PLOS ONE

RESEARCH ARTICLE

A Neighborhood-Wide Association Study (NWAS): Example of prostate cancer aggressiveness

Shannon M. Lynch[1]*, Nandita Mitra[2], Michelle Ross[2], Craig Newcomb[2], Karl Dailey[2], Tara Jackson[2], Charnita M. Zeigler-Johnson[3], Harold Riethman[4], Charles C. Branas[2,5], Timothy R. Rebbeck[6]

1 Fox Chase Cancer Center, Cancer Prevention and Control, Philadelphia, Pennsylvania, United States of America, **2** University of Pennsylvania, Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America, **3** Thomas Jefferson University, Philadelphia, Pennsylvania, United States of America, **4** Old Dominion University, Norfolk, Virginia, United States of America, **5** Columbia University, Mailman School of Public Health, New York, New York, United States of America, **6** Dana Farber Cancer Institute and Harvard TH Chan School of Public Health, Boston, Massachusetts, United States of America

* shannon.lynch@fccc.edu

Citation: Lynch SM, Mitra N, Ross M, Newcomb C, Dailey K, Jackson T, et al. (2017) A Neighborhood-Wide Association Study (NWAS): Example of prostate cancer aggressiveness. PLoS ONE 12(3): e0174548. https://doi.org/10.1371/journal.pone.0174548

Editor: Esmaeil Ebrahimie, Flinders University, AUSTRALIA

Received: June 14, 2016

Accepted: March 11, 2017

Published: March 27, 2017

Data Availability Statement: Data cannot be made publicly available for ethical and legal reasons. This data includes geocoded data at the census tract level linked to individual, anonymized records and releasing PA registry data is against the data use agreement. The U.S. Census data for this analysis can be downloaded from socialexplorer.com. Researchers can request Pennsylvania State Cancer Registry Data from the Pennsylvania Department of Health. Current contact information: James Rubertone Department of Health | Bureau of Health Statistics and Registries 555 Walnut Street,

Abstract

Purpose

Cancer results from complex interactions of multiple variables at the biologic, individual, and social levels. Compared to other levels, social effects that occur geospatially in neighborhoods are not as well-studied, and empiric methods to assess these effects are limited. We propose a novel Neighborhood-Wide Association Study(NWAS), analogous to genome-wide association studies(GWAS), that utilizes high-dimensional computing approaches from biology to comprehensively and empirically identify neighborhood factors associated with disease.

Methods

Pennsylvania Cancer Registry data were linked to U.S. Census data. In a successively more stringent multiphase approach, we evaluated the association between neighborhood (n = 14,663 census variables) and prostate cancer aggressiveness(PCA) with n = 6,416 aggressive (Stage≥3/Gleason grade≥7 cases) vs. n = 70,670 non-aggressive (Stage<3/Gleason grade<7) cases in White men. Analyses accounted for age, year of diagnosis, spatial correlation, and multiple-testing. We used generalized estimating equations in Phase 1 and Bayesian mixed effects models in Phase 2 to calculate odds ratios(OR) and confidence/credible intervals(CI). In Phase 3, principal components analysis grouped correlated variables.

Results

We identified 17 new neighborhood variables associated with PCA. These variables represented income, housing, employment, immigration, access to care, and social support. The

6th Floor | Harrisburg, PA 17101-1914 Phone: 717.547.3690 | Fax: 717.772.3258. | jrubertone@pa.gov

Funding: This work was supported by grants from the Public Health Service (P50-CA105641, P60-NM006900 and R01-CA85074 to TRR and F31-AG039986 to SML). The authors have no competing financial interests to report.

Competing interests: The authors have declared no competing interests exist.

top hits or most significant variables related to transportation (OR = 1.05;CI = 1.001–1.09) and poverty (OR = 1.07;CI = 1.01–1.12).

Conclusions

This study introduces the application of high-dimensional, computational methods to large-scale, publically-available geospatial data. Although NWAS requires further testing, it is hypothesis-generating and addresses gaps in geospatial analysis related to empiric assessment. Further, NWAS could have broad implications for many diseases and future precision medicine studies focused on multilevel risk factors of disease.

Introduction

Cancer likely results from complex interactions of factors at the macro-environmental, individual, and biologic levels[1]. Identifying relevant factors within each level for joint studies is a challenge, particularly at the macro-environmental level, defined here by the neighborhood in which a person lives. Studies of cancer that evaluate neighborhood consider a limited number of variables based on prior knowledge; this affects comparability and consistency across studies and makes etiologic inferences difficult. A lack of empirical assessment is a well-cited limitation of neighborhood studies[2], and empiric approaches from biology could be applied to the macro-environmental level. For example, genome-wide association studies (GWAS) have driven population-based cancer research for the past several years [3]. GWAS use high-throughput, low-cost technology and readily available genome-mapping to evaluate the role of millions of genetic markers for a variety of diseases using an agnostic approach[4]. These approaches are hypothesis-generating, and the clinical implications of GWAS are starting to have translational impact[5].

Applying concepts from GWAS, environmental-wide association studies (EWAS) were subsequently developed to study the effect of exposures at the individual level (e.g., pesticides), and to provide insights for gene-environment interaction studies[6]. However, neighborhood factors have not been comprehensively studied using these approaches. Borrowing concepts from GWAS and EWAS, we propose the neighborhood-wide association study (NWAS) as a novel, empirical approach to evaluate the effect of multiple neighborhood-level exposures on disease outcomes and to address gaps in neighborhood research. The objective of this method is to apply informatics approaches to the study of neighborhood through the systematic identification of neighborhood factors that may be associated with disease phenotypes. With NWAS, we aim to generate hypotheses in order to inform gene-environment studies and potentially more precisely identify neighborhoods at high risk for poor cancer outcomes.

We introduce the NWAS approach and demonstrate how agnostic, high-dimensional data analyses can be used to identify neighborhood characteristics associated with high grade/high stage, aggressive prostate cancer. There are at least two hypotheses that may explain the role of neighborhood in prostate cancer aggressiveness. First, unfavorable neighborhood environments may exert a biological effect on prostate cancer aggressiveness. Neighborhood environment could affect prostate cancer severity under a chronic stress hypothesis, in which residents from disadvantaged neighborhoods experience greater emotional stress and constant "wear and tear" on the body that can affect cancer initiation and progression[7] [8, 9]. Second, unfavorable neighborhood environments may be correlated with factors related to health care access, particularly screening behaviors and practices. Because screening can detect cancer at

earlier stages, people living in less favorable neighborhoods may have less access to care that lead to later (i.e., more aggressive) cancers at the time of diagnosis[10–13]. These two hypotheses are not mutually exclusive of one another and could both be acting through neighborhood-level influences. Given few individual-level risk factors for prostate cancer have been identified [14] and only a few studies have investigated neighborhood effects on prostate cancer using *a priori* variable selection approaches[10–13], empiric assessments of the effect of neighborhood on aggressive prostate cancer are warranted.

Materials and methods

Study population

Anonymized data from the Pennsylvania (PA) Department of Health Cancer Registry identified prostate cancer patients diagnosed from 1995 to 2005. The registry included variables related to prostate cancer tumor stage and grade, age at diagnosis, year of diagnosis, and race/ethnicity. We focused only on Caucasian prostate cancer cases in this analysis (n = 80,575). Race-specific analyses were also conducted in GWAS to account for population stratification [3, 15]. We excluded cases with a P.O. Box address (n = 112). We also excluded those missing tumor grade or stage (n = 3371), age (n = 2), or year of diagnosis (n = 4). A total of 77,086 men were included in the analysis (S1 File).

Neighborhood variables

Residential addresses of prostate cancer patients were geocoded at the census tract level and assigned a Federal Information Processing Standard (FIPS) code[16] using Arc GIS software. The FIPS code was linked to the 2000 U.S. Census using Microsoft Visual Studio 2008. Prostate cancer cases were linked to the neighborhood variable values of the census tract in which they live, and cases residing in the same census tract were assumed to have the same neighborhood characteristics.

All 24,634 census tract variables available in the 2000 U.S. Census Summary File 1 (SF1) and Summary File 3 (SF3) were downloaded from Social Explorer (http://www.socialexplorer.com). The SF1 form is distributed to every household in the U.S. SF1 collects demographic data about each person within the housing unit, such as age, gender, and race, as well as general housing information related to occupancy and tenure. The SF3 form is distributed to 5% of all housing units in the U.S and includes more specific questions related to socioeconomic status and physical environment characteristics, such as migration, language ability, disability, veterans status, vehicle availability, kitchen and plumbing facilities [17, 18]. All SF1 and SF3 variables were evaluated for missing data (S1 File; S1 and S2 Figs; S2 and S3 Files). Variables with greater than 10% missingness (n = 8,092) and modal values that comprised over 95% of the data (n = 1,879) were excluded. 14,663 census variables were left for analysis.

Outcome definition

All incident, White prostate cancer cases among men residing in Pennsylvania from 1995–2005 were included in this study. Incident cases were identified according to ICD-0-3 site and morphology coding. We assumed complete case ascertainment, given medical facilities are required by law to report all diagnosed prostate cancer cases[19] We created a combined "prostate cancer aggressiveness" variable for our primary outcome that was defined by 6,416 cases with a high tumor stage(stage 3 or 4) and high tumor grade(grade 7+), compared to 70,670 controls(<Stage 3 or <Gleason 7)[20, 21](S1 File).

Statistical analysis

The NWAS consists of methodologic steps derived from GWAS and EWAS[22] [6, 23]. First, we consider all publically-available Year 2000 U.S. Census variables, which serve as neighborhood "loci", measured across cases and controls, for associations with prostate cancer aggressiveness after adjustments for multiple comparisons[6]. Second, we account for spatial effects which assume that nearby neighborhoods have similar characteristics[23], an effect that is not consistently accounted for in neighborhood studies and is considered a limitation[2]. Third, we account for linkage disequilibrium statistically and consider the high degree of correlation among census variables [22]. Fourth, dimension reduction techniques were applied across 3 analytical phases, where each phase included progressively more stringent statistical criteria to minimize false positives. All models were adjusted for age at diagnosis and year of diagnosis.

Phase 1. The goal of phase 1 of the NWAS analysis was to identify an initial set of neighborhood-level variables associated with unfavorable prostate cancer prognosis, accounting for non-independence of observations within neighborhoods. For each neighborhood variable, a Generalized Estimating Equation (GEE) approach with a logit link function, robust standard error, and assumption of an exchangeable correlation matrix was used to estimate an odds ratio (OR) and 95% confidence interval[24]. P-values were Bonferroni-corrected to account for multiple comparisons[25]; corrected p-values less than 0.05 were considered to be statistically significant. Phase 1 analyses were conducted using SAS 12.0 statistical software.

Phase 2. The purpose of the statistical methods in Phase 2 was to further evaluate those variables that reached statistical significance in Phase 1 by accounting for spatial variability in our data. To accomplish this, we specified a Bayesian hierarchical logistic regression model in which we allow for both global and local smoothing using two sets of random effects (see S1 File).

We defined the geographic region as county since each geographic area must include at least 1 case and 1 control. Neighborhood variables were Z-score transformed in order to compare odds ratios from many regressions[6]. To address the large multiple testing problem, our significance threshold is set to 0.05/n for n the number of variables identified in Phase 1, which corresponds to a Bonferroni corrected threshold of 0.05. Significance is determined by the exclusion of 0 in the (100–0.05/n)% credible intervals (CI). Phase 2 analyses were conducted using Integrated Nested Laplace Approximations (INLA) [26] in R statistical software version 3.1.3.

Phase 3. The goal of the statistical methods in Phase 3 was to identify variable groups that reflect correlated neighborhood concepts. Paralleling the idea of haplotype blocks (SNPS in high linkage disequilibrium) in GWAS that represent similar gene regions[6], we clustered together significant variables from Phase 2 using principal components analysis[27]. Like post-GWAS fine mapping approaches[28], the most significant variable within each component (i.e., the variable with the tightest credible interval from Phase 2) was considered the best representation of that principal component (i.e. gene region). The most significant variables, or "top hits" were selected from each individual principal component that together explained 90% of the variance. Thus, there were as many top hits as principal components. Phase 3 analyses used STATA/SE 12.0 statistical software. This study utilized existing data sources which allow for waivers of informed consent and was approved by the Institutional Review Board of the University of Pennsylvania under protocol 817734.

Results

Of the reported 3,135 census tracts in PA in 2000, 3,037 (97%) census tracts are represented in our study sample (S1 File). Aggressive prostate cancer cases were clustered in urban areas,

namely Pittsburgh and Philadelphia (S1 File). The average age of the study population was 69.2 (standard deviation (sd): 9.4) and mean year of diagnosis was 2000. The average age of aggressive cases was 69.8 (sd: 10.4) and of nonaggressive cases was 68.8 (sd: 9.0).

Fig 1 summarizes Phases 1–3 study methods (Fig 1a) and findings (Fig 1b). In Phase 1, from 14, 663 variables, we identified 434 unique census variables significantly associated with prostate cancer aggressiveness at Bonferroni-corrected significance levels (S1 and S4 Files). After Phase two, 217 unique variables were still significant at Bonferroni-corrected credible intervals (S5 File). The average amount of residual variability in aggressive prostate cancer risk across the 217 Bayesian models in Phase 2 was 34% (range: 14%-50%), which is considered large. In Phase 3, 17 uncorrelated principal components were identified from these 217 neighborhood variables. Components 1–8 explain 80% of the variance, and related to poverty (Component 1), white only neighborhood characteristics (Component 2), household/ housing unit poverty status (Component 3), households and living alone (component 4), rented houses built before 1939 (Component 5), civilian population (Component 6), household income above $60K (Component 7), and immigration (Component 8) (Fig 1b). 76 of 217 Phase 2 variables loaded on Component 1 and 51 loaded on Component 2 after the Phase 3 principal components analysis. The top 10% of significant variables from Phase 2 loaded on Component 1 (S5 File).

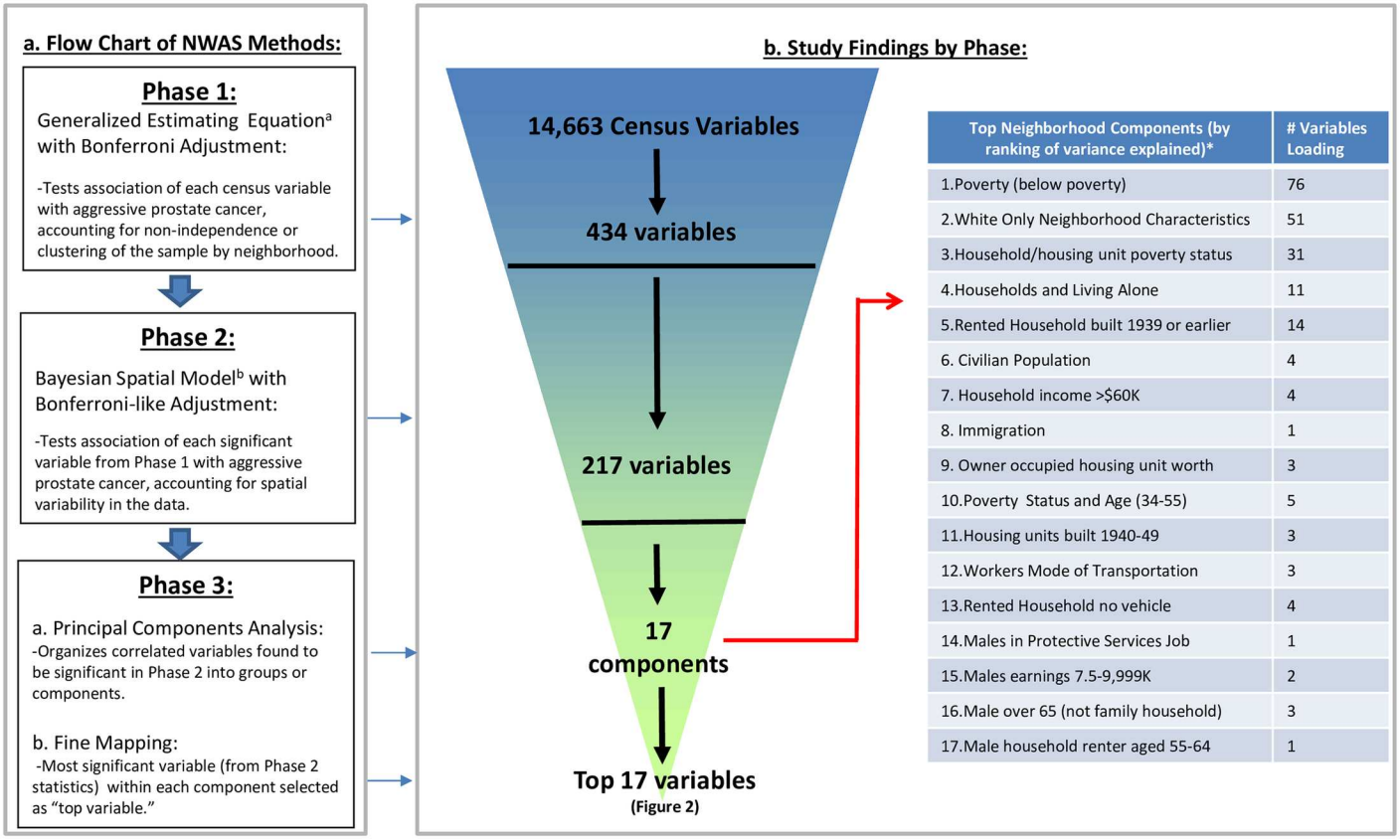

Fig 1. a. Study flow chart of NWAS statistical methods b. Overview of study findings by methodological phase.

[a] $\text{Logit}(p) = \alpha+\beta_{i0}x_{age}+\beta_{i1}x_{\text{year of diagnosis}}+\beta_{i2}x_{\text{neighborhood variable (i, j)}}+\varepsilon_{ij}$;
where i = individual cancer cases; j = census tracts (Phase 1)

[b] $\text{Logit}(p) = \alpha+\beta_{i0}x_{age}+\beta_{i1}x_{\text{year of diagnosis}}+\beta_{2}x_{\text{neighborhood variable (i, j)}}+V_{(j)}+U_{(j)}$
where i = individual prostate cancer cases; j = county, $V_{(j)}$ are independent non-spatial random effects and $U_{(j)}$ are spatially structured random effects (Phase 2).

*These 17 components explain 90% of the variance

https://doi.org/10.1371/journal.pone.0174548.g001

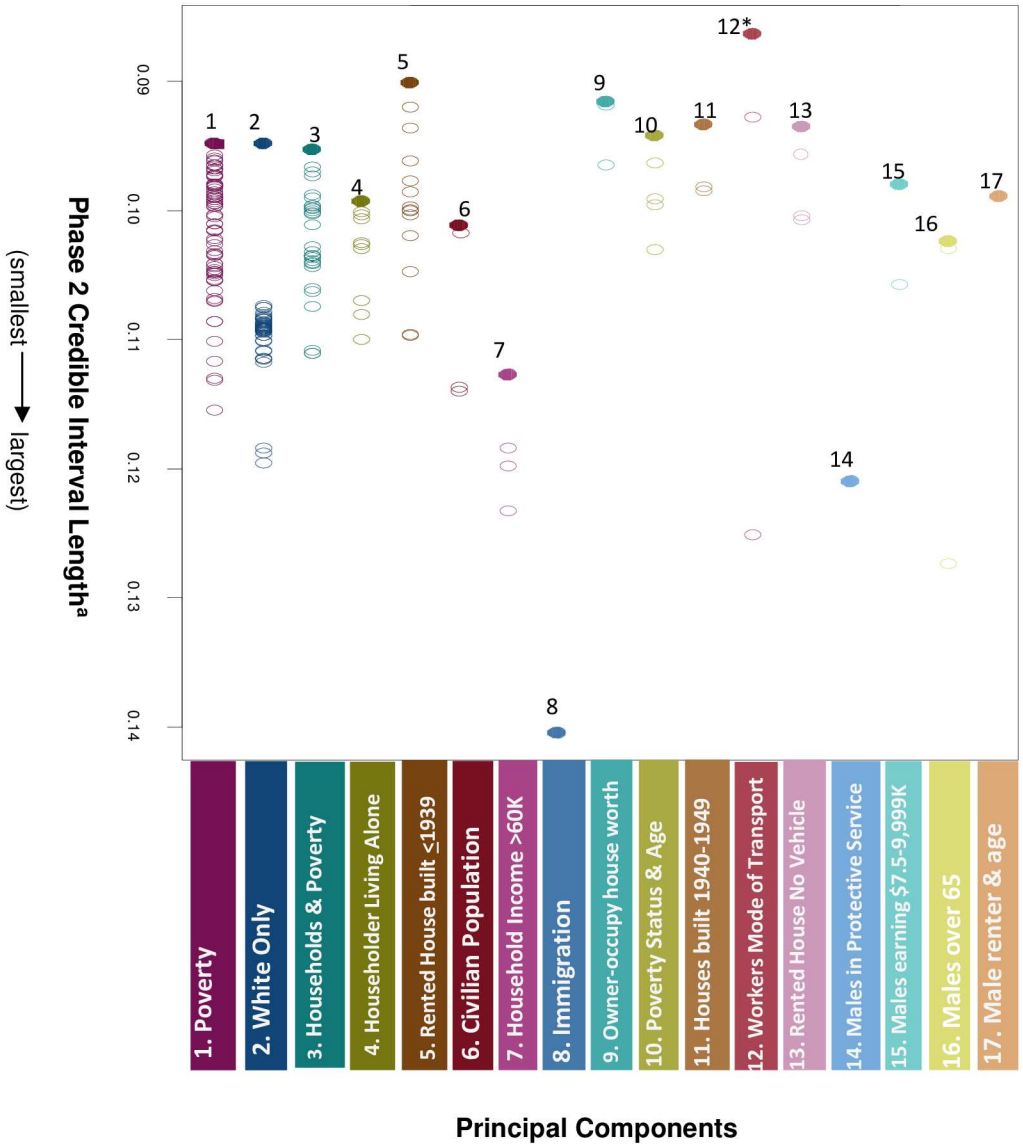

Fig 2. Phase 3-Principal components and fine mapping analysis to identify top hits. Dots represent single neighborhood variables from Phase 2 (n = 217 total dots). Open dots are color-coded to their respective component (from Phase 3-Principal Components analysis). Closed-colored dots represent the most significant variable within each component (Phase 3-Fine Mapping) and corresponding statistics are provided by component number in Fig 3. *Top hit based on statistical significance from Phase 2 data. [a] Statistical significance determined by Bonferroni-corrected confidence intervals from the Phase 2 Bayesian model, i.e. smaller credible interval length indicates greater statistical significance.

https://doi.org/10.1371/journal.pone.0174548.g002

The top 17 most significant variables within each of the 17 principal components are presented in Figs 2 and 3. The most significant variable from Component 1 represented Non-Hispanic Whites aged 6–11 for whom poverty status was determined (OR = 1.07, CI = 1.01–1.12). Seven of the top 17 hits or variables from Components 1–3, 7, 9–10, 15 were related to socioeconomic status. One top hit related to employment, specifically male protective service occupations such as fire-fighting and law enforcement (OR = 0.94, CI = 0.89–0.99), and one related to immigration status (OR = 0.93, CI = 0.87–0.99). Two were associated with physical environment (aggregate income of occupied, rented housing units built 1940–1949 with a householder

			Phase 1				Phase 2			Phase 3
	Mean (sd[a])	Range	Odds Ratio	CI[b]	p-value	Rank	Odds Ratio	CI[b]	Rank	Component Loading
Census Variable										
%White alone population for whom poverty status is determined age 6-11 years (pct_sf3_pct075a006)	0.38 (0.53)	0-20.6	1.09	1.05-1.12	0.03	207	1.07	1.01-1.12	12	1
%White, Non-hispanics where poverty status determined aged 18-64 below poverty level in 1999 (pct_SF3_p159i007)	4.4 (4.7)	0-100	1.01	1.01-1.02	.003	112	1.06	1.02-1.12	11	2
% Male Nonfamily households below poverty level (pct_sf3_p092021)	1.6 (1.9)	0-35.4	1.03	1.02-1.04	0.03	382	1.06	1.01-1.11	14	3
%Male householder living alone (nonfamily household) (pct_SF3_h019093)	4.69 (3.9)	0-39.5	1.02	1.01-1.02	0.03	386	1.07	1.01-1.12	61	4
% Renter occupied housing unit built 1939 or earlier with householder aged 15-24 years (pct_SF3_hct005083)	0.75 (1.4)	0-31.8	1.05	1.03-1.06	0.003	101	1.07	1.02-1.11	2	5
Imputed civilian non-institutionalized population 5 years and older (pct_sf3_p120002)	6.39 (2.6)	0-63.2	1.02	1.01-1.03	0.02	340	1.06	1.01-1.11	91	6
%Household income $60K-74,999 (pct_SF3_p052012)	10.9 (3.6)	0-25.1	0.98	0.97-0.99	0.048	433	0.95	0.89-0.99	202	7
%Foreign born naturalized citizen at or above poverty level (pct_SF3_pct051020)	2.0 (2.1)	0-18.9	0.96	0.94-0.97	8.8 X 10-6	8	0.93	0.87-0.99	217	8
%Household income of $10K-19,999 with owner-occupied housing unit value of $10K-19,999 (pct_SF3_hct017019)	0.34 (1.1)	0-23.0	1.06	1.04-1.08	2.7 X 10-5	13	1.05	1.00-1.10	3	9
% where the ratio of income to poverty level in 1999 for persons aged 45-54 years, under 0.50 (pct_sf3_pct050102)	0.33 (0.41)	0-12.6	1.18	1.12-1.25	6X10-5	17	1.08	1.03-1.13	10	10
Aggregate household income in 1999 of renter occupied housing units with householders 15-34 that were built 1940-1949 (pct_sf3_hct015042)	0.67 (1.1)	0-28.3	1.06	1.03-1.08	0.001	71	1.06	1.01-1.11	7	11
%Workers 16 years and over taking public transportation, namely trolley or street cars, to work (pct_SF3_p030007)	0.12 (0.64)	0-14.6	1.10	1.06-1.13	0.0001	23	1.05	1.001-1.09	1	12
%Renter occupied housing units with householder aged 55-64 with no vehicle available (pct_SF3_h045025)	0.54 (0.99)	0-15.1	1.06	1.04-1.09	0.003	120	1.07	1.02-1.12	8	13
%Male Protective Service Occupations: fire-fighting, prevention, and law enforcement workers (pct_SF3_p050026)	0.89 (0.93)	0-16.7	0.93	0.90-0.96	0.04	416	0.94	0.89-0.99	213	14
%Males with earnings of $7500-9,999 in 1999 (pct_SF3_p084006)	1.50 (0.96)	0-37.1	1.07	1.04-1.10	0.01	255	1.05	1.001-1.10	41	15
%Male householder over 65 living alone in nonfamily household (pct_SF1_p030012)	7.2 (2.5)	0-34.5	1.03	1.02-1.04	.005	160	1.07	1.02-1.13	100	16
% 1 unit detached or attached household renters aged 55-64 years (pct_sf3_hct004093)	0.74 (0.75)	0-9.3	1.09	1.05-1.12	0.03	383	1.06	1.01-1.12	53	17

Fig 3. Summary of neighborhood variable "top hits" associated with aggressive prostate cancer by phase. [a]Standard deviation (sd); [b]Confidence or Credible Interval (CI).

https://doi.org/10.1371/journal.pone.0174548.g003

aged 15–34 (OR = 1.06, CI = 1.01–1.11) and percent (%) renter occupied housing units built 1939 or earlier with householder aged 15–24 years (OR = 1.07, CI = 1.02–1.11). Two variables related to social support (%male householder living alone (OR = 1.06, CI = 1.01–1.11) and % male householder over 65 living alone in nonfamily household (OR = 1.07, CI = 1.02–1.13)). The top hit (most significant variable from Phase 2) was %workers >16 years taking trolley or street car public transportation to work (OR = 1.05, CI = 1.001–1.09).

Discussion

We used a novel NWAS to assess the association of 14,663 neighborhood variables with prostate cancer aggressiveness from the PA Cancer Registry. Through a series of progressively more stringent phases, model adjustments, and dimension reduction techniques, we identified the top 17 neighborhood variables associated with aggressive prostate cancer. These findings confirm some previous associations, but also provide new insights into the role of neighborhood in prostate cancer and suggest the potential value of NWAS to inform public health interventions and multilevel studies.

Previous studies of neighborhood and prostate cancer suggest that neighborhoods with poor socioeconomic (SES) circumstances are related to high-grade prostate cancer[29], independent of individual-level exposures[12, 20]. In these studies, SES was measured with deprivation scores and single, *a priori* selected U.S. census variables related to education, income, poverty, housing[30] and employment (S1 File). Our findings support that neighborhood income and poverty (Components 1, 2, 3, 7, 8, 9, 10, 15), employment (Component 14) and housing (Components 3, 4, 5, 9, 11, 13, 16, 17) relate to prostate cancer aggressiveness. However, neighborhood education was not an important determinant of aggressive prostate cancer here, suggesting education could be a confounder rather than a main effect in neighborhood studies, but more study is warranted.

Immigration (Component 9; %foreign-born naturalized citizens at or above poverty) was a significant NWAS finding. Studies of neighborhoods with higher rates of foreign-born immigrants have shown associations with decreased risk for cancer[31]. Even if individuals are diagnosed with late-stage prostate cancer, survival is improved for those who live in high ethnically homogeneous enclaves, suggesting the strong role social support, alone and in conjunction with poverty, may play in prostate cancer progression[31, 32].

The top hit in this analysis related to taking public transportation to work. This variable, as well as not owning a vehicle, relate more to urban, as compared to rural settings, and are also often used as surrogates for access to medical care[33, 34]. Access to care is often cited as a cause of disparity in prostate cancer treatment[34] and survival[33] across both urban and rural settings. Higher cancer incidence and mortality rates are noted in more urban settings, and cases arising from rural environments often are diagnosed at later disease stages[35]. Thus, NWAS findings are plausible and consistent with previously identified sociodemographic domains[2, 36].

From a methodologic standpoint, NWAS provides a new, agnostic approach to neighborhood and contextual variable selection[2]. In the past, one study might define poverty as proportion of households below poverty, another as percentage receiving public assistance. This inconsistency in the choice of neighborhood or contextual variables has limited the ability to make etiologic inferences across studies[27]. Further, previous neighborhood studies often select census variables that represent fewer socioeconomic parameters, for instance, % population below poverty[17][10]. Our NWAS identified more complex, joint effect variables that combined race, age, and poverty information with household or renter status. These more complex variables could provide insights into disease etiology and suggest that interactions

may exist among demographic domains that are often considered individually in current neighborhood studies[10]. For example, percentage of male nonfamily householders living alone AND percentage of male nonfamily householders living alone over 65 appear to represent similar social concepts. However, NWAS separated these variables into two different components. Variables related to single resident households are used as markers of social support [37, 38], and it is possible that they could represent separate or potentially dynamic changes in the role of social support across the lifespan. Thus, given the specificity of NWAS top hits, it is possible they could be used alone or in combination, in future multilevel investigations and to more precisely identify and target geographic areas that are associated with one or more unfavorable NWAS characteristics for disease interventions. However, NWAS is a new methodologic approach and the etiologic significance of the NWAS hits would need to be investigated within those neighborhoods that exhibit these unfavorable characteristics. Further, the utility of NWAS findings for neighborhood risk assessments will need to be determined through comparisons with existing variable selection methods in future studies.

While NWAS methods can be extended in a variety of ways, the current formulation described here has limitations. Area-level data analyses assume individuals residing within the same geographic area experience similar circumstances. In reality, non-residential experience (e.g., work) and individual-level characteristics (e.g., biology, behavior, risk factors) also impact health states. Thus, future NWAS should be conducted in study populations that can adjust for or directly study individual-level SES factors[29]. In addition, standardized data processing, aggregation methods, and geographic boundaries used in administrative datasets can suffer from systematic reporting bias and missingness [6, 16]. Based on our missing data assessments, bias is likely non-differential (S1 File), but future NWAS studies should investigate missingness using both spatial autocorrelation and imputation techniques[39], as well as evaluate the effects of aggregation and geospatial boundary selection using interpolation and point-based, boundary-free approaches [40, 41] [42, 43].

The NWAS approach described here features many of the methodologic requirements previously proposed for GWAS or EWAS studies[23]. A hallmark of GWAS has been replication of discovery findings in comparable study populations. The NWAS presented here focuses on discovery and minimization of false positives through statistical adjustments, without a separate replication population. Under certain circumstances, a single discovery phase[44] and other biologic or functional-based approaches may be favored over statistical replication[45]. For example, the frequency or percentages of census variables may vary by geography, which can bias estimates of association. Interactions between variables (as indicated by the more complex, joint effect variables in the NWAS) are also likely to vary by geography[45]. While comparisons across geographical areas may be undertaken, use of independent datasets to validate findings may mask real differences between these geographies, and may not be appropriate in the NWAS setting. This is a topic that requires further exploration.

This NWAS study demonstrates that high-dimensional data analysis can be applied to large, publically-available datasets and can yield biologically plausible results. This is the first study to systematically, agnostically, and comprehensively evaluate the role of neighborhood-level factors in prostate cancer using "big data" methods. Although NWAS approaches should be tested in other study populations, NWAS addresses methodological limitations in current neighborhood studies, while capitalizing on methodologic approaches used for precision medicine[46] [47]. Further, coupling an NWAS approach with individual-level risk factor information could have implications for multilevel, health disparity studies, as well as precision public health initiatives aimed at identifying and targeting geographic areas in need of intervention efforts across disease sites.

Supporting information

S1 Fig. Year 2000 SF1 variable missingness overview for the state of Pennsylvania.
(TIF)

S2 Fig. Year 2000 SF3 variable missingness overview for the state of Pennsylvania.
(TIF)

S1 File. Methods and data analysis.

1. Data Cleaning
2. Data Mapping
3. Neighborhood-wide Association Study (NWAS) Methods Detail
4. Table. Examples of Neighborhood Methods used in Prostate Cancer Research
(DOCX)

S2 File. Summary of year 2000 U.S. census SF1-Pennsylvania state prostate cancer registry join.
(XLSX)

S3 File. Summary of year 2000 U.S. census SF3-Pennsylvania state prostate cancer registry join.
(XLSX)

S4 File. Phase 1 results.
(XLSX)

S5 File. Phases 2 and 3 results.
(XLSX)

Acknowledgments

These data were supplied by the Bureau of Health Statistics and Research, Pennsylvania Department of Health, Harrisburg, Pennsylvania. The Pennsylvania Department of Health specifically disclaims responsibility for any analyses, interpretations, or conclusions.

Author Contributions

Conceptualization: SML NM TRR CB HR TJ.

Data curation: SML KD CN.

Formal analysis: SML KD CN MR.

Funding acquisition: SML TRR.

Investigation: SML.

Methodology: SML NM MR TJ.

Project administration: SML.

Resources: TJ CZJ.

Software: CN KD.

Supervision: SML CZJ.

Validation: CN KD MR CZJ.

Visualization: SML.

Writing – original draft: SML.

Writing – review & editing: SML NM MR TJ CB CZJ TRR HR.

References

1. Lynch SM, Rebbeck TR. Bridging the Gap between Biologic, Individual, and Macroenvironmental Factors in Cancer: A Multilevel Approach. Cancer Epidemiology Biomarkers & Prevention. 2013; 22 (4):485–95. https://doi.org/10.1158/1055-9965.epi-13-0010
2. Sampson RJ, Morenoff JD, Gannon-Rowley T. Ann Rev Sociol 2002; 28:443–478. "Assessing Neighborhood Effects": Social Processes and New Directions in Research. Ann Rev Sociol. 2002;28:443–78.
3. Eeles RA, Kote-Jarai Z, Giles GG, Olama AA, Guy M, Jugurnauth SK, et al. Multiple newly identified loci associated with prostate cancer susceptibility. Nat Genet. 2008; 40(3):316–21. https://doi.org/10.1038/ng.90 PMID: 18264097
4. Varghese JS, Easton DF. Genome-wide association studies in common cancers—what have we learnt? Current Opinion in Genetics & Development. 2010; 20(3):201–9.
5. Li H, Achour I, Bastarache L, Berghout J, Gardeux V, Li J, et al. Integrative genomics analyses unveil downstream biological effectors of disease-specific polymorphisms buried in intergenic regions. Npj Genomic Medicine. 2016; 1:16006. http://www.nature.com/articles/npjgenmed20166#supplementary-information. PMID: 27482468
6. Patel CJ, Bhattacharya J, Butte AJ. An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus. PLoS One. 2010; 5(5):e10746. https://doi.org/10.1371/journal.pone.0010746 PMID: 20505766
7. Hill TD. Neighborhood disorder, psychological distress, and heavy drinking. Social Science & Medicine. 2005; 61(5):965–75.
8. Geronimus AT, Hicken M, Keene D, Bound J. "Weathering" and Age Patterns of Allostatic Load Scores Among Blacks and Whites in the United States. Am J Public Health. 2006; 96:826–33. https://doi.org/10.2105/AJPH.2004.060749 PMID: 16380565
9. McEwen B. Protective and Damaging effects of stress Mediators. N Engl J Med. 1998; 338:171–9. https://doi.org/10.1056/NEJM199801153380307 PMID: 9428819
10. Ziegler-Johnson C, Weber A, Glanz K, Spangler E, Rebbeck TR. Gender- and Ethnic-specific Associations with Obesity: Individual and Neighborhood-level Factors. J Natl Med Assoc. 2013; 105(2):173–82. PMID: 24079218
11. Byers TE, Wolf HJ, Bauer KR, Bolick-Aldrich S, Chen VW, Finch JL, et al. The impact of socioeconomic status on survival after cancer in the United States. Cancer. 2008; 113(3):582–91. https://doi.org/10.1002/cncr.23567 PMID: 18613122
12. Carpenter W, Howard D, Taylor Y, Ross L, Wobker S, Godley P. Racial differences in PSA screening interval and stage at diagnosis. Cancer Causes and Control. 2010; 21(7):1071–80. https://doi.org/10.1007/s10552-010-9535-4 PMID: 20333462
13. Lyratzopoulos G, Barbiere JM, Greenberg DC, Wright KA, Neal DE. Population based time trends and socioeconomic variation in use of radiotherapy and radical surgery for prostate cancer in a UK region: continuous survey. BMJ. 2010;340.
14. Crawford ED. Epidemiology of Prostate Cancer. Urology. 2003; 62:3–12.
15. Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet. 2007; 39(5):645–9. https://doi.org/10.1038/ng2022 PMID: 17401363
16. United States Census Bureau. 2000 Census Technical Documentation for SF 3. United States Department of Commerce. 2007.
17. Diez Roux AV, Jacobs DR, Kiefe CI. Neighborhood characteristics and components of the insulin resistance syndrome in young adults: the coronary artery risk development in young adults (CARDIA) study. Diabetes Care. 2002; 25(11):1976–82. PMID: 12401742
18. Robert SA, Strombom I, Trentham-Dietz A, Hampton JM, McElroy JA, Newcomb PA, et al. Socioeconomic risk factors for breast cancer: distinguishing individual- and community-level effects. Epidemiology. 2004; 15(4):442–50. PMID: 15232405

19. Part III, Chapter 27 Communical and Noncommunicable Diseases, Section 27.31, Act 224 of 1980 The Pennsylvania Cancer Control, Prevention, and Research Act (1980).
20. Zeigler-Johnson C, Tierney A., Rebbeck TR, Rundle A. Prostate Cancer Severity Associations with Neighborhood Deprivation. Prostate Cancer 2011.
21. SEER Program Coding and Staging Manual 2000. In: National Cancer Institute. Surveillance E, and Endpoints Research (SEER). editor. Bethesda, MD2000.
22. Eeles RA, Kote-Jarai Z, Al Olama AA, Giles GG, Guy M, Severi G, et al. Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. Nat Genet. 2009; 41 (10):1116–21. http://www.nature.com/ng/journal/v41/n10/suppinfo/ng.450_S1.html. PMID: 19767753
23. Pearson TA, Manolio TA. HOw to interpret a genome-wide association study. JAMA. 2008; 299 (11):1335–44. https://doi.org/10.1001/jama.299.11.1335 PMID: 18349094
24. Hubbard AE, Ahern J, Fleischer NL, Laan MVd, Lippman SA, Jewell N, et al. To GEE or Not to GEE: Comparing Population Average and Mixed Models for Estimating the Associations Between Neighborhood Risk Factors and Health. Epidemiology. 2010; 21(4):467–74. https://doi.org/10.1097/EDE.0b013e3181caeb90 PMID: 20220526
25. Aickin M, Gensler H.. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. Am J Public Health. 1996; 86:726–8. PMID: 8629727
26. Ru H, Martino S. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. JR Statist Soc B. 2008; 71(2):319–92.
27. Messer L, Laraia B, Kaufman J, Eyster J, Holzman C, Culhane J., et al.. The development of a standard neighborhood deprivation index. Journal of Urban Health. 2006; 83(6):1041–62. https://doi.org/10.1007/s11524-006-9094-x PMID: 17031568
28. Meuwissen TH, Goddard ME. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics. 2000; 155(1):421–30. PMID: 10790414
29. Diez Roux AV, Mair C. Neighborhoods and health. Annals of the New York Academy of Sciences. 2010; 1186(1):125–45.
30. Thomson H, Thomas S, Sellstrom E, Petticrew M. Housing Improvements for Health and Associated Socio-economic Outcomes. Cochrane Database Syst Reve. 2013.
31. Schupp CW, Press DJ, Gomez SL. Immigration factors and prostate cancer survival among Hispanic men in California: does neighborhood matter? Cancer. 2014; 120(9):1401–8. https://doi.org/10.1002/cncr.28587 PMID: 24477988
32. Carriere GM, Sanmartin C, Bryant H, Lockwood G. Rates of cancer incidence across terciles of the foreign-born population in Canada from 2001–2006. Can J Public Health. 2013; 104(7):e443–9. PMID: 24495818
33. Guidry JJ, Aday LA, Zhang D, Winn RJ. Transportation as a barrier to cancer treatment. Cancer Pract. 1997; 5(6):361–6. PMID: 9397704
34. Patel K, Kenerson D, Wang H, Brown B, Pinkerton H, Burress M, et al. Factors influencing prostate cancer screening in low income African Americans in Tennessee. J Health Care Poor Underserved. 2010; 21(1 Suppl):114–26. https://doi.org/10.1353/hpu.0.0235 PMID: 20173288
35. Monroe AC, Ricketts TC, Savitz LA. Cancer in Rural versus Urban Populations: A Review. J Rural Health. 1992; 8(3):212–20. PMID: 10121550
36. Dietz RD. The estimation of neighborhood effects in the social sciences: An interdisciplinary approach. Social Science Research. 2002; 31:539–75.
37. Galster GC. The Mechanisms of Neighbourhood Effects: Theory, Evidence, and Policy Implications. In: van Ham M, Manley D, Bailey N, Simpson L, Maclennan D, editor. Neighbourhood Effects Research: New Perspectives. Dordrecht: Springer Netherlands; 2012. p. 23–56.
38. Thompson EE, Krause N. Living Alone and Neighborhood Characteristics as Predictors of Social Support in Late Life. J Gerontol B Psychol Sci Soc Sci. 1998; 53(6):S354–64. PMID: 9826977
39. Paternoster L, Zhurov Alexei I, Toma Arshed M, Kemp John P, St. Pourcain B, Timpson Nicholas J, et al. Genome-wide Association Study of Three-Dimensional Facial Morphology Identifies a Variant in PAX3 Associated with Nasion Position. The American Journal of Human Genetics. 90(3):478–85. https://doi.org/10.1016/j.ajhg.2011.12.021 PMID: 22341974
40. Geronimus AT. Invited Commentary: Using Area-based Socioeconomic Measures—Think Conceptually, Act Cautiously. American Journal of Epidemiology. 2006; 164(9):835–40. https://doi.org/10.1093/aje/kwj314 PMID: 16968860
41. Krieger N, Chen JT, Waterman PD, Soobader M-J, Subramanian SV, Carson R. Geocoding and Monitoring of US Socioeconomic Inequalities in Mortality and Cancer Incidence: Does the Choice of Area-

based Measure and Geographic Level Matter?: The Public Health Disparities Geocoding Project. American Journal of Epidemiology. 2002; 156(5):471–82. PMID: 12196317

42. Branas CC, Cheney RA, MacDonald JM, Tam VW, Jackson TD, Ten Have TR. A Difference-in-Differences Analysis of Health, Safety, and Greening Vacant Urban Space. American Journal of Epidemiology. 2011; 174(11):1296–306. https://doi.org/10.1093/aje/kwr273 PMID: 22079788

43. Kondo MC, Keene D, Hohl BC, MacDonald JM, Branas CC. A Difference-In-Differences Study of the Effects of a New Abandoned Building Remediation Strategy on Safety. PLoS ONE. 2015; 10(7): e0129582. https://doi.org/10.1371/journal.pone.0129582 PMID: 26153687

44. Thomas DC, Casey G, Conti DV, Haile RW, Lewinger JP, Stram DO. Methodological Issues in Multistage Genome-wide Association Studies. Statistical science: a review journal of the Institute of Mathematical Statistics. 2009; 24(4):414–29.

45. Aslibekyan S, Claas SA, Arnett DK. To Replicate or Not to Replicate: The Case of Pharmacogenetic Studies: Establishing Validity of Pharmacogenomic Findings: From Replication to Triangulation. Circulation: Cardiovascular Genetics. 2013; 6(4):409–12.

46. Moore JH, Ritchie MD. The Challenges of Whole-Genome Approaches to Common Diseases. JAMA. 2004; 291.

47. Department of Health and Human Services. National Institutes of Health. Help me Understand Genetics: Precision Medicine. 2017. National Library of Medicine. Bethesda, MD. https://ghr.nlm.nih.gov/primer/precisionmedicine/initiative.