CANCER IN PHILADELPHIA NEIGHBORHOODS

TECHNICAL APPENDIX

Technical Appendix

This appendix contains the definition of terms and methodology for the report "Cancer in Philadelphia Neighborhoods."

The Technical Appendix includes the definition of terms, description of data sources and data quality, modeling approach, information about data suppression, and caution in interpretation.

Section 1. Definition of Terms

Neighborhoods: Philadelphia neighborhoods are groups of census tracts. The neighborhood definitions used in this report are described in section 2.1.

Age-specific incidence/mortality rate: Defined as the number of events (new cancer cases or deaths) in a specified age group per 100,000 population within the same age group.

Direct method of age-adjustment: A method of age standardization in which age-specific rates are applied to a standard population to derive a summary age-adjusted rate.[1] In this method, population weights derived from the age distribution of the standard population are used to combine the modeled age-specific incidence or mortality rates.

Age-adjusted incidence/mortality rate and standard population: Crude rates are influenced by the underlying age distribution of the population. For instance, a geographic area with a relatively older population has a higher crude rate because the incidence and mortality rates for most cancers increase with increasing age. To account for this, a standard population and age specific rates are used to derive an age-adjusted rate using the direct method of adjustment. This ensures the comparability of rates from one year to another, or from one geographic unit to another.

We have used the 2000 U.S. standard population to calculate the age-adjusted rates by the direct method (as described above). The 2000 U.S. standard million population is based on the population for each single year of age as reported in the Census P25-1130 series. Single years were combined to form 19 age groups: 0 years, 1-4 years, 5-9 years, 10-14 years, 15-19 years, 20-24 years, 25-29 years, 30-34 years, 35-39 years, 40-44 years, 45-49 years, 50-54 years, 55-59 years, 60-64 years, 65-69 years, 70-74 years, 75-79 years, 80-84 years, and 85 years and above.

Due to small sample sizes, for the analysis, we collapsed the 19 age-groups in the standard population into the following 6 age groups and used them to calculate the age-adjusted rates: 00-34 years, 35-44 years, 45-54 years, 55-64 years, 65-74 years, and 75 years and above. The proportion of population in those 6 age groups serve as the weights in computing age-adjusted rates (Table 1.1). Note that the standard million weights are not race- or sex-specific and do not adjust for race or sex differences between geographic units or populations.

Table 1.1. Age distributions and age-adjustment weights based on the 2000 U.S. standard population

Age Group	Population	Adjustment Weight
All Ages	274,633,642	1.0
00-34 years	134,273,319	0.488919
35-44 years	44,659,185	0.162613
45 to 54 years	37,030,152	0.134834
55-64 years	23,961,506	0.087247
65-74 years	18,135,514	0.0660
75 years and above	16,573,966	0.060349

Prevalence rates: Proportions of persons who completed screening or reported a cancer-related risk factor (referred to as prevalences or prevalence rates) were estimated from the survey data. Prevalence rates were age-adjusted to the Philadelphia population age distribution using the direct method and the American Community Survey (ACS) 2014-2018 5-year aggregate population estimates, as described in section 3.2. As noted in Table 1.2 specific estimates are limited to specific age groups. Neighborhoods were then aggregated across the city to derive city-wide estimates.

Table 1.2. Age groups used in age-adjustment of prevalence estimates, by outcome measure and population subgroup.

	Age Groups
All Risk Factors	18-34 years, 35-44 years, 45-64 years, 65 years and above
Cervical Cancer Screening	18-34 years, 35-44 years, 45-64 years
Breast Cancer Screening and Colorectal Cancer Screening	50-59 years, 60-64 years, 65-74 years

Model-based estimates: Model-based estimates are derived from a statistical model that was used to generate smoothed estimates of cancer incidence, mortality, and prevalence rates. These estimates are the median of the samples of the posterior distribution from a Bayesian model as described in section 3. The model-based estimate for incidence/mortality rate is adjusted to the U.S. 2000 standard million population. The model-based estimate of the prevalence rate for cancer screening and risk factors is adjusted to the Philadelphia ACS 2014-2018 population.

Credible interval: In Bayesian statistics, the interval within which an unobserved parameter value falls with some probability is referred as the credible interval. The 95% credible intervals in this report are constructed using values corresponding to the 2.5th and the 97.5th percentiles of the posterior distribution.

Section 2. Sources of Data and Data Quality

2.1. Neighborhood Definition

Neighborhoods were designed to be small enough to represent meaningful distinctions within Philadelphia, while being large enough to have adequate data. We started with predefined neighborhood definitions initially created for the Southeastern Pennsylvania Household Health Survey administered by the Public Health Management Corporation (PHMC). PHMC identified 45 neighborhoods in Philadelphia based on groupings of contiguous census tracts using 2000 Census Tract boundaries. These neighborhoods were then aligned to 2010 Census Tract boundaries and excluded tracts designated as special land-use tracts with little or no residential population and special characteristics such as large parks or employment area (n=12 tracts). From these initial PHMC neighborhoods, we modified boundaries using local knowledge. Due to large population size, we separated the Center City neighborhood into two: Center City East and Center City West. This process results in 46 neighborhoods, containing approximately 4 to 16 census tracts per neighborhood, with a median population of 31,851 (range: 19,864-54,652) based on American Community Survey (ACS) 2014-2018 population estimates. These neighborhood definitions have been used in other reports for the Philadelphia Department of Public Health.[2]

2.2. Incidence

The cancer incidence data in this report are for 2012-2016 and from the Pennsylvania Cancer Registry (PCR), a part of the National Program of Cancer Registries (NPCR) that has been collecting cancer incidence data since 1985. PCR collects and houses cancer data obtained from the reports of medical facilities and laboratories where the cancer is diagnosed and/or treated. These data include information on cancer patients such as their demographic characteristics, as well as the type of cancer, the site where it first appeared (primary site), the extent of disease (stage), and the treatments patients received. The International Classification of Diseases for Oncology, 3rd revision (ICD-O-3) codes were used to classify primary sites.[3] In addition to all types of cancers together, we investigated six specific cancer sites with ICD-O-3 codes listed below:

- *Colorectal Cancer*: C180, C181, C182, C183, C184, C185, C186, C187, C188, C189, C260, C199, C209.
- *Lung and Bronchus:* C340-C349
- *Liver Cancer:* C220.
- *Prostate Cancer*: C619
- *Breast Cancer:* C500-C509

For the above six cancers, we excluded histology types 9050-9055 (mesothelioma), 9140 (Kaposi Sarcoma), 9590-9992 (some lymphoid and haematopoietic cancers) to be consistent with National Cancer Institute, Surveillance, Epidemiology, and End Results Program (SEER) reporting and to include only cases that were microscopically confirmed. We limited our analyses to malignant, primary site (invasive) cancers using the behavior code provided in the cancer registry, and excluding unknown, benign or uncertain behavior cancers. All incident cases with a behavior code of 3 are deemed reportable to the NPCR and SEER cancer registries. The residential addresses were geocoded using the ESRI Business Analyst 2016 geocoder. Only the cases with residential addresses that were matched to Philadelphia 2010 census tracts were included. If a person had more than one primary tumor, each tumor was counted as a separate case. Thus, our analysis of cancer incidence in this report is not based on the number of individuals diagnosed with cancers but rather on the number of primary tumors diagnosed. The race was determined based on primary race. The cancer registry has five items to denote race (race1-race5), from which we used the primary race (race1) to categorize cases into racial groups. Primary races other than white and blacks were considered as 'Other' races.

Cancers in patients of unknown or missing age, sex, and race were omitted from analysis which amounted to loss of 2.59 % of incident cases.

[1] Colorectal cancer captures the cancers of colon, rectum and rectosigmoid junction.
[2] Liver cancer does not include cancer of intrahepatic bile duct.
[3] Although it is possible for males to develop breast cancer, we limited our analysis to female breast cancer.

2.3. Mortality

The cancer mortality data in this report are for years 2012-2016 and were supplied by the Bureau of Health Statistics & Registries, Pennsylvania Department of Health, Harrisburg, Pennsylvania. Death certificates are typically completed by medical staff or funeral directors and sent to state vital statistics office. A standard death certificate contains the information on underlying (immediate) and contributing causes of deaths coded according to the International Statistical Classification of Disease and related Health Problems (ICD) as required by World Health Organization (WHO) regulations for its member nations. Underlying causes of deaths in our study period (2012-2016) were coded according to the 10th revision of ICD Classification. We included all deaths where the malignant cancer was the underlying cause of death (ICD-10 codes: C00-C97). The ICD-10 codes for the six site-specific cancers we investigated are presented below:

- *Colorectal Cancer:* C18, C26.0, C19-C20, C21
- *Lung and Bronchus*: C340
- *Liver Cancer*: C22.0, C22.2-C22.4, C22.7, C22.9
- *Prostate Cancer*: C61
- *Breast Cancer:* C50

Residential addresses were geocoded using the ESRI Business Analyst 2016 geocoder. Only records geocoded to Philadelphia census tracts were included. The race information was obtained from the 'race' variable in the death certificates. Deceased with the race designation other than 'White' and 'Black' were considered as 'Other' races. Death records with unknown or missing age, sex, and race were omitted from the analysis which comprises 1.15% of deaths.

2.4. Population data for incidence and mortality

For incidence and mortality rates, the race, sex, and age-group specific population estimates for Philadelphia census tracts was obtained from the American Community Survey 5-year aggregate (ACS 2010-2014, 2011-2015, 2012-2016, 2013-2017, 2014-2018). The population for White alone comes from the ACS table 'B01001A' and the population for Black alone comes from the ACS table 'B0100B'. The population for Other races was obtained by subtracting the white alone and black alone population from the total population obtained from ACS table 'B01001'. We assumed that the mid-points of these 5 surveys correspond to years of our mortality and incidence data which includes years between 2012 and 2016. To estimate the rates pooled for 2012-2016, we summed the population from these surveys for each census tract, by age group, sex, and race. These census tract level cumulative populations were then aggregated to the Philadelphia neighborhoods to obtain neighborhood-level population estimates.

2.5. Cancer Screening and Risk Factors

The primary data source for cancer screening and cancer-related risk factors was the Public Health Management Corporation's Southeastern Pennsylvania Household Health Survey (SEPAHHS).[4] SEPAHHS is a series of cross-sectional surveys that collect data on health and social well-being on more than 10,000 households in Bucks, Chester, Delaware, Montgomery, and Philadelphia Counties administered by PHMC. The survey contains information about local residents' health status, health behaviors, and access to care. The data were collected through a random digit dialing telephone survey, which since 2008 has included cell phone users. For this study, we used SEPAHHS data on adult respondents (18 years and older) in Philadelphia County collected in 2015 and 2018, pooled together. Cancer screening and related risk factors were measured using information self-reported as part of the telephone questionnaire. Survey questions and definitions are found in Table 2.

Table 2. Survey questions and definitions for self-reported cancer screening, risk factor measures, and race/ethnicity

Measure	Description	SEPAHHS Questions	Responses
Cancer screening			
Colorectal Cancer Screening	Having a colonoscopy or sigmoidoscopy in the past 10 years. Only using data for adults aged 50-74.	About how long has it been since you last had a colonoscopy or a sigmoidoscopy? These tests are performed to screen for colorectal cancer.	Yes: in the past ten years or less No: more than 10 years ago or never
Breast Cancer Screening	Having a mammogram within the past 2 years. Only using data for females aged 50-74.	About how long has it been since you last had a mammogram?	Yes: in the past two years or less No: more than two years ago or never
Cervical Cancer Screening	Having a pap smear within the past 3 years. Only using data for females aged 18-64.	About how long has it been since you last had a Pap smear test?	Yes: in the past three years or less No: more than three years ago or never
Risk Factors			
Obesity	Obesity as body mass index of 30 kg/m2 or higher	BMI category computed from respondent's self-reported weight and height.	Yes: obese (BMI 30 or greater) No: overweight; normal weight; underweight
Diabetes	Respondent told by a doctor or health professional that they have or had diabetes.	Have you ever been told by a doctor or other health professional that you had diabetes?	Yes: has been diagnosed with diabetes No: no diabetes; pre-diabetes; borderline diabetes; diabetes only during pregnancy
Current Smoking	Currently smoking cigarettes	Do you now smoke cigarettes? Do you now smoke cigarettes every day, some days, or not at all?	Yes: smokes cigarettes every day or some days No: does not smoke cigarettes
Physical Activity	Exercising at least 30 minutes for at least 3 days per week	Thinking about the past month, how many times per week did you participate in any physical activities for exercise that lasted for at least one half-hour, such as walking, basketball, dance, rollerblading or gardening?	Yes: three or more times per week No: less than three times per week
Fruit/Vegetable Consumption	Eating five or more servings of fruits or vegetables per day meeting United States Department of Agriculture recommendations[5].	How many servings of fruits and vegetables do you eat on a typical day? A serving of a fruit or vegetable is equal to a medium apple, half a cup of peas or half a large banana.	Yes: five or more servings per day No: less than five servings per day

Measure	Description	SEPAHHS Questions	Responses
Binge Drinking	Males: drinking five or more alcoholic beverages on one occasion in the past month. Females: drinking four or more alcoholic beverage in the past month. Only from survey year 2018.	During the past 30 days, on how many different days did you have (five/four) or more drinks on at least one occasion? One drink is equivalent to a 12-ounce beer, a 5-ounce glass of wine, or a drink with one shot of liquor. Men are asked how many days they have had 5 or more drinks. Women are asked how many days they have had 4 or more drinks.	Yes: at least one day of four or more drinks for females, five or more drinks for males No: zero days of drinking four or more drinks for females, five or more drinks for males
Sugar-Sweetened Beverage Consumption	Drinking at least one sugar-sweetened beverage (soda or juice) per day.	During the PAST MONTH, how many times per day, week, or month did you drink SODA such as Coke or 7-Up? Do not include diet soda. During the PAST MONTH, how many times per day, week, or month did you drink FRUIT DRINKS or BOTTLED TEAS such as Snapple, Hugs, lemonade, or Kool-Aid? Do not include diet drinks.	Yes: drinks one or more soda per day; OR drinks one or more fruit drink or bottled tea per day No: drinks soda less than once per day; AND drinks fruit drinks or bottled tea less than once per day
Demographic Variables			
Race	Race	Which of these groups would you say best represents your race?	White Black Other: Asian or Pacific Islander, American Indian or Alaska Native, Biracial or multiracial, Hispanic/Latino (voluntary), Something else (specify)

2.6. Population data for cancer screening and risk factors

The modeling approach we used to derive the smoothed and adjusted prevalence estimates from the survey data requires the use of census tract level population data (see section 3.2). Census tract-level population count data were obtained from the American Community Survey (ACS) 2014-2018 5-year aggregate population estimates by race, age, and sex.[6] Population estimates were obtained from the ACS 2014-2018 tables for sex by age: B01001 for the total population, B01001A for White alone, and B01001B for Black or African American alone. The population estimates for Other races was obtained by subtracting the White alone and Black alone population from the total population. Population counts were aggregated by sex and race into age groups 18-34 years, 35-44 years, 45-64 years, and 65 years and over.

Since colorectal cancer screening and breast cancer screening are only recommended for individuals between ages 50 and 74, the analyses for these screening outcomes require the use of age groups that fall within these bounds. The cut points used for age groups in the ACS population data by race and sex (including ages 45-54 years, 55-64 years, and 65-74 years) are mismatched to the cancer screening age bounds (50 and 74 years) and the age groups used in the PHMC survey (50-59 years, 60-64 years, and 65-74 years). Thus, we imputed population counts by race and sex for the required cancer screening age groups.

To impute the needed cancer screening population counts, we assumed a uniform distribution of the population across the single-year ages within a broad age group (e.g., age group 45-54 years). We first determined the total population count by sex and race within each census tract for the available age groups (45-54 years, 55-64 years, and 65-74 years). Then, the 10-year age group population count estimates for ages 45-54 years and 55-64 years were divided in half to produce 5-year age group estimates. These 5-year age groups were recombined as needed and the population estimates were summed, resulting in imputed estimates of tract-level population counts by race and sex for the required cancer screening age groups (50-59 years, 60-64 years, and 65-74 years). These census tract level cumulative populations were then aggregated to the Philadelphia neighborhoods to obtain neighborhood-level population estimates.

2.7. Neighborhood level demographic and socioeconomic position indicators

Several census tract-level demographic and socioeconomic position indicators were derived from American Community Survey (ACS) 2014-2018 5-year aggregate estimates. Tract-level ACS counts were summed over tracts that were aggregated into neighborhoods as described in Section 1 to produce sex-specific and total neighborhood-level numerators and denominators used to calculate proportion estimates as shown in Table 3. Neighborhood-level median household income was calculated as a weighted median of the census tract-level ACS estimates, using the number of households per census tracts as weights. Median household income is not available as sex-specific.

Table 3. Demographic and socioeconomic position indicators definitions

Measure	Description	American Community Survey data table	Numerator details	Denominator details
Total population	Total number of persons	B01001 (SEX BY AGE)	total population	n/a
Age under 18 years	Percent of persons less than 18 years of age	B01001 (SEX BY AGE)	individuals aged 0 to 17 years	Entire population
Age 18-44 years	Percent of persons 18-44 years of age	B01001 (SEX BY AGE)	individuals aged 18 to 44 years	Entire population
Age 45-64 years	Percent of persons 45-64 years of age	B01001 (SEX BY AGE)	individuals aged 45 to 64 years	Entire population
Age 65 years and older	Percent of persons 65 years of age and older	B01001 (SEX BY AGE)	individuals aged 65 years and older	Entire population
Race/ethnicity: Hispanic	Percent of persons who are Hispanic/Latino ethnicity	B01001I (SEX BY AGE (HISPANIC OR LATINO)) B01001 (SEX BY AGE)	Individuals reporting as Hispanic/Latino ethnicity	Entire population
Race/ethnicity: Non-Hispanic Black	Percent of persons who are non-Hispanic/ Latino Black	B01001B (SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE)) B01001 (SEX BY AGE)	Individuals reporting as Black or African American alone	Entire population
Race/ethnicity: Non-Hispanic Asian/Pacific Islander	Percent of persons who are non-Hispanic/ Latino Asian, Native Hawaiian or Pacific Islander	B01001D (SEX BY AGE (ASIAN ALONE)) B01001E (SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE)) B01001 (SEX BY AGE)	Individuals reporting as Combined race categories: Asian alone, Native Hawaiian or Pacific Islander alone	Entire population
Race/ethnicity: Non-Hispanic white	Percent of persons who are non-Hispanic/ Latino white	B01001A (SEX BY AGE (WHITE ALONE)) B01001 (SEX BY AGE)	Individuals reporting as white alone	Entire population

Measure	Description	American Community Survey data table	Numerator details	Denominator details
Median household income	The weighted median of tract-level values of median household income, weighted by the number of households per census tract	B19013 (MEDIAN HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2018 INFLATION-ADJUSTED DOLLARS)) B11001 (HOUSEHOLD TYPE (INCLUDING LIVING ALONE))	n/a	n/a
Poverty	Percent of persons who are living below the federal poverty level	B17001 (POVERTY STATUS IN THE PAST 12 MONTHS BY SEX BY AGE)	individuals who are below poverty level	Entire population with poverty level determined
Education: High school degree or less	Percent of persons aged 25 and older who have high school degree or less education	B15002 (SEX BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 25 YEARS AND OVER)	individuals aged 25 and older with high school diploma or equivalent or less than HS education	Adults aged 25 and older
Education: Bachelor's degree or higher	Percent of persons aged 25 and older who have Bachelor's degree or higher	B15002 (SEX BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 25 YEARS AND OVER)	individuals aged 25 and older with 4-year bachelor degree and more advanced professional and graduate degrees	Adults aged 25 and older
Unemployment	Percent of civilian work force aged 16 and older who are unemployed	B23001 (SEX BY AGE BY EMPLOYMENT STATUS FOR THE POPULATION 16 YEARS AND OVER) B23025 (EMPLOYMENT STATUS FOR THE POPULATION 16 YEARS AND OVER)	unemployed civilian work force population aged 16 and older	Population aged 16 and older who are part of the civilian work force (not armed forces)
Uninsured	Percent of persons who are uninsured	B27001 (HEALTH INSURANCE COVERAGE STATUS BY SEX BY AGE)	individuals with no health insurance	Entire population

3. Modeling Approach

3.1. Incidence and Mortality Rates

Due to small counts of deaths and incident cases for some neighborhoods and types of cancer, estimates of incidence or mortality rates are often highly variable. We fit Bayesian models using a Markov Chain Monte Carlo (MCMC) algorithm to obtain stable estimates of cancer incidence and mortality rates in Philadelphia neighborhoods using the 5 years of mortality data between 2012 and 2016 in Philadelphia neighborhoods. These models use a similar approach as Quick 2020.[7]

For each combination of age group a $\{a=1,2,...,N_a; N_a=5\}$, race/ethnicity r $\{r = 1, ..,N_r; N_r=3\}$, sex s $\{s = 1,..,N_s; N_s=2\}$, and neighborhood i $\{i=1,2,..., N_i; N_i=46\}$ with corresponding population n_{iars}, we assume that the number of events (death counts or incidence cases) for any given cancer site comes from a Poisson distribution with underlying true rate n_{iars}.

$$y_{iars} \mid \lambda_{iars} \sim Pois\,(n_{iars}\,\lambda_{iars})$$

We then used linear regression to model the natural logarithm of the rate as follows:

$$ln\,\lambda_{iars} = \beta_{0;ars} + z_{irs}$$

where $\beta_{0;ars}$ is the age, race, and sex-specific intercept for all neighborhoods and z_{irs} is the random effect that varies by neighborhood, race, and sex.

We begin by modeling the intercepts, where we assume a multivariate normal distribution:

$$\boldsymbol{\beta}_{0;\bullet rs} \sim MVN(\boldsymbol{\mu}_{\bullet s}, \boldsymbol{G}_{\beta_0;s})$$

where $\boldsymbol{\beta}_{0;\bullet rs} = \left(\beta_{0;1rs}, \ldots, \beta_{0;N_a rs}\right)$ denotes the vector of N_a age-specific intercept parameters for each combination of race and sex. The mean vector $\boldsymbol{\mu}_{\bullet s} = \left(\mu_{1s}, \ldots, \mu_{N_a s}\right)$ with elements μ_{as} plays the role of non-race-specific general age/sex intercept estimates. We specified all μ_{as} to have an uninformative uniform prior. Specifically, for the analysis of all cancers combined, we specified $\mu_{as} \sim Uniform\ (-5,5)$ and other site-specific cancers, we specified $\mu_{as} \sim Uniform\ (-15,15)$.

While we want to allow different age groups within each combination of race and sex to have different intercept estimates, we do believe that these values may be correlated with one another. As such, we define $\boldsymbol{\beta}_{0;\bullet rs}$ to have the age-group covariance structure $\boldsymbol{G}_{\beta_0;s}$, which is a sex-specific N_a x N_a square matrix. The elements $g_{\beta_0;s;a,a'}$ correspond to the covariance between age groups a and a' for sex.

$$\boldsymbol{G}_{\beta_0;s} = \begin{bmatrix} g_{\beta_0;s;1,1} & \cdots & g_{\beta_0;s;1,N_a} \\ \vdots & \ddots & \vdots \\ g_{\beta_0;s;N_a,1} & \cdots & g_{\beta_0;s;N_a,N_a} \end{bmatrix}$$

The sex-specific age-covariance structure $G_{\beta_0;s}$ is not race-specific. All N_r race groups use the same estimated age covariance structures, therefore allowing groups to borrow information from each other in estimating the intercepts $\boldsymbol{\beta}_0$. The prior distribution of $G_{\beta_0;s}$ is an inverse Wishart with the scale matrix $R_{G'}$ a diagonal matrix of size $N_a X N_a$ and degrees of freedom chosen to be $\nu = N_a + 1.$

The elements $G_{a,a}$, of matrix R_G are chosen to reduce prior informativeness:

$$g_{a,a'} = \begin{cases} 0, & \text{if } a \neq a' \\ 0.5, & \text{if } a = a' \end{cases}$$

While the intercepts $\boldsymbol{\beta}_0$ estimate the city-level estimates for age, race, and sex group, we include the random effect term Z_{irs} to capture the neighborhood level effect of race and sex on the mortality and incidence rates.

We use a similar approach to estimate the random effects as we did for the intercepts. We define vectors of $N_r N_s$ random effects $z_{i\bullet\bullet}$ for each neighborhood, having elements Z_{irs} representing the random effect for each combination of neighborhood, race, and sex.

The vector of random effects $z_{i\bullet\bullet}$ with terms Z_{irs} is assumed to have multivariate normal distribution, centered around a $N_r N_s$ vector of zeros, with a covariance structure G_z:

$$z_{i\bullet\bullet} \sim MVN(0, G_z)$$

The covariance structure G_z is a $N_r N_s$ x $N_r N_s$ matrix having elements $g_{z;rs,r's'}$ where the off-diagonal elements represents the covariances between race-sex groups, rs and $r's'$. The covariance structure G_z is assumed to have an inverse Wishart prior distribution with the scale matrix G_{z0} and degrees of freedom chosen to be $\nu = N_r N_s + 1$. The elements $g_{z0;rs,r's'}$ of G_{z0} are chosen to be minimally informative.

$$g_{z0;rs,r's'} = \begin{cases} 0, & \text{if } r \neq r' \text{ or } s \neq s' \\ 0.5, & \text{if } r = r' \text{ and } s = s' \end{cases}$$

The mortality and incidence rate for each combination of sex, race, age, for each neighborhood were obtained by running an MCMC algorithm with two chains for 150,000 iterations in WinBUGS using the R package R2WinBUGS to draw samples from the posterior distribution.[8] Because MCMC algorithms require the specification of (often) arbitrarily chosen initial values, the first batch of 100,000 samples were discarded as part the so-called "burn-in" period. We then "thinned" the remaining 60,000 iterations' worth of samples from each chain by a factor of 10 to reduce the autocorrelation between the samples. This produced a final set of 4,000 posterior samples per chain on which our estimates are based.

Age-adjusted rate estimates, denoted $\hat{\lambda}_{\cdot irs}$, were calculated by computing the weighted average of the age-specific rates, $\lambda_{\cdot irs}$, where the weights, w_a, correspond to the 2000 U.S. standard population. These calculations are based on each set of posterior samples i.e.,

$$\hat{\lambda}_{\cdot irs}^{(\ell)} = \sum_{a=1}^{N_a} w_a \times \lambda_{iars}^{(\ell)},$$

where $\ell = 1, \ldots, L$ and L denotes the total number of posterior samples.

Using the estimated age-adjusted rates $\hat{\lambda}_{\cdot irs}$ and neighborhood, race, and sex-specific population n_{irs}, we estimated the sex-specific mortality rates, $\hat{\lambda}_{\cdot irs}^{(\ell)}$ for each neighborhood as follows:

$$\hat{\lambda}_{is}^{(\ell)} = \frac{\sum_{r=1}^{N_r} \left[\hat{\lambda}_{\cdot irs}^{(\ell)} \times n_{irs} \right]}{\sum_{r=1}^{N_r} n_{irs}} .$$

Similarly, we estimated the sex-pooled, neighborhood-level rates, $\hat{\lambda}_{\cdot irs}^{(\ell)}$ as follows:

$$\hat{\lambda}_{i}^{(\ell)} = \frac{\sum_{s=1}^{N_s} \sum_{r=1}^{N_r} \left[\hat{\lambda}_{\cdot irs}^{(\ell)} \times n_{irs} \right]}{\sum_{s=1}^{N_s} \sum_{r=1}^{N_r} n_{irs}} .$$

These neighborhood rates are adjusted to the age distribution of the 2000 U.S. standard million population and reflect the race composition of each neighborhood.

Furthermore, we estimated the sex-pooled, city-level rates $\hat{\lambda}_{Philly}^{(\ell)}$ as follows:

$$\hat{\lambda}_{Philly}^{(\ell)} = \frac{\sum_{i=1}^{N_i} \sum_{s=1}^{N_s} \sum_{r=1}^{N_r} \left[\hat{\lambda}_{\cdot irs}^{(\ell)} \times n_{irs} \right]}{\sum_{i=1}^{N_i} \sum_{s=1}^{N_s} \sum_{r=1}^{N_r} n_{irs}} .$$

3.2. Cancer Screening and Risk Factor Prevalence Rates

PHMC provides survey weights in the SEPAHHS datasets which adjust the prevalence estimates from sampled respondents to represent the full Philadelphia population at the time of survey administration (i.e., the 2015 survey is adjusted to the 2015 population). In lieu of using the provided survey weights, we chose to use a modeling approach which considers key demographic and spatial variables that PHMC used when the weights were created. Using this modelling approach, we obtained smoothed estimates of prevalence for each measure by sex and race, adjusted by age, within each neighborhood. We chose to model at the neighborhood level rather than by census tracts due to the potential for highly unstable model estimates resulting from small sample sizes at the tract level. Neighborhoods were additionally aggregated across the city to derive city-wide estimates. These models use a similar approach as Quick 2020[7].

To calculate prevalence estimates, we first used Bayesian statistical modeling to derive prevalence estimates for each combination of race, sex, age, and neighborhood, using data pooled across the desired survey years. The model estimates were then aggregated over all races and age groups to create sex-specific prevalence estimates for each neighborhood and for the whole city. Estimates from all neighborhoods were age-standardized to the Philadelphia 2014-2018 ACS population to allow for meaningful comparisons across different neighborhood populations. The modeling approach is described in detail below.

Models were run separately for each measure. Each model included all survey respondents with non-missing values for the binary outcome $\boldsymbol{Y}$ (e.g. obesity, smoking, etc.) and the categorical model parameters of age group a $\{a=1,...,N_a\}$, race (white, Black, other) $\{r=1,...,N_r; N_r=3\}$., sex $s\{s=1,...,N_s; N_s=1 \text{ or } 2\}$, and neighborhood $i\{i=1,...,N_i\}$.

Let us assume that the outcome, y, for the j^{th} individual, given their specific neighborhood i, age group a, race r, sex s, has a Bernoulli distribution:

$$y_{j,iars} \sim Bernoulli\left(\theta_{iars}\right)$$

where θ_{iars} denotes the probability that outcome $y_{j,iars}=1$ and thus represents our estimate of the prevalence of the outcome.

To model θ_{iars}, we assume a logistic regression model of the form:

$$logit\left(P\left(y_{j,iars}=1 \,\middle|\, \boldsymbol{\beta}_0, \boldsymbol{z}\right)\right) = logit\,\theta_{iars} = \beta_{0;ars} + z_{irs},$$

where $\beta_{0;ars}$ is a intercept parameter that varies for each combination of age, race, and sex and z_{irs} is a random effect that varies by neighborhood, race, and sex.

We expect that the sample size used for estimating each group-specific parameter will be quite small, leading to potentially unstable model estimates. We address this by defining covariance structures which allow for group-specific estimates to borrow strength from each other.

We begin by modeling the intercepts, where we assume a multivariate normal distribution:

$$\boldsymbol{\beta}_{0;\bullet rs} \sim MVNorm\left(\boldsymbol{\mu}_{\bullet rs}, \boldsymbol{G}_{\beta_0;s}\right)$$

where $\beta_{0;\bullet rs} = \left(\beta_{0;1,rs}, \ldots, \beta_{0;N_a,gs}\right)$ denotes the vector of N_a age-specific intercept parameters for each combination of race and sex. The mean $\mu_{\bullet s} = \left(\mu_{1s}, \ldots, \mu_{N_a s}\right)$ with elements μ_{as} plays the role of non-race-specific general age/sex intercept estimates. We specified all μ_{as} to have an uninformative prior, μ_{as}~*Uniform(-5,5)* .

While we want to allow different age groups within each combination of race and sex to have different intercept estimates, we do believe that these values may be related to one another. As such, we define $\boldsymbol{\beta}_{0;\bullet rs}$ to have the age group covariance structure $\boldsymbol{G}_{\beta_0;s}$, which is a sex-specific N_a *x* N_a square matrix. The elements $g_{\beta_0;s;a,a'}$ correspond to the covariance between age groups *a* and a_1 for sex.

$$\boldsymbol{G}_{\beta_0;s;\bullet\bullet} = \begin{bmatrix} g_{\beta_0;s;1,1} & \cdots & g_{\beta_0;s;1,N_a} \\ \vdots & \ddots & \vdots \\ g_{\beta_0;s;N_a,1} & \cdots & g_{\beta_0;s;N_a,N_a} \end{bmatrix}$$

The sex-specific age-covariance structure $\boldsymbol{G}_{\beta_0;s}$ is not race/ethnicity-specific. All N_r race/ethnicity groups use the same estimated age covariance structures, therefore allowing groups to borrow information from each other in estimating the intercepts $\boldsymbol{\beta}_0$. The prior distribution of $\boldsymbol{G}_{\beta_0;s}$ is an inverse Wishart on the matrix $\boldsymbol{G}_{\beta_0;0}$, with degrees of freedom chosen to be $\nu = N_a$. The elements $g_{\beta_0;s;a,a'}$ of matrix $\boldsymbol{G}_{\beta_0;0}$ are chosen to reduce prior informativeness:

$$g_{\beta_0;0;a,a'} = \begin{cases} 0, & \text{if } a \neq a' \\ 2, & \text{if } a = a'. \end{cases}$$

While the intercepts $\boldsymbol{\beta}_0$ estimate the city-level effect of age, race, and sex on the probability of the outcome, we are interested in differences in outcome prevalence among the many neighborhoods of Philadelphia. The random effects $\boldsymbol{z}$ capture the neighborhood level effect of race and sex on the prevalence estimate.

We use a similar approach to estimate the random effects as we did for the intercepts. For each neighborhood, we define a vector of $N_r N_s$ random effects $\boldsymbol{z}_{i\bullet\bullet}$. The elements z_{irs} represent the random effect for each combination of neighborhood, race, and sex. The vector $\boldsymbol{z}_{i\bullet\bullet}$ is assumed to have multivariate normal distribution, centered around a vector of $N_r N_s$ zeros, with a covariance structure $\boldsymbol{G}_z$:

$$\boldsymbol{z}_{i\bullet\bullet} \sim MVN\left((0,\ldots,0), \boldsymbol{G}_z\right).$$

The covariance structure $\boldsymbol{G}_z$ is a $N_r N_s$ *x* $N_r N_s$ matrix, where elements $g_{z;rs,r's'}$ denote the covariance between the estimates for two groups with race and sex combinations *rs* and $r^1 s^1$. We assume $\boldsymbol{G}_z$ to have an inverse Wishart prior distribution on the scale matrix $\boldsymbol{G}_{z0}$ and degrees of freedom chosen to be $\nu = N_r N_s$. The elements $g_{z0;rs,r's'}$ of $\boldsymbol{G}_{z0}$ are chosen to be minimally informative:

$$g_{z0;rs,r's'} = \begin{cases} 0, & \text{if } r \neq r' \text{ or } s \neq s' \\ 1, & \text{if } r = r' \text{ and } s = s'. \end{cases}$$

Each model was run in WinBUGS for a minimum of 80,000 iterations, up to 100,000 iterations. Discarding the initial 10,000 iterations for a burn-in period and thinning by a factor of 10 yielded *L = 7,000 to 9,000* posterior estimates of $\boldsymbol{\beta}_0$ and $\boldsymbol{z}$. For the ℓ^{th} iteration, where ℓ =1,...,*L*, the estimate of the prevalence, $\hat{\theta}^{(\ell)}_{iars;}$, is computed from $\boldsymbol{\beta}_0^{(\ell)}$ and $\boldsymbol{z}^{(\ell)}$ using the following formula:

$$\hat{\theta}^{(\ell)}_{iars} = \frac{\exp\left(\beta^{(\ell)}_{0;ars} + z^{(\ell)}_{irs}\right)}{1 + \exp\left(\beta^{(\ell)}_{0;ars} + z^{(\ell)}_{irs}\right)}$$

Let n_{irs} be the population count in neighborhood i for age group a, race/ethnicity r, and sex s, as obtained from the ACS data. Neighborhood-level and city-level prevalence estimates were aggregated from prevalence estimates $\hat{\theta}^{(\ell)}_{iars}$ using population-based weighting as shown in the formulas below.

Philadelphia city-level age group population proportions within sex:

$$prop_{\text{age;Philly},as} = \frac{n_{\text{Philly},as}}{n_{\text{Philly},s}} = \frac{\sum_{i=1}^{N_i}\sum_{r=1}^{N_r} n_{iars}}{\sum_{a=1}^{N_a}\sum_{i=1}^{N_i}\sum_{r=1}^{N_r} n_{iars}}$$

Neighborhood-level age-adjusted age/sex/race population counts:

$$n_{\text{adjusted};iars} = n_{isr} \times prop_{\text{age;Philly},as} = \frac{\left(\sum_{a=1}^{N_a} n_{iars}\right) \times \left(\sum_{i=1}^{N_i}\sum_{r=1}^{N_r} n_{iars}\right)}{\left(\sum_{a=1}^{N_a}\sum_{i=1}^{N_i}\sum_{r=1}^{N_r} n_{iars}\right)}$$

Neighborhood-level sex-specific age-adjusted prevalence estimates:

$$\hat{\theta}^{(\ell)}_{is} = \frac{\sum_{a=1}^{N_a}\sum_{r=1}^{N_r}\left[\hat{\theta}^{(\ell)}_{iars} \times n_{\text{adjusted};iars}\right]}{\sum_{a=1}^{N_a}\sum_{r=1}^{N_r} n_{iars}}$$

Neighborhood-level combined-sex (males and females) age-adjusted prevalence estimates:

$$\hat{\theta}^{(\ell)}_{i} = \frac{\sum_{s=1}^{N_s}\sum_{a=1}^{N_a}\sum_{r=1}^{N_r}\left[\hat{\theta}^{(\ell)}_{iars} \times n_{\text{adjusted};iars}\right]}{\sum_{s=1}^{N_s}\sum_{a=1}^{N_a}\sum_{r=1}^{N_r} n_{iars}}$$

These neighborhood prevalences are adjusted to the age distribution of Philadelphia using the American Community Survey (ACS) 2014-2018 5-year aggregate population estimates and reflect the race composition of each neighborhood.

Philadelphia-level sex-specific prevalence estimates:

$$\hat{\theta}^{(\ell)}_{\text{Philly},s} = \frac{\sum_{i=1}^{N_i}\sum_{a=1}^{N_a}\sum_{r=1}^{N_r}\left[\hat{\theta}^{(\ell)}_{iars} \times n_{iars}\right]}{\sum_{i=1}^{N_i}\sum_{a=1}^{N_a}\sum_{r=1}^{N_r} n_{iars}}$$

Philadelphia-level combined-sex (males and females) prevalence estimates:

$$\hat{\theta}^{(\ell)}_{\text{Philly}} = \frac{\sum_{i=1}^{N_i}\sum_{a=1}^{N_a}\sum_{r=1}^{N_r}\sum_{s=1}^{N_s}\left[\hat{\theta}^{(\ell)}_{iars} \times n_{iars}\right]}{\sum_{i=1}^{N_i}\sum_{a=1}^{N_a}\sum_{r=1}^{N_r}\sum_{s=1}^{N_s} n_{iars}}$$

For each outcome prevalence estimate $\hat{\theta}$, the point estimate and lower and upper bounds of the 95% credible interval were determined by taking the 50th, 2.5th, and 97.5th percentiles, respectively, of the estimates from all L posterior samples.

These city prevalence rates reflect the sex and race distribution of Philadelphia using the American Community Survey (ACS) 2014-2018 5-year aggregate population estimates.

4. Significant Difference from City Level Estimate

We deemed the estimates of incidence, mortality, screening, or prevalence for a neighborhood significantly different from the city if the median city-level estimate is not contained in the neighborhood estimate's 95% credible interval. For example, if $\hat{\lambda}_i^{(\ell)}$ represents the neighborhood-level posterior samples and $\hat{\lambda}_{Philly}^{(\ell)}$ represents the city-level posterior samples, we counted the number of samples in $\hat{\lambda}_i^{(\ell)}$ that are greater than median of $\hat{\lambda}_{Philly}^{(\ell)}$ and then divided by the total number of samples (or iterations in MCMC sampling) L, where l =1,..., L. If this quantity is greater than 0.975 or less than 0.025, we call the estimate for this neighborhood significantly different than the city.

It should be noted that making comparisons to the median of the city-level estimates effectively assumes that there is no uncertainty in our city-level estimates

5. Data Suppression

Due to privacy concerns and the reliability of rates based on small counts, we chose to present model-based estimates rather than estimate crude rates. Incidence and mortality estimates are considered reliable if the posterior median $\hat{\lambda}$ is greater than the width of the 95% credible interval. Risk factor and cancer screening estimates are considered reliable if both the posterior median $\hat{\theta}$ and 1- $\hat{\theta}$ are greater than the width of the 95% credible interval. Estimates that are not considered reliable are suppressed.

6. Notes of Caution in Interpretation

The model-based incidence and mortality rates in this report are adjusted to the 2000 standard million population and may not be comparable to the rates obtained by standardizing to another standard population or to those obtained by different models. Cancer screening and risk factor prevalence rates are age-standardized to the Philadelphia ACS 2014-2018 population.

Cancer screening and risk factors are self-reported which may be less accurate than physician diagnoses or objective measurements. It has been found that survey respondents tend to underreport weight [9], alcohol intake[10], and tobacco use[11], and overreport physical activity and they may not be aware of underlying health conditions.[12]

References

1. Curtin, L.R., & Klein, R.J. , Direct standardization (age-adjusted death rates). 1995, US Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Center for Health Statistics: Hyattsville, MD.
2. Health., P.D.o.P. Close to Home: The Health of Philadelphia's Neighborhoods. 2019 8/29/2019 [cited 2020 July 1, 2020]; Available from: https://www.phila.gov/media/20190801133844/Neighborhood-Rankings_7_31_19.pdf.
3. PA Department of Health. Table 1: ICD-O-3 codes used to determine the primary site of cancer incidence. August 2019]; Available from: https://www.health.pa.gov/topics/HealthStatistics/EDDIE/Documents/Cancer_County_State_(ICD-O_Codes).pd.
4. Public Health Management Corporation. Public Health Management Corporation Community Health Data Base's Southeastern Pennsylvania Household Health Survey. 2018 01 September 2019]; Available from: http://www.chdbdata.org/.
5. Lin, B.-H., Reed, Jane, Lucier, Gary. U.S. Fruit and Vegetable Consumption. 2004 [cited 2019; 792-2:[Available from: https://www.ers.usda.gov/webdocs/publications/42566/15230_aib792-2_1_.pdf?v=0.
6. Bureau., U.S.C. Census 2000 Summary File 1 and Summary File 3 - United States. 2001 September 01, 2019]; Available from: https://www.census.gov/main/www/cen2000.html.
7. Quick, H., et al., Trends in Tract-Level Prevalence of Obesity in Philadelphia by Race-Ethnicity, Space, and Time. Epidemiology, 2020. 31(1): p. 15-21.
8. Sturtz, S., Ligges, U. & Gelman, A.E.,, R2WinBUGS: a package for running WinBUGS from R. 2005.
9. Wen M, K.-J.L., Sex and ethnic differences in validity of self-reported adult height, weight and body mass index. Ethn Dis, 2012. 22: p. 72-78.
10. Feunekes GI, w.t.V.P., van Staveren WA, Kok FJ., Alcohol intake assessment: the sober facts. Am J Epidemiol, 1999. 150: p. 105-112.
11. Klein JD, T.R., Sutter EJ., Self-reported smoking in online surveys: prevalence estimate validity and item format effects. Med Care, 2007. 45: p. 691-695.
12. Pierannunzi C, H.S., Balluz L, A systemic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS), 2004-2011. BMC Med Res Methodol, 2013. 13: p. 49.

DREXEL UNIVERSITY
Urban Health Collaborative
Dornsife School of Public Health

Urban Health Collaborative
Dornsife School of Public Health, Drexel University
3600 Market Street, 7th Floor
Philadelphia, PA 19104
uhc@drexel.edu