treatment Centers	0.0027	Access to health
Prevention Quality Index - African American	0.0018	Access to health
Prevention Quality Index	0.0017	Access to health
Rural/Urban	0.0013	Social
Number of Urologists	0.0013	Access to health
Race/Ethnicity	0.0007	Race/Ethnicity
Social Vulnerability	0.0004	Social

Supplementary Table 2. Variable Importance Measures for Age 50 – 54 at Time of Diagnosis: African American Subsample

Stage 1/2		
Label	VIMP	Category
State	0.0002	Macro
Gleason Score	0.1034	Tumor
PSA	0.0496	Tumor
Median Family Income	0.0045	Social
Number of Doctors	0.0029	Access to Health
Number of Radiation/Oncologists	0.0023	Access to Health
Medicare/Medicaid Dual Enrollment	0.0009	Social
Prevention Quality Index - African American	0.0008	Access to Health
Social Vulnerability	0.0007	Social
Stage 3		
State	0.0156	Macro
Diagnosis Year	0.0021	Macro
Gleason Score	0.1162	Tumor
PSA	0.0248	Tumor
Number of Radiation/Oncologists	0.0075	Access to Health
Race/Ethnicity	0.0022	Race/Ethnicity
Number of Chemotherapy Treatment Centers	0.0019	Access to Health
Prevention Quality Index - African American	0.0015	Access to Health
Prevention Quality Index	0.0013	Access to Health
Stage 4		
GleasonCat	0.0928	Tumor
PSA	0.0537	Tumor
Rural/Urban	0.0013	Social
Prevention Quality Index	0.0013	Access to Health
Number of Chemotherapy Treatment Centers	0.001	Access to Health
Number of Urologists	0.0008	Access to Health
Median Family Income	0.0007	Social

Supplementary Table 3. Variable Importance Measures for Age 55 – 69 at Time of Diagnosis

Stage 1/2		
Variable	VIMP	Category
Diagnosis Year	0.0074	Macro
State	0.0039	Macro
Gleason Score	0.0846	Tumor
PSA	0.0466	Tumor
Race/Ethnicity	0.004	Race/Ethnicity
Medicare/Medicaid Dual Enrollment	0.0012	Social
Rural/Urban	0.0012	Social
Prevention Quality Index	0.001	Access to health
Social Vulnerability	0.001	Social
Median Family Income	0.0009	Social
Number of Urologists	0.0006	Access to health
Number of Radiation/Oncologists	0.0005	Access to health
Number of Chemotherapy Treatment Centers	0.0005	Access to health
Number of Doctors	0.0005	Access to health
Prevention Quality Index - African American	0.0001	Access to health
Stage 3		
Diagnosis Year	0.0066	Macro
State	0.0057	Macro
Gleason Score	0.1153	Tumor
PSA	0.0227	Tumor
Median Family Income	0.0033	Social
Number of Doctors	0.002	Access to health
Medicare/Medicaid Dual Enrollment	0.0016	Social
Social Vulnerability	0.0016	Social
Number of Radiation/Oncologists	0.0013	Access to health
Prevention Quality Index - African American	0.0011	Access to health
Rural/Urban	0.0006	Social
Prevention Quality Index	0.0003	Access to health
Stage 4		
Diagnosis Year	0.0062	Macro
State	0.0046	Macro
Gleason Score	0.0996	Tumor
PSA	0.0577	Tumor
Median Family Income	0.0051	Social
Number of Doctors	0.0031	Access to health

Race/Ethnicity	0.0026	Race/Ethnicity
Number of Radiation/Oncologists	0.0022	Access to health
Rural/Urban	0.0012	Social
Social Vulnerability	0.0011	Social
Number of Chemotherapy Treatment Centers	0.001	Access to health
Number of Urologists	0.0008	Access to health
Prevention Quality Index - African American	0.0006	Access to health
Prevention Quality Index	0.0004	Access to health
Medicare/Medicaid Dual Enrollment	0.0004	Social

Supplementary Table 4. Variable Importance Measures for Age 55 – 69 at Time of Diagnosis: African American Subsample

Stage 1/2		
Label	VIMP	Category
Diagnosis Year	0.0045	Macro
State	0.004	Macro
Gleason Score	0.0875	Tumor
PSA	0.0639	Tumor
Median Family Income	0.0055	Social
Race/Ethnicity	0.005	Race/Ethnicity
Social Vulnerability	0.004	Social
Prevention Quality Index	0.0038	Access to Health
Prevention Quality Index - African American	0.0029	Access to Health
Number of Radiation/Oncologists	0.0019	Access to Health
Rural/Urban	0.0015	Social
Number of Urologists	0.0009	Access to Health
Number of Doctors	0.0007	Access to Health
Medicare/Medicaid Dual Enrollment	0.0004	Social
Number of Chemotherapy Treatment Centers	0.0001	Access to Health
Stage 3		
Diagnosis Year	0.0141	Macro
State	0.0117	Macro
Gleason Score	0.0749	Tumor
PSA	0.0453	Tumor
Race/Ethnicity	0.0047	Race/Ethnicity
Number of Doctors	0.0015	Access to Health
Median Family Income	0.0013	Social
Social Vulnerability	0.0008	Social
Medicare/Medicaid Dual Enrollment	0.0008	Social
Number of Chemotherapy Treatment Centers	0.0005	Access to Health
Prevention Quality Index - African American	0.0004	Access to Health
Prevention Quality Index	0.0002	Access to Health
Number of Radiation/Oncologists	0.0001	Access to Health
Stage 4		
Diagnosis Year	0.0004	Macro
State	0.0063	Macro
Gleason Score	0.0816	Tumor
PSA	0.0537	Tumor
Number of Doctors	0.003	Access to Health
Median Family Income	0.0025	Social
Number of Radiation/Oncologists	0.0021	Access to Health
Prevention Quality Index	0.0018	Access to Health
Rural/Urban	0.0008	Social

Social Vulnerability	0.0007	Social
Number of Chemotherapy Treatment Centers	0.0005	Access to Health
Prevention Quality Index - African American	0.0003	Access to Health
Number of Urologists	0.0001	Access to Health

Supplementary Table 5. Variable Importance Measures for Age 70+ at Time of Diagnosis

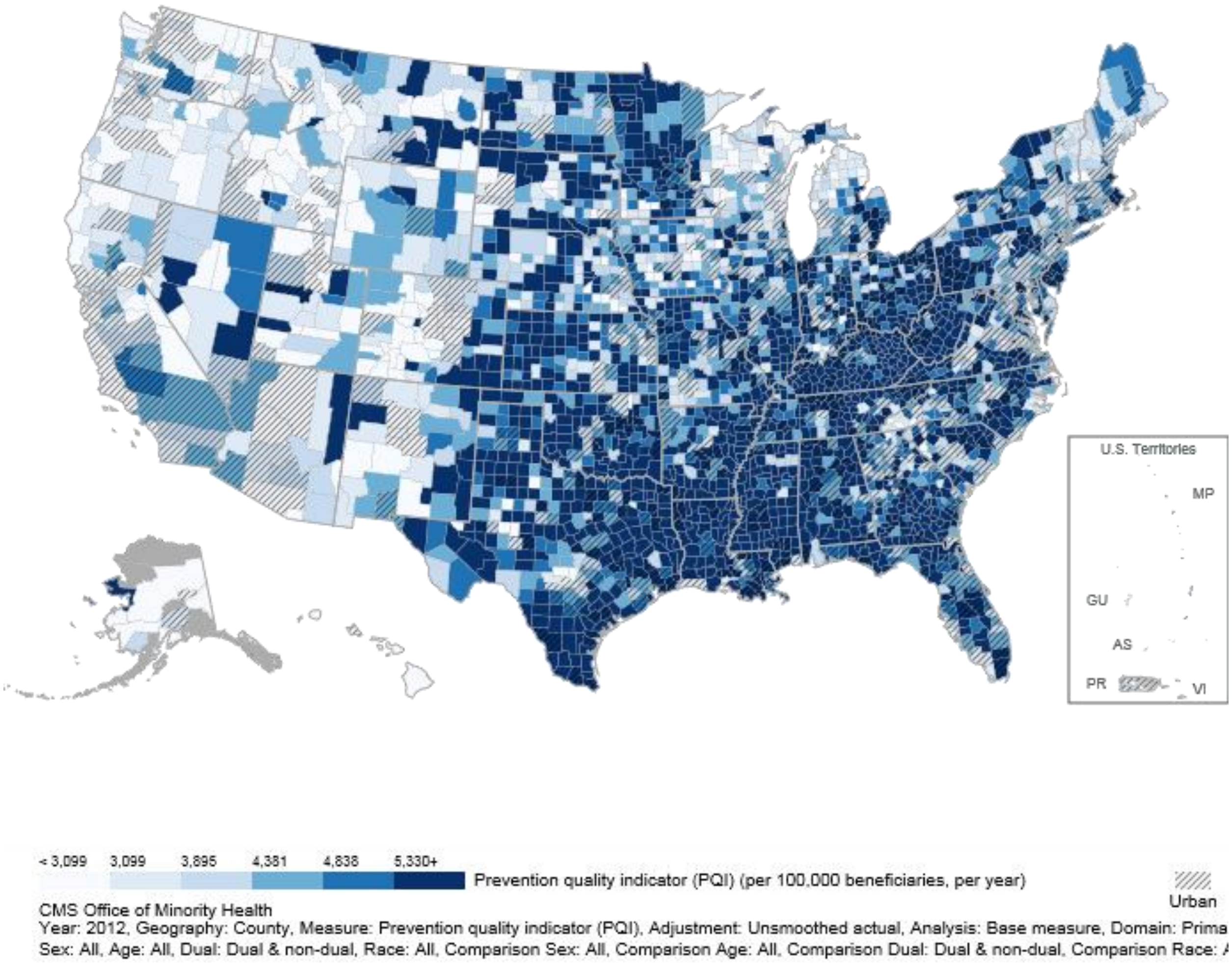
Stage 1/2		
Variable	VIMP	Category
State	0.009	Macro
Diagnosis Year	0.0056	Macro
Gleason Score	0.1106	Tumor
PSA	0.0554	Tumor
Median Family Income	0.0097	Social
Number of Doctors	0.007	Access to Health
Number of Radiation/Oncologists	0.0043	Access to Health
Race/Ethnicity	0.0043	Race/Ethnicity
Social Vulnerability	0.0022	Social
Number of Urologists	0.0021	Access to Health
Number of Chemotherapy Treatment Centers	0.0019	Access to Health
Rural/Urban	0.0019	Social
Prevention Quality Index - African American	0.0011	Access to Health
Medicare/Medicaid Dual Enrollment	0.0009	Social
Prevention Quality Index	0.0007	Access to Health
Stage 3		
State	0.0084	Macro
Diagnosis Year	0.0038	Macro
Gleason Score	0.0947	Tumor
PSA	0.0301	Tumor
Median Family Income	0.007	Social
Race/Ethnicity	0.0059	Race/Ethnicity
Number of Doctors	0.0049	Access to Health
Prevention Quality Index - African American	0.0023	Access to Health
Prevention Quality Index	0.0022	Access to Health
Number of Urologists	0.002	Access to Health
Number of Chemotherapy Treatment Centers	0.0018	Access to Health
Rural/Urban	0.0013	Social
Social Vulnerability	0.0012	Social
Number of Radiation/Oncologists	0.0007	Access to Health
Medicare/Medicaid Dual Enrollment	0.0003	Social
Stage 4		
State	0.0068	Macro
Diagnosis Year	0.0031	Macro
GleasonCat	0.086	Tumor
PSA	0.0209	Tumor

Race/Ethnicity	0.0031	Race/Ethnicity
Number of Doctors	0.0027	Access to Health
Median Family Income	0.0027	Social
Number of Radiation/Oncologists	0.0023	Access to Health
Number of Chemotherapy Treatment Centers	0.0012	Access to Health
Prevention Quality Index - African American	0.0012	Access to Health
Social Vulnerability	0.0012	Social
Rural/Urban	0.0011	Social
Number of Urologists	0.0009	Access to Health
Prevention Quality Index	0.0007	Access to Health
Medicare/Medicaid Dual Enrollment	0.0004	Social

Supplementary Table 6. Variable Importance Measures for Age 70+ at Time of Diagnosis: African American Subsample

Stage 1/2		
Label	VIMP	Category
Diagnosis Year	0.0067	Macro
State	0.0059	Macro
Gleason Score	0.0945	Tumor
PSA	0.0715	Tumor
Median Family Income	0.0129	Social
Number of Doctors	0.0059	Access to health
Number of Radiation/Oncologists	0.0052	Access to health
Race	0.0037	Race/Ethnicity
Number of Urologists	0.0029	Access to health
Number of Chemotherapy Treatment Centers	0.0023	Access to health
Prevention Quality Index	0.0021	Access to health
Social Vulnerability	0.0021	Social
Prevention Quality Index - African American	0.0019	Access to health
Rural/Urban	0.0015	Social
Medicare/Medicaid Dual Enrollment	0.0002	Social
Stage 3		
State	0.0106	Macro
Gleason Score	0.1074	Tumor
PSA	0.0366	Tumor
Median Family Income	0.0044	Social
Prevention Quality Index - African American	0.0011	Access to health
Rural/Urban	0.001	Social
Medicare/Medicaid Dual Enrollment	0.0004	Social
Stage 4		
Diagnosis Year	0.0006	Macro
Gleason Score	0.0891	Tumor
PSA	0.0174	Tumor
Median Family Income	0.002	Social
Number of Urologists	0.0011	Access to health
Number of Radiation/Oncologists	0.0007	Access to health
Rural/Urban	0.0005	Social
Number of Doctors	0.0003	Access to health
Race	0.0002	Race/Ethnicity
Number of Chemotherapy Treatment Centers	0.0001	Access to health

Supplementary Figure 1. Prevention Quality Index by County 2012. *Source CMS Office of Minority Health https://data.cms.gov/mapping-medicare-disparities*.


U.S. Territories
MP
GU
AS
PR
VI
< 3,099
3,099
3,895
4,381
4,838
5,330+
Prevention quality indicator (PQI) (per 100,000 beneficiaries, per year)
Urban

CMS Office of Minority Health
Year: 2012, Geography: County, Measure: Prevention quality indicator (PQI), Adjustment: Unsmoothed actual, Analysis: Base measure, Domain: Prima
Sex: All, Age: All, Dual: Dual & non-dual, Race: All, Comparison Sex: All, Comparison Age: All, Comparison Dual: Dual & non-dual, Comparison Race: A