Annals of Epidemiology 73 (2022) 9–16

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

# A Flexible Method for Identifying Spatial Clusters of Breast Cancer Using Individual-Level Data

Maria E. Kamenetsky, PhD[a], Amy Trentham-Dietz, PhD[a,b], Polly Newcomb, PhD[c], Jun Zhu, PhD[d], Ronald E. Gangnon, PhD[a,b,e,*]

[a] Department of Population Health Sciences, University of Wisconsin-Madison, United States
[b] Carbone Cancer Center, School of Medicine and Public Health, University of Wisconsin-Madison, United States
[c] Fred Hutchinson Cancer Research Center, United States
[d] Department of Statistics, University of Wisconsin-Madison, United States
[e] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, United States

ARTICLE INFO

*Article history:*
Received 4 February 2022
Revised 16 May 2022
Accepted 10 June 2022
Available online 27 June 2022

*Keywords:*
spatial cluster detection
spatial cluster
spatial epidemiology
breast cancer
Lasso
case-control

ABSTRACT

Prior research has shown that cancer risk varies by geography, but scan statistics methods for identifying cancer clusters in case-control studies have been limited in their ability to identify multiple clusters and adjust for participant-level risk factors. We develop a method to identify geographic patterns of breast cancer odds using the Wisconsin Women's Health Study, a series of 5 population-based case-control studies of female Wisconsin residents aged 20-79 enrolled in 1988-2004 (cases=16,076, controls=16,795). We create sets of potential clusters by overlaying a 1 km grid over each county-neighborhood and enumerating a series of overlapping circles. Using a two-step approach, we fit a penalized binomial regression model to the number of cases and trials in each grid cell, penalizing all potential clusters by the least absolute shrinkage and selection operator (Lasso). We use BIC to select the number of clusters, which are included in a participant-level logistic regression model. We identify 15 geographic clusters, resulting in 23 areas of unique geographic odds ratios. After adjustment for known risk factors, confidence intervals narrowed but breast cancer odds ratios did not meaningfully change; one additional hotspot was identified. By considering multiple overlapping spatial clusters simultaneously, we discern gradients of spatial odds across Wisconsin.



## Introduction

Breast cancer is a multi-faceted disease and spatial surveillance of breast cancer diagnoses can be used to target screening programs in geographic areas with high cancer burdens. Spatial and temporal patterns can be used to identify differences in geographic risk [15,18,25,33,34,46], and these differences may be driven by many factors including environmental risk factors [10,27,32,41] or structural access to health care [3,4,7,8,12,17,19,22,28]. Identifying geographic areas of elevated breast cancer risk remains of interest to researchers as spatial cluster detection methods continue to develop and our understanding of risk factors and cancer etiology expands. Previous studies found clusters of breast cancer in Cape Cod, Massachusetts [34] and Marin County, California [9,48], whereas in Wisconsin one study suggested breast cancer risk was elevated in the North Shore Milwaukee area [39].

Due to privacy concerns and data availability, many health data are aggregated to a polygon leading to a loss in location precision. Kernel-based methods have estimated the risk surface, but have not adjusted for additional covariates [21]. Studies that have used scanning windows such as the spatial scan [23] or spline-based approaches in case-control studies have either been performed across relatively small study regions[15,33,36,46,47], have separately reported results from sub-regions [38], or have been limited in the number of cases [1]. In this study, we investigate patterns of geographic odds of breast cancer in Wisconsin in the period between 1988-2004 using data from the Wisconsin Women's Health Study

*Abbreviations:* BIC, Bayesian information criterion; BMI, Body mass index ($kg/m^2$); CI, Confidence interval; km, Kilometers; Lasso, least angle shrinkage and selection operator; SD, Standard deviation; WWHS, Wisconsin Women's Health Study; WTM, Wisconsin Transverse Mercator.

* Corresponding author.
*E-mail address:* ronald@biostat.wisc.edu (R.E. Gangnon).

https://doi.org/10.1016/j.annepidem.2022.06.006
1047-2797/

(WWHS). We use a large case-control study across the state of Wisconsin to identify spatial clusters and estimate the geographic risk surface. Other spatial scan methods have used scanning windows to create sets of potential clusters, where the most likely cluster is assessed using Monte Carlo-based hypothesis testing. Where approaches based on SaTScan [24] and FlexScan [43] require sequential deletion to identify multiple clusters, our method obtains the coefficient paths of all potential clusters and selects the final number of clusters using information criteria. Our use of the Lasso is computationally-efficient and identifies distinct clusters with geographic odds ratios different from the background odds of breast cancer, even after adjusting for known participant-level covariates.

## Methods

### *Study Population*

We analyzed data from participants in the Wisconsin Women's Health Study (WWHS). WWHS is comprised of five case-control studies of breast cancer in Wisconsin women aged 20-79 years old spanning 1988-2004. WWHS case women were identified from the Wisconsin mandatory cancer registry as having first invasive breast cancer. Controls were randomly selected from driver's license lists (<65 years old) and Medicare beneficiary files (65-79 years old), and age-matched to cases in 5-year age groups. The five study time periods differed slightly in their age eligibility criteria due to the primary scientific aims of each wave of data collection: 1988-1991 (study participants between ages 20-74 years old), 1992-1995 (ages 50-79 years old), 1997-2000 and 2001-2004 (ages 20-69 years old). Telephone interviews were conducted from September 1988 through May 2004 and collected information on breast cancer including reproductive histories, alcohol and tobacco use, family history of breast cancer, and contraceptive and post-menopausal hormone use for all study participants. The WWHS full sample collection has been described elsewhere [29,30].

In our sample, there were 16,076 eligible case women (response rate: 85% [30]) and 16,795 eligible control women (response rate: 87%). After excluding 47 women with no geographic coordinates nor identified county, there were 16,075 case and 16,749 control women in the final analytic sample. Available data for each woman included age, height, weight, body mass index ($kg/m^2$, BMI), age at menopause, parity (number of full-term pregnancies, continuous), drinks of alcohol per week, age at first full-term birth, race, education level, family history of breast cancer, and post-menopausal hormone use. Age was taken at time of diagnosis for cases and at a reference age for controls, defined as age at time of interview minus the average time from diagnosis to interview among similarly-aged cases. All other variables were ascertained for the reference age. There was some missing data, with the largest percent missing (14.4%) for age at menopause.

To account for missing data, multiple imputation using predictive mean matching using the `Hmisc` R package [16] was used. We used 14 multiply-imputed datasets for the analysis based on the fraction of observations with missing data. The imputation model included breast cancer status, coordinates, age, race, education, age at menopause, menopause status, parity, BMI, drinks per week, age at first birth, as well as breast cancer stage at diagnosis (local, regional, distant), height, weight, and post-menopausal hormone use.

### *Geocoding*

Study participant addresses were geocoded across Wisconsin based on participant mailing address at time of interview. A five-step geocoding strategy achieved a 97% match rate [30]. Participant latitude and longitude coordinates were projected into easting and northing coordinates using the Wisconsin Transverse Mercator (WTM) projection to facilitate Euclidean distance calculations[11].

The Wisconsin Department of Health Services has identified counties along the Minnesota-Wisconsin border (Barron, Bayfield, Buffalo, Burnett, Dunn, Eau Claire, Pepin, Pierce, Polk, and St. Croix) as having under-reported case counts along the border [13], likely due to treatment-seeking in Minnesota.

### *Statistical Analyses*

#### *Neighborhood-Level Spatial Cluster Model*

For each of the 72 counties in Wisconsin, we created a neighborhood based on counties that share a common border of any length in order to take into consideration border effects. Each county-neighborhood was divided into small grid cells created by overlaying a high-resolution 1 km grid. Cases and controls were assigned to their respective grid cells.

We considered potential spatial clusters as moving circular windows centered at cell centroids centered inside the focal county [23,26]. Potential clusters were constructed using ordered Euclidean distances from each grid cell centroid up to a maximum radius, $r_{max}$. We set $r_{max}$ to 10 km, which reflected the median great-circle distance between home and work for Wisconsin residents [45]. Figure 1 demonstrates the creation of neighborhood-counties and circular potential clusters for a single grid cell using Milwaukee county. We obtained the full set of potential clusters by enumerating over all combinations of circular windows from 0 to $r_{max}$, allowing for potential clusters to overlap spatially. Grid cells in each county-neighborhood can be considered members of a potential cluster if the focal cell of the potential cluster remains inside the county of interest. For a single county-neighborhood $b$ in $b = 1, \ldots, 72$ county-neighborhoods, let $t_{bg}$ be the sum of cases and controls in each 1 km grid cell in and $p_{bg}$ be the probability of being a case in the $g$th grid cell centered at the $b$th county centroid. Then

$$Y_{bg} \sim \text{Binomial}(t_{bg}, p_{bg}) \tag{1}$$

where $Y_{bg}$ is the number of cases in grid cell $g$ in neighborhood $b$ and

$$\log\left(\frac{p_{bg}}{1-p_{bg}}\right) = \alpha + \sum_{j=1}^{K_b} \theta_j \mathbb{1}\{d(z_{bg}, c_j) \le r_j\}, \tag{2}$$

where $\alpha$ is the background geographic odds of breast cancer in the county-neighborhood for the grid cells not belonging to any active clusters. The spatial clustering component is $\sum_{j=1}^{K_b} \theta_j \mathbb{1}\{d(z_j, c_j) \le r_j\}$, where $K_b$ is the total number of potential spatial clusters in county-neighborhood $b$, $\theta_j$ is the log geographic odds ratio inside cluster $j$, and $\mathbb{1}\{\cdot\}$ is the indicator function that takes 1 if the Euclidean distance $d(\cdot)$ between a cell with center $z_{bg}$ and cluster centered at $c_j$ is less than or equal to radius $r_j$, and is 0 otherwise. For simplicity of notation, we let $x_{gj} = \mathbb{1}\{d(z_{bg}, c_j) \le r_j\}$.

We used regularization based on the Lasso penalty to identify spatial cluster(s) in each county-neighborhood[20]. To select the clusters in the study region, we minimized the following penalized loglikelihood function:

$$f(\alpha, \boldsymbol{\theta}) = -\ell(\alpha, \boldsymbol{\theta}) + \lambda \sum_{j=1}^{K_b} |\theta_j| \tag{3}$$

where $\ell(\alpha, \boldsymbol{\theta})$ is the binomial loglikelihood function and $\lambda$ is a tuning parameter which controls the amount of shrinkage and goes from 1 to $L$, where $\lambda_1, \ldots, \lambda_L$ are monotonically decreasing. The Lasso regularization procedure begins with the null model with

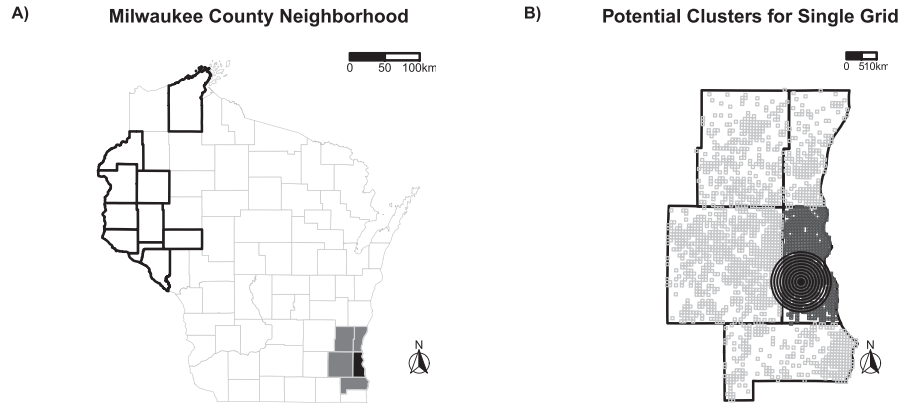


**Fig. 1.** (A) The Milwaukee county-neighborhood contains Ozaukee, Washington, Waukesha, and Racine counties. (B) In the Milwaukee county-neighborhood, there were a total of 2170 grid cells (541 in the focal county of Milwaukee). Grid cells with zero cases and controls were omitted. There were a total of 123,977 circular potential clusters centered at focal grid cells (black cells) inside Milwaukee county. Cells in the county-neighborhood (grey cells) can be members of a potential cluster only if the focal cell is inside Milwaukee county.

no clusters. Variable selection is performed by shrinking $\theta_j$'s to zero as they reach the penalty and are dropped from the active set, which is the set of clusters allowed into the model at $\lambda_l$. This results in coefficient paths for each $\theta_j$ cluster over $\lambda_l$ tuning parameters. As $\lambda$ gets smaller, more clusters are allowed to enter the model. This procedure was repeated for each county-neighborhood in the analysis, giving a total of $K$ potential clusters across all 72 county-neighborhoods.

*Selection of Number of Clusters*

We evaluated model fit using an information-theoretic approach. At each $\lambda_l$, we counted the number of parameters in the active set by the number of predictors in the model. The criterion we used to identify the number of clusters in the study region is Bayesian information criterion (BIC) [42], defined as:

$$\mathrm{BIC}(k,\lambda) = -2\ell(\hat{\alpha},\hat{\boldsymbol{\theta}};\lambda) + (k_\lambda + 1)\log(n^*), \tag{4}$$

where $\ell(\hat{\alpha},\hat{\boldsymbol{\theta}})$ is the loglikelihood based on the active set for a given $\lambda$, and $k_\lambda$ is the number of clusters selected by the model in the active set at each $\lambda$, and $n^*$ is the effective sample size (the smaller of the number of cases and the number of controls). Previous work has shown using simulation studies that BIC better maintains the false positive rate near 0% under the null when no clusters are present in the study region [20]. We selected either 0 or $k$ cluster(s) in each county-neighborhood, with geographic odds ratios that differ from the estimated background breast cancer rate.

*Participant-Level Spatial Cluster Model*

For each participant in the study, we defined a case-control identifier, $Y_i$, to be:

$$Y_i = \begin{cases} 1, & \text{if } i\text{th participant is a breast cancer case} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Let $p_i$ be the probability of the $i$th participant being a case. We used logistic regression to model the log odds of being a case as a function of the identified clusters as well as additional covariates:

$$\log(p_i/\{1-p_i\}) = \alpha + \sum_{j=1}^{k}\theta_j x_{ij} + \sum_{p=1}^{P}\beta_p u_{ip}, \tag{6}$$

where $\alpha$ is the intercept; $\sum_{j=1}^{k}\theta_j x_{ij}$ is the spatial clustering component, $k$ are the selected spatial clusters across the county-neighborhoods. Participant $i$ can be a member of multiple overlapping clusters or no clusters; $u_{ip}$ are $P$ covariates for each participant $i$, associated with regression parameters $\beta_p$, which can be known or unknown.

Multivariable logistic regression was used to model the probability of being a breast cancer case. The set of identified clusters from the cluster identification step were included in the participant-level spatial cluster models as indicator variables. Age-adjusted and fully-adjusted regression models were fitted separately to the 14 imputed data sets. The fully-adjusted model included age, family history of breast cancer, parity, BMI, drinks of alcohol per week, age at first birth, age at menopause, education, and race. Age, age at first birth, age at menopause, and BMI were centered and scaled. Prior literature has demonstrated how racist practices are associated with place of residence as well as disparities in breast cancer incidence [3] due to discrimination [2,31,44]. We capture such racism-driven social determinants of health by including race as a proxy, which is associated with place of residence and breast cancer risk and not on the causal pathway.

After adjustments for missing data via multiple imputation, point estimates, standard errors, and geographic odds ratio estimates were pooled for statistical inference using Rubin's method [40]. Rubin's method pools regression coefficients and standard errors across models performed on each of the imputed datasets and considers within and between imputation variance to derive confidence intervals. Analyses were performed using the `clusso`[20], `sf` [35], `sp`[6], `spdep`[6], and `rgeos`[5] packages in the `R` statistical software[37].

## Results

*Descriptive Results*

The mean age of women in the study was 57.7 (SD: 10.5) for cases and 57.2 (SD: 10.3) for controls and 80% of cases and 90% of controls had no family history of breast cancer. For women who had a full-term pregnancy, on average cases had 3.0 children (SD: 1.6) and were 24.2 years old at first birth (SD: 4.6), while controls had on average 3.2 children (SD: 1.5) and were 23.5 years old at first birth (SD: 4.3). Most case and control women were white (96%, 95%), and the largest education group had achieved at least a grade 12 education in both cases (43%) and controls (43%). Descriptive statistics for the analytical sample can be found in Table 1. Figure 2 shows the county of residence for cases (left) and controls (right).

**Table 1**
Baseline Characteristics of the Analyzed Participants Wisconsin Women's Health Study (n=32,824).

| All Women | | Controls | Cases | Missing % |
|---|---|---|---|---|
| N (100%) | | 16749 | 16075 | |
| Age (years) (mean (SD)) | | 57.16 (10.25) | 57.68 (10.49) | 0.0 |
| BMI (kg/$m^2$) (mean (SD)) | | 26.25 (5.48) | 26.38 (5.43) | 2.2 |
| Height (m) (mean (SD)) | | 1.64 (0.06) | 1.64 (0.06) | 1.3 |
| Weight (kg) (mean (SD)) | | 70.50 (15.44) | 71.31 (15.13) | 1.5 |
| Stage (%) | Local | 0 (0.0) | 10107 (66.6) | |
| | Regional | 0 (0.0) | 4706 (31.0) | |
| | Distant | 0 (0.0) | 354 (2.3) | |
| Race (%) | White | 15,981 (96.5) | 15,463 (97.3) | |
| | Black | 297 (1.8) | 219 (1.4) | 1.2 |
| | Hispanic | 110 (0.7) | 79 (0.5) | |
| | Other | 170 (1.0) | 125 (0.8) | |
| Education (%) | < High School | 1,852 (11.2) | 1,683 (10.5) | 1.0 |
| | High School | 7,272 (43.9) | 6,966 (43.8) | |
| | Some College 1-3 | 4,204 (25.4) | 3,786 (23.8) | |
| | Bachelor's | 2,254 (13.6) | 2,371 (14.9) | |
| | Graduate Degree | 993 (6.0) | 1,105 (6.9) | |
| **Post-Menopausal Women** | | **Controls** | **Cases** | **Missing** |
| N (67.91%) | | 11,423 | 10,868 | |
| Age at Menopause (years) (mean (SD)) | | 47.52 (6.75) | 48.38 (6.26) | 14.4 |
| Post-Menopausal Hormone Use (%) | Never | 6,343 (60.9) | 5,796 (58.6) | |
| | Former | 1,168 (11.2) | 1,144 (11.6) | |
| | Current | 2912 (27.9) | 2949 (29.8) | 8.9 |
| **Gave Birth** | | **Controls** | **Cases** | **Missing** |
| N (87.22%) | | 14,769 | 13,861 | |
| Parity (mean (SD)) | | 3.20 (1.73) | 2.99 (1.60) | 0.0 |
| Age at First Birth (years) (mean (SD)) | | 23.50 (4.28) | 24.15 (4.55) | 0.3 |
| **Self-Reported Drinkers** | | **Controls** | **Cases** | **Missing** |
| N (80.19%) | | 13,325 | 12,995 | |
| Drinks per Week (mean (SD)) | | 3.24 (6.93) | 3.66 (6.37) | 0.0 |

Abbreviations: SD, standard devation; kg, kilograms, $m^2$, meters squared.

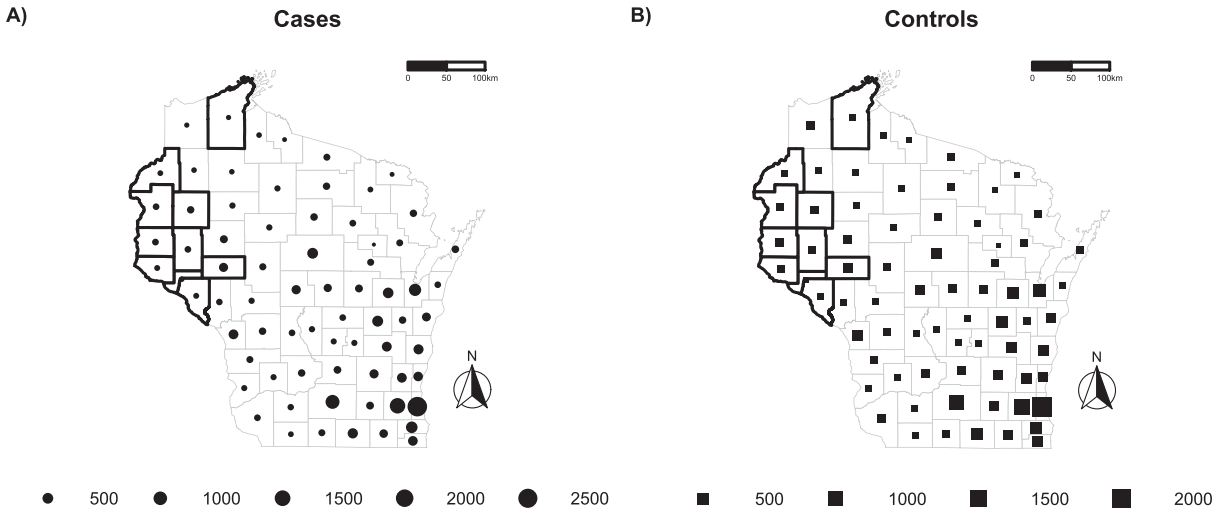


**Fig. 2.** Breast cancer cases (A) and controls (B) with circles/squares centered at each county centroid and are proportional to the number of cases or controls. Black outlines counties with under-reported case counts.

*Cluster Identification*

Ten of the 72 county-neighborhoods in Wisconsin were identified as having one or more clusters, with a total of 15 unique clusters across the state (Figure 3). Eight of the 15 clusters were identified either in or next to counties likely to have under-reported counts. Three clusters were identified across Grant and Lafayette counties which are on the border with Minnesota as well as Iowa.

There were four additional counties of elevated odds identified: one in Marathon county and three across Waukesha and Milwaukee counties.

*Cluster Estimation*

After adjusting for age, two geographic clusters with elevated breast cancer odds above the state background rate were identified in Marathon and Milwaukee county (Table 2). The Milwaukee county area was represented by three clusters, all with a 9.8 km radius, which produced seven areas of unique geographic odds ratios: one in Greenfield (OR = 1.17, 95% CI 1.06-1.30) and the remaining in West Allis. In West Allis, the geographic odds ratio in three of the areas were different from the background state-level odds of breast cancer with no clusters in the region (OR = 1.20,

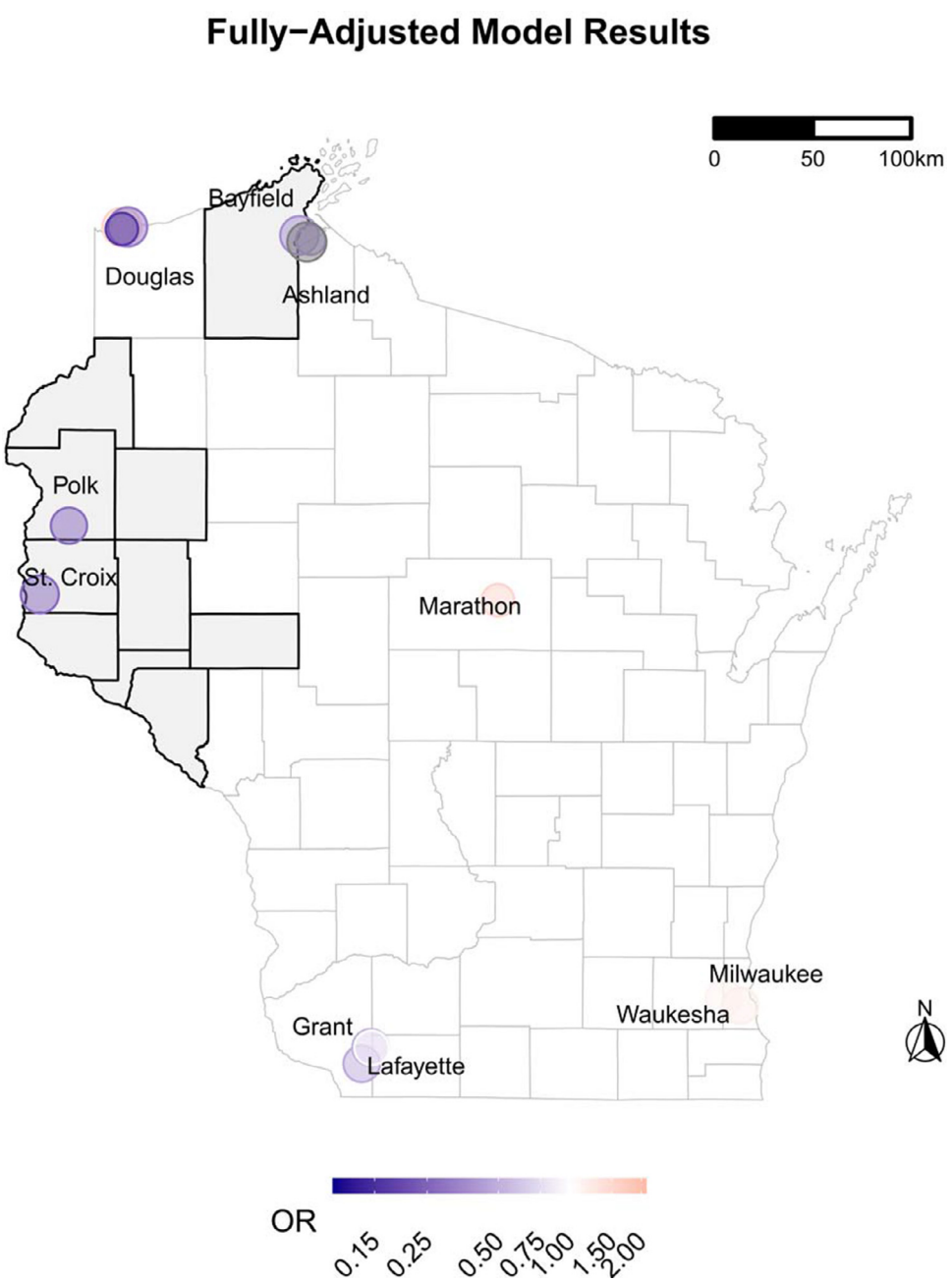


**Fig. 3.** Clusters identified by BIC. Under-reporting counties are indicated in grey. Circular clusters are allowed to overlap.

95% CI= 1.09-1.32; OR = 1.32, 95% CI = 1.00-1.75; OR = 1.41, 95% CI 1.25-1.59).

There were 11 clusters with reduced breast cancer odds identified with cluster radii ranging from 7.6 km to 10 km. In Douglas, Ashland, Grant, and Lafayette counties, there was a single area with a relatively large number of cases and controls that resulted in reduced geographic odds ratios different from the background: Superior (OR = 0.08, 95% CI = 0.04-0.15), Ashland (OR = 0.32, 95% CI = 0.17-0.61, and Platteville (OR = 0.25, 95% CI = 0.12-0.51)). Across Polk and St. Croix counties, there were two areas (Alden OR = 0.29, 95% CI=0.14-0.61 and Hudson OR = 0.28, 95% CI=0.17-0.47) with reduced geographic odds ratios.

All of the geographic areas identified as having risk of breast cancer different from the background level of risk after adjustment for age persisted after adjustment for other known risk factors. After full adjustment, the cluster estimates and confidence intervals of elevated geographic odds ratio in Weston and in Greenfield did not meaningfully change; the geographic odds ratio in one area of West Allis increased negligibly from 1.32 (95% CI = 1.00-1.75) to 1.35 (95% CI = 1.02-1.79) under the fully-adjusted model.

While confidence intervals either widened or remained the same for most areas of reduced geographic odds ratios, in some areas they narrowed slightly. These included Hudson (OR=0.27, 95% CI = 0.16-0.44), Alden (OR = 0.27, 95% CI = 0.13-0.57), Smelser (OR = 0.49, 95% CI = 0.19-1.22), and one area in Platteville (OR = 0.51, 95% CI = 0.06-4.56) (Table 2).

We mapped the geographic odds ratio surface and calculated 95% confidence bounds for estimates of the geographic odds ratios, holding all other covariates equal. Figure 4 shows this surface across Wisconsin using control women with random noise added to their locations. The two center columns identify the counties with clusters. The two point plots map the unique geographic odds

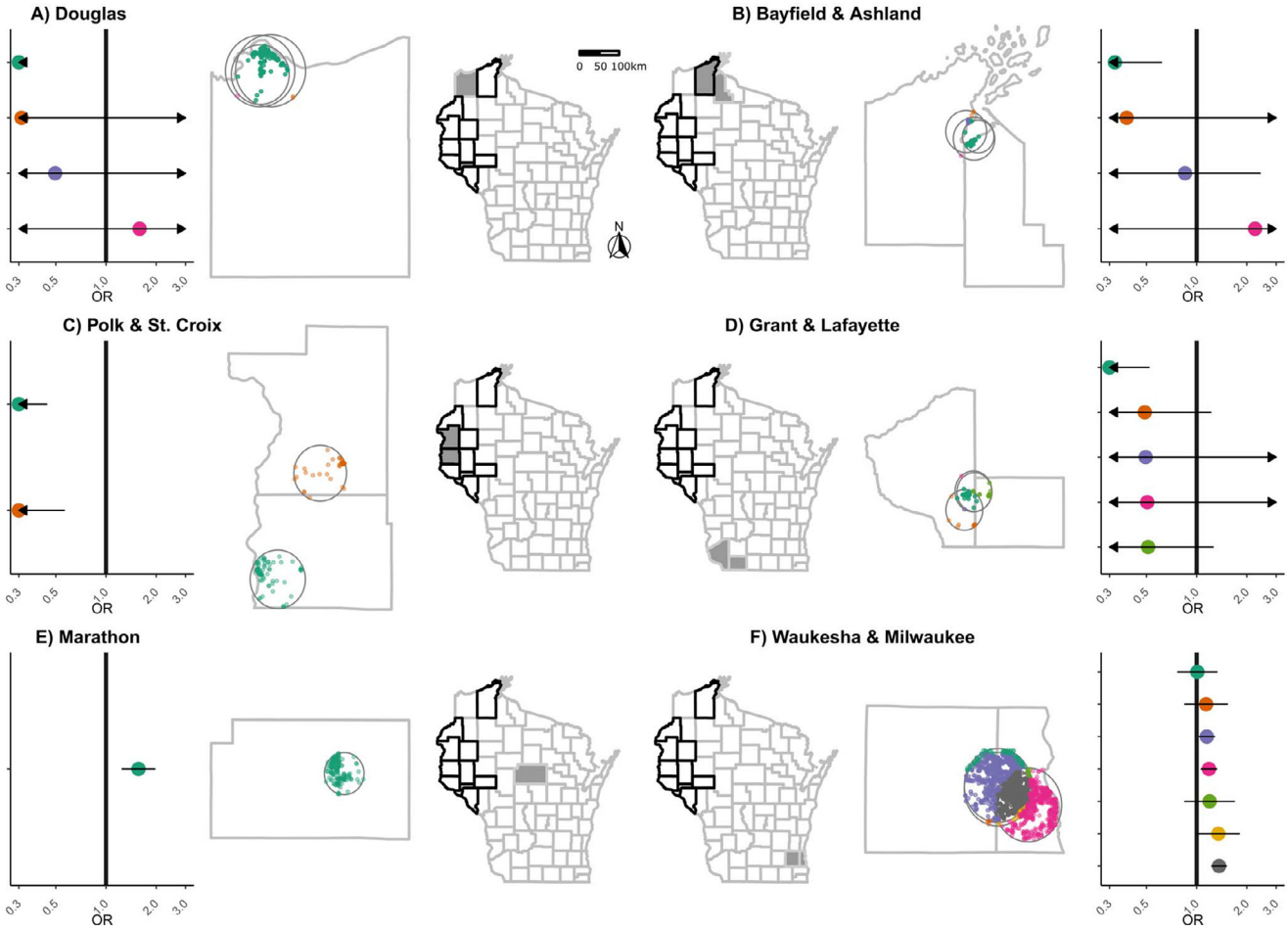


**Fig. 4.** Each panel (A-F) consists of (left) fitted odds ratio (OR) estimates and 95% CI for control women representing odds ratios in counties where clusters (in black circles) were identified. Arrows on confidence intervals indicate upper bounds greater than 3 or lower bounds less than 0.3 (see Table 2 for estimates); (center) control women encoded by the unique odds ratio estimates mapped to each county or set of counties; (right) map of Wisconsin identifying the set of counties in each panel in grey. Black outline indicates counties with under-reported case counts.

ratios. For example in the bottom right corner (Waukesha and Milwaukee counties), there were three total clusters identified. However with the overlap of these clusters, there were seven unique areas of geographic odds ratios. In the outermost panel, the unique geographic odds ratios and their respective 95% confidence bounds identify the areas in color on their respective maps.

## Discussion

In this study, we developed an approach to smooth participant-level geographic odds of breast cancer and identify areas that differ from the background odds of breast cancer across the state using a large case-control study. Weston (Marathon county) and areas of Greenfield and West Allis (Waukesha and Milwaukee counties) were identified as having elevated geographic odds ratios of breast cancer, which persisted even after adjustment for multiple established risk factors. Our results are consistent with previous literature in identifying areas of elevated risk near Milwaukee [4,14,39]. The highest odds of breast cancer identified in West Allis (1.41, 95% CI=1.25-1.59) are consistent with other cluster investigation studies of breast cancer that found odds ratios of 1.32-1.55 in Cape Cod, Massachusetts [34]. Using our approach, we also identified a new area of elevated geographic odds ratio in Marathon county that had previously not been identified.

The three clusters of reduced odds identified in Ashland and Bayfield counties are near Lake Superior. The clusters identified in Douglas county, though not a county with under-reported case counts, are identified because Superior is adjacent to Duluth, Minnesota and patients likely seek treatment at major hospitals and health care systems that serve Northern Minnesota and Northern Wisconsin. The two clusters of reduced odds identified across Grant and Lafayette counties border Minnesota and Iowa are likely identified due to treatment-seeking across states.

This study of spatial clustering is unique in using a large case-control study to identify spatial clusters across a broad geographic study region while adjusting for known risk factors at the participant level. By using the Lasso, we do not make assumptions on the size nor locations of the clusters and estimate the geographic odds inside the identified clusters. Such maps are informative to public health officials in identifying areas with opportunity for intervention. The elevated geographic odds ratios may be driven by increased need for access to care especially in more segregated areas of Milwaukee county, communities that are more homogeneous in certain risk factors such as areas with large Ashkenazi Jewish populations, or exposure to external risk factors not considered.

We chose covariates to include in the regression models *a priori*, which included a combination of reproductive and socioeconomic factors as well as personal behaviors. We did not adjust for post-menopausal hormone use and found that including post-menopausal hormone use did not affect the clusters identified nor the point estimates.

**Table 2**
Unique Geographic Odds Ratio Estimates Wisconsin, 1997-2000

| | Age-Adjusted | | Fully-Adjusted | | | |
|---|---|---|---|---|---|---|
| City | OR | (95% CI) | OR | (95% CI) | Cases | Controls |
| Douglas County | | | | | | |
| **Superior** | 0.08 | (0.04, 0.15) | 0.08 | (0.04, 0.15) | 33 | 423 |
| Superior | 0.34 | (0.02, 5.58) | 0.50 | (0.03, 7.59) | 2 | 2 |
| Superior | 0.35 | (0.02, 5.75) | 0.31 | (0.02, 6.06) | 0 | 2 |
| Superior | 0.97 | (0.05, 20.93) | 1.59 | (0.07, 36.56) | 0 | 1 |
| Ashland & Bayfield Counties | | | | | | |
| Ashland | 0.32 | (0.03, 3.44) | 0.38 | (0.03, 4.19) | 0 | 3 |
| **Ashland** | 0.32 | (0.17, 0.61) | 0.32 | (0.17, 0.63) | 39 | 126 |
| Ashland | 0.77 | (0.27, 2.18) | 0.85 | (0.3, 2.42) | 14 | 16 |
| Ashland | 2.39 | (0.2, 28.15) | 2.24 | (0.19, 26.78) | 0 | 1 |
| Polk & St. Croix Counties | | | | | | |
| **Hudson** | 0.28 | (0.17, 0.47) | 0.27 | (0.16, 0.44) | 21 | 79 |
| **Alden** | 0.29 | (0.14, 0.61) | 0.27 | (0.13, 0.57) | 10 | 36 |
| Grant & Lafayette Counties | | | | | | |
| **Platteville** | 0.25 | (0.12, 0.51) | 0.25 | (0.12, 0.52) | 27 | 93 |
| Smelser | 0.43 | (0.05, 3.84) | 0.49 | (0.05, 4.63) | 0 | 2 |
| Platteville | 0.48 | (0.2, 1.16) | 0.51 | (0.21, 1.25) | 6 | 24 |
| Smelser | 0.51 | (0.21, 1.25) | 0.49 | (0.19, 1.22) | 3 | 10 |
| Platteville | 0.58 | (0.07, 5.02) | 0.51 | (0.06, 4.56) | 1 | 1 |
| Marathon County | | | | | | |
| **Weston** | 1.58 | (1.31, 1.91) | 1.57 | (1.29, 1.9) | 293 | 196 |
| Waukesha & Milwaukee Counties | | | | | | |
| West Allis | 1.07 | (0.82, 1.39) | 1.01 | (0.77, 1.33) | 93 | 80 |
| West Allis | 1.13 | (0.85, 1.49) | 1.14 | (0.84, 1.54) | 11 | 4 |
| **Greenfield** | 1.17 | (1.06, 1.3) | 1.18 | (1.05, 1.34) | 605 | 529 |
| **West Allis** | 1.20 | (1.09, 1.33) | 1.15 | (1.03, 1.29) | 1202 | 1044 |
| West Allis | 1.25 | (0.94, 1.66) | 1.20 | (0.89, 1.6) | 14 | 14 |
| **West Allis** | 1.32 | (1, 1.75) | 1.35 | (1.02, 1.79) | 42 | 26 |
| **West Allis** | 1.41 | (1.25, 1.59) | 1.36 | (1.19, 1.56) | 2133 | 1632 |

Abbreviations: OR, odds ratio; CI, confidence interval [a] Full adjustment includes age, family history, parity, BMI, drinks of alcohol per week, age at first birth (for women who had given birth), age at menopause (for women who had gone through menopause), race, and education. [b] **Bold** indicates geographic areas significantly different from the background based on the fully-adjusted model. [c] The same city may be listed multiple times and be the center city for multiple clusters because clusters are allowed to overlap with varying radii.

One known limitation in this study is the under-reporting of Wisconsin cancer cases treated in Minnesota facilities [13] (counties along the Minnesota-Wisconsin border). Due to state statutes, the Minnesota Cancer Surveillance System does not allow data sharing with other state cancer registries. The Wisconsin Cancer Reporting System has voluntary contractual agreements with Minnesota facilities to report Wisconsin resident cases directly to the Wisconsin cancer registry, but compliance varies. The under-reporting of cases along the Minnesota border likely biases results for those counties downward. In addition, geographic residence at time of interview is only a proxy for exposure and we encourage the development of new methods to more fully examine residential mobility and length of residence in relation to breast cancer risk.

By exploring breast cancer in Wisconsin, we have identified geographic areas with differing geographic risk from the rest of the state. Of the 23 areas of unique geographic odds identified, 5 were areas of elevated odds and 18 were areas of reduced odds. Our study is based on a rigorous epidemiologic case-control study design, with detailed information about study participants. With these areas detected, future studies can further investigate factors that may be driving these differences in odds, including exploration of breast cancer subtypes.

## Sources of Funding

The results reported herein correspond to specific aims of grant P30 CA014520 to investigator Dr. Amy Trentham-Dietz from the National Cancer Institute. This work was also supported by grant R01 CA047147 to investigator Dr. Polly Newcomb from the National Cancer Institute.

## Data and Computing Code Availability

The data used for this analysis contain protected health information and cannot be made publicly available. Example analytic code is posted on GitHub at the following URL: https://github.com/mkamenet3/SpatialClustersBreastCancerIndividualData

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

We are indebted to the women who participated in this study. The authors would like to thank Julie McGregor for study support, John Hampton for statistical assistance, and Drs. Jane McElroy, Patrick Remington, and Stephanie Robert for scientific leadership of the parent study.

## References

[1] Airoldi C, Magnani C, Lazzarato F, Mirabelli D, Tunesi S, Ferrante D. Environmental asbestos exposure and clustering of malignant mesothelioma in community : a spatial analysis in a population - based case – control study. Environmental Health 2021;20(103):1–13. doi:10.1186/s12940-021-00790-3.
[2] Amirikia KC, Mills P, Bush J, Newman LA. Higher Population-Based Incidence Rates of Triple-Negative Breast Cancer Among Young African-American Women: Implications for Breast Cancer Screening Recommendations. Cancer 2011;117:2747–53. doi:10.1002/cncr.25862.
[3] Ashing KT, Jones V, Bedell F, Phillips T, Erhunmwunsee L. Calling Attention to the Role of Race-Driven Societal Determinants of Health on Aggressive Tumor

Biology : A Focus on Black Americans. JCO Oncology Practice 2021:1–9. doi:10.1200/OP.21.00297.
[4] Beyer KMM, Zhou Y, Matthews K, Hoormann K, Bemanian A, Laud PW, Nattinger AB. Breast and Colorectal Cancer Survival Disparities in Southeastern Wisconsin. Wisconsin Medical Journal 2016;115(1):17–22.
[5] Bivand R., Rundel C.. rgeos: Interface to Geometry Engine - Open Source ('GEOS'); 2019. https://cran.r-project.org/package=rgeos.
[6] Bivand RS, Pebesma E, Gomez-Rubio V. Applied spatial data analysis with {R}, Second edition. Springer, NY; 2013. http://www.asdar-book.org/
[7] Celaya M, Onega T, Gui J, Riddle B, Cherala S, Rees J. Breast cancer stage at diagnosis and geographic access to mammography screening (New Hampshire, 1998-2004). Rural and Remote Health 2010;10.
[8] Chandak A, Nayar P, Lin G. Rural-Urban Disparities in Access to Breast Cancer Screening : A Spatial Clustering Analysis. Journal of Rural Health 2019;35:229–35. doi:10.1111/jrh.12308.
[9] Clarke CA, Glaser SL, West DW, Ereman RR, Erdmann CA, Barlow JM, Wrensch MR. Breast cancer incidence and mortality trends in an affluent population : Marin County, California, USA, 1990 – 1999. Breast Cancer Research 2002;7:1–7.
[10] Coyle YM. The effect of environment on breast cancer risk. Breast Cancer Research and Treatment 2004;84:273–88.
[11] Wisconsin: Coordinate Reference Systems. Danielsen D, Koch T, editors. second. Madison, Wisconsin: Wisconsin State Cartographer's Office; 2009.
[12] Foote M. Racial Disparities in Cancer Incidence and Mortality : Wisconsin and United States, 1996-2000. Wisconsin Medical Journal 2003;1102(5).
[13] Foote M. Wisconsin Cancer Data Bulletin Wisconsin Northwestern Counties with Low Cancer Incidence Rates Due to Underreporting of Cancer Cases Treated at Minnesota Facilities Information from the Wisconsin Cancer Reporting System Wisconsin Cancer Data Bulletin. Tech. Rep.. Wisconsin Department of Health Services; 2019.
[14] Goodman M, Naiman JS, Goodman D, Lakind JS. Cancer clusters in the USA: What do the last twenty years of state and federal investigations tell us? Critical Reviews in Toxicology 2012;42(6):474–90. doi:10.3109/10408444.2012.675315.
[15] Han D, Rogerson PA, Bonner MR, Nie J, Vena JE, Muti P, Trevisan M, Freudenheim JL. Assessing spatio-temporal variability of risk surfaces using residential history data in a case control study of breast cancer. International Journal of Health Geographics 2005;4(9). doi:10.1186/1476-072X-4-9.
[16] Harrell Jr F.E., with contributions from Charles Dupont, others. M.. Hmisc: Harrell Miscellaneous; 2020. https://cran.r-project.org/package=Hmisc.
[17] Henry KA, Mcdonald K, Sherman R, Kinney AY, Stroup AM. Association Between Individual and Geographic Factors. Journal of Women's Health 2014;23(8). doi:10.1089/jwh.2013.4668.
[18] Jacquez GM, Greiling DA. Local clustering in breast, lung and colorectal cancer in Long Island, New York. International Journal of Health Geographics 2003;12:1–12.
[19] Jewett PI, Gangnon RE, Elkin E, Hampton JM, Jacobs EA, Malecki K, LaGro J, Newcomb PA, Trentham-Dietz A. Geographic access to mammography facilities and frequency of mammography screening. Annals of Epidemiology 2018;28(2):65–71.e2. doi:10.1016/j.annepidem.2017.11.012.
[20] Kamenetsky ME, Lee J, Zhu J, Gangnon RE. Regularized Spatial and Spatio-Temporal Cluster Detection. Spatial and Spatio-temporal Epidemiology 2022;41:100462. doi:10.1016/j.sste.2021.100462.
[21] Kelsall JE, Diggle PJ. Non-Parametric Estimation of Spatial Variation in Relative Risk. Statistics in Medicine 1995;14:2335–42.
[22] Khan-Gates JA, Ersek JL, Eberth JM, Adams SA, Pruitt SL. Geographic Access to Mammography and Its Relationship to Breast Cancer Screening and Stage at Diagnosis : A Systematic Review. Women's Health Issues 2015;25(5):482–93. doi:10.1016/j.whi.2015.05.010.
[23] Kulldorff M. A spatial scan statistic. Communications in Statistics - Theory and Methods 1997;26(6):1481–96. doi:10.1080/03610929708831995.
[24] Kulldorff M. SaTScan User Guide V9.4. Tech. Rep.. Harvard Medical School and Harvard Pilgrim Health Care Institute; 2015.
[25] Kulldorff M, Feuer EJ, Miller BA, Freedman LS. Breast Cancer Clusters in the Northeast United States : A Geographic Analysis. American Journal of Epidemiology 1997;146(2):161–70.
[26] Kulldorff M, Nagarwalla N. Spatial disease clusters: Detection and inference. Statistics in Medicine 1995;14(8):799–810. doi:10.1002/sim.4780140809.
[27] Laden F, Hunter DJ. Environmental Risk Factors and Female Breast Cancer. Annual Review of Public Health 1998;19:101–23.
[28] Lipscomb J, Fleming ST, Trentham-Dietz A, Kimmick G, Wu X-C, Morris CR, Zhang K, Smith RA. What Predicts an Advanced-Stage Diagnosis of Breast Cancer ? Sorting Out the Influence of Method of Detection, Access to Care, and Biologic Factors. Cancer Epidemiology, Biomarkers & Prevention 2016;25:613–24. doi:10.1158/1055-9965.EPI-15-0225.
[29] McElroy JA, Gangnon RE, Newcomb PA, Kanarek MS, Anderson H, Brook JV, Trentham-Dietz A, Remington PL. Risk of breast cancer for women living in rural areas from adult exposure to atrazine from well water in Wisconsin. Journal of Exposure Science and Environmental Epidemiology 2007;17:207–14. doi:10.1038/sj.jes.7500511.
[30] Mcelroy JA, Remington PL, Robert SA, Newcomb PA. Geocoding Addresses from a Large Population-based Study : Lessons Learned. Epidemiology 2003;14(4):399–407. doi:10.1097/01.EDE.0000073160.79633.cl.
[31] Minas TZ, Kiely M, Ajao A, Ambs S. An overview of cancer health disparities : new approaches and insights and why they matter. Carcinogenesis 2021;42(1):2–13. doi:10.1093/carcin/bgaa121.
[32] Nickels S, Truong T, Hein R, Stevens K, Buck K, Behrens S, Eilber U, Schmidt M, Haberle L, Vrieling A, Gaudet M, Figueroa J, Schoof N, Spurdle AB, Rudolph A, Fasching PA, Hopper JL, Makalic E, Schmidt DF, Southey MC, Beckmann MW, Ekici AB, Fletcher O, Gibson L, Silva IdS, Peto J, Humphreys MK, Wang J, Cordina-Duverger E, Menegaux F, Nordestgaard BG, Bojesen SE, Lanng C, Anton-Culver H, Ziogas A, Bernstein L, CLarke CA, Brenner H, Muller H, Arndt V, Stegmaier C, Brauch H, Bruning T, Harth V, Network TG, Mannermaa A, Kataja V, Kosma V-M, Hartikainen JM, KConFab, Group AM, Lambrecths D, Smeets D, Neven P, Paridaens R, Flesch-Janys D, Obi N, Wang-Gorhke S, Couch FJ, Olson JE, Vachon CM, Giles GG, Severi G, Baglietto L, Offit K, John EM, Miron A, Andrulis IL, Knight JA, Glendon G, Mulligan AM, Chanock SJ, Lissowska J, Liu J, Cox A, Cramp H, Connley D, Balasubramanian S, Dunning AM, Shah M, Trentham-Dietz A, Newcomb P, Titus L, Egan K, Cahoon EK, Rajaraman P, Sigurdson AJ, Doody MM, Guenel P, Pharoah PDP, Schmidt MK, Hall P, Easton DF, Garcia-Closas M, Milne RL, Chang-Claude J. Evidence of Gene – Environment Interactions between Common Breast Cancer Susceptibility Loci and Established Environmental Risk Factors. PLOS Genetics 2013;9(3). doi:10.1371/journal.pgen.1003284.
[33] Ozonoff A, Webster T, Vieira V, Weinberg J, Ozonoff D, Aschengrau A. A Global Cluster detection methods applied to the Upper Cape Cod cancer data. Environmental Health 2005;4(19). doi:10.1186/1476-069X-4-19.
[34] Paulu C, Aschengrau A, Ozonoff D. Exploring Associations between Residential Location and Breast Cancer Incidence in a Case- Control Study. Environmental Health Perspectives 2002;110(5):471–8.
[35] Pebesma E. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 2018;10(1):439–46. doi:10.32614/RJ-2018-009.
[36] Pesarsick J, Gwilliam M, Adeniran O, Rudisill T, Smith G, Hendricks B. Annals of Epidemiology Original article Identifying high-risk areas for nonfatal opioid overdose : a spatial case-control study using EMS run data. Annals of Epidemiology 2019;36:20–5. doi:10.1016/j.annepidem.2019.07.001.
[37] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria; 2019. https://www.r-project.org/.
[38] Ramis R, Gómez-barroso D, Tamayo I, García-Pérez J, Morales A, Romaguera EP, Lopez-Abente G. Spatial Analysis of Childhood Cancer : A Case / Control Study. PLOS ONE 2015:1–15. doi:10.1371/journal.pone.0127273.
[39] Remington PL, Park S. Breast Cancer Incidence and Mortality in Milwaukee's North Shore Communities. Wisconsin Medical Journal 1997;96(3).
[40] Rubin DB. Multiple Imputation for Nonresponse in Surveys. Hoboken, New Jersey: John Wiley & Sons, Inc; 1987.
[41] Rudolph A, Song M, Brook MN, Milne RL, Mavaddat N, Michailidou K, Bolla MK, Wang Q, Dennis J, Wilcox AN, Hopper JL, Southey MC, Keeman R, Fasching PA, Beckmann MW, Gago-Dominguez M, Castelao JE, Guenel P, Truong T, Bojesen SE, Flyger H, Brenner H, Arndt V, Brauch H, Bruning T, Mannermaa A, Kosma V-M, Lambrecths D, Keupers M, Couch FJ, Vachon C, Giles GG, MacInnis RJ, Figueroa J, Brinton L, Czene K, Brand JS, Gabrielson M, Humphreys K, Cox A, Cross SS, Dunning AM, Orr N, Swerdlow A, Hall P, Pharoah PDP, Schmidt MK, Easton DF, Chatterjee N, Chang-Claude J, Garcia-Closas M. Joint associations of a polygenic risk score and environmental risk factors for breast cancer in the Breast Cancer Association Consortium. International Journal of Epidemiology 2018;47(2):526–36. doi:10.1093/ije/dyx242.
[42] Schwarz G. Estimating the Dimension of a Model. The Annals of Statistics 1978;6(2):461–4. doi:10.1214/aos/1176344136.
[43] Takahashi K., Yokoyama T., Tango T.. FlexScan: Software for the felxible spatial scan statistic. 2009. https://sites.google.com/site/flexscansoftware/.
[44] Taylor TR, Williams CD, Makambi KH, Mouton C, Harrell JP, Cozier Y, Palmer JR, Rosenberg L, Adams-Campbell LL. Racial Discrimination and Breast Cancer Incidence in US Black Women The Black Women' s Health Study. American Journal of Epidemiology 2007;166(1):46–54. doi:10.1093/aje/kwm056.
[45] US Department of Transporation FHA. 2017 National Household Travel Survey. Tech. Rep.; 2017. http://nhts.ornl.gov
[46] Vieira V, Webster T, Weinberg J, Aschengrau A, Ozonoff D. A Global Spatial analysis of lung, colorectal, and breast cancer on Cape Cod: An application of generalized additive models to case-control data. Environmental Health 2005;4(11). doi:10.1186/Received.
[47] Webster T, Vieira V, Weinberg J, Aschengrau A. Method for mapping population-based case-control studies : an application using generalized additive models. International Journal of Health Geographics 2006;5(26):1–10. doi:10.1186/1476-072X-5-26.
[48] Wrensch M, Chew T, Farren G, Barlow J, Belli F, Clarke C, Erdmann CA, Lee M, Moghadassi M, Peskin-mentzer R, Jr CPQ, Souders-mason V, Spence L, Suzuki M, Gould M. Risk factors for breast cancer in a population with high incidence rates. Breast Cancer Research 2003;5(4). doi:10.1186/bcr605.