HHS Public Access

Author manuscript

Cancer Epidemiol Biomarkers Prev. Author manuscript; available in PMC 2022 July 01.

Published in final edited form as:

Cancer Epidemiol Biomarkers Prev. 2022 January ; 31(1): 66–76. doi:10.1158/1055-9965.EPI-21-0838.

Phenotype Discovery and Geographic Disparities of Late-Stage Breast Cancer Diagnosis across U.S. Counties: A Machine Learning Approach

Weichuan Dong[1,2,3,4], **Wyatt P. Bensken**[1,3], **Uriel Kim**[1,2,3], **Johnie Rose**[1,2,3], **Nathan A. Berger**[1,5], **Siran M. Koroukian**[1,2,3]

[1]Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, Ohio, 44106, USA

[2]Center for Community Health Integration, Case Western Reserve University School of Medicine, Cleveland, Ohio, 44106, USA

[3]Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine, Cleveland, Ohio, 44106, USA

[4]Department of Geography, Kent State University, Kent, Ohio, 44240, USA

[5]Center for Science, Health, and Society, Case Western Reserve University School of Medicine, Cleveland, Ohio, 44106, USA

Abstract

Background: Disparities in the stage at diagnosis for breast cancer have been independently associated with various contextual characteristics. Understanding which combinations of these characteristics indicate highest risk, and where they are located, is critical to targeting interventions and improving outcomes for patients with breast cancer.

Methods: The study included women diagnosed with invasive breast cancer between 2009 and 2018 from 680 U.S. counties participating in the Surveillance, Epidemiology, and End Results program. We used a machine learning approach called Classification and Regression Tree (CART) to identify county 'phenotypes', combinations of characteristics that predict the percentage of breast cancer patients presenting with late-stage disease. We then mapped the phenotypes and compared their geographic distributions. These findings were further validated using an alternate machine learning approach called random forest.

Results: We discovered seven phenotypes of late-stage breast cancer. Common to most phenotypes associated with high risk of late-stage diagnosis were high uninsured rate, low mammography use, high area deprivation, rurality, and high poverty. Geographically, these phenotypes were most prevalent in southern and western states, while phenotypes associated with lower percentages of late-stage diagnosis were most prevalent in the northeastern states and select metropolitan areas.

Corresponding Author: Weichuan Dong, 10900 Euclid Ave, WG-43, Cleveland, Ohio, 44106, USA, Telephone: 216-368-3445, weichuan.dong@case.edu.

The authors report no conflicts of interest.

Conclusions: The use of machine learning methods of CART and random forest together with geographic methods offers a promising avenue for future disparities research.

Impact: Local interventions to reduce late-stage breast cancer diagnosis, such as community education and outreach programs, can use machine learning and geographic modeling approaches to tailor strategies for early detection and resource allocation.

Keywords

Breast cancer; Late stage; Disparity; Machine learning; Geographic Information Systems

Introduction

Due to increased screening and awareness of symptoms in the past few decades, the rate of advanced stage breast cancer (for which prognosis is markedly poorer) has decreased dramatically (1). However, breast cancer remains the second leading cause of cancer death in US women (1). Additionally, segments of our population continue to bear a disproportionate burden and experience high rates of late-stage diagnosis.

Previous studies have linked place-based disparities in late-stage breast cancer (LSBC) to numerous contextual characteristics, including socioeconomic status (2, 3), neighborhood deprivation (4, 5), access to screening services (5-9), and availability of primary care physicians (10-12). Studies of LSBC traditionally have considered all predictor variables independently using a parametric regression framework (2-8, 10). However, such a framework does not lend itself to identifying homogeneous subgroups, nor is it efficient in detecting effect measure modification in predicting LSBC. For example, one variable might be important in predicting LSBC in a certain subgroup of a population but may not be as important in other subgroups of the population. Several studies have recognized this issue by stratifying their populations into subgroups by factors such as race/ethnicity (4, 11), income or poverty level (11, 12), and urban-rural status (4, 13). However, the decision on which attributes to stratify the data may be subjective, and the selection of threshold values often are arbitrary. In addition, the complexity of a model increases substantially when considering stratifications along multiple dimensions. Therefore, more advanced solutions are needed to explore the correlates of LSBC among different subgroups of the population.

In this study, we identified ‘phenotypes’, or combinations of county characteristics that predict the percentage of women with breast cancer presenting with late-stage disease. We applied the machine learning technique known as classification and regression tree (CART) with a broad range of county-level characteristics harvested from various sources. This resulted in the classification of counties into phenotypes based on the most important predictors of LSBC. We then examined the geographic distribution of counties with high-risk phenotypes. These findings were further examined using random forest.

Identifying specific clusters of characteristics associated with late-stage diagnosis acknowledges the complex relationships among selected drivers of cancer disparities. It also offers researchers and practitioners a better framework for addressing disparities across heterogeneous and more highly specified groups.

Materials and Methods

Data Source

This study used cancer incidence data from the Surveillance, Epidemiology, and End Results (SEER) program, a resource from the National Cancer Institute. Data in the SEER program cover 34.6% of the United States population with 97% completeness within SEER regions (14). These data cover geographically diverse regions of the country and are broadly representative of the U.S. population along the dimensions of poverty and education (15). The SEER*Stat software was used to query and extract the data. Given the deidentified nature of the data, the Case Western Reserve University Institutional Review Board determined that this work did not involve human subjects research and was thus exempted from review.

Study Population and Variables of Interest

The study included 20 of the 21 registries in the SEER program. We excluded the Alaska Native Tumor Registry since it does not cover cancer cases of all demographic groups (16). The included registries cover the U.S. states of California (Greater California, Los Angeles, San Francisco-Oakland, and San Jose-Monterey registries), Connecticut, Georgia (Atlanta, Greater Georgia, and Rural Georgia registries), Hawaii, Idaho, Iowa, Kentucky, Louisiana, Massachusetts, New Jersey, New Mexico, New York, Utah, and metropolitan areas of Detroit and Seattle-Puget Sound. A total of 732 counties or equivalents (i.e., parishes in Louisiana; for convenience, ‘counties’ is used in the rest of the text) from the 20 SEER registries were included in the study. The outcome of interest was the county-level percentage of LSBC among women diagnosed with invasive breast cancer during a ten-year period between 2009 and 2018. For Massachusetts, only cases from 2009 through 2017 were included because the stage variable in 2018 was not available at the time of the study. Further explanation of the Massachusetts data can be found in Supplemental Figure S1 and Supplemental Figure S2. We used the “Combined Summary Stage” variable from the SEER*Stat software, which classifies tumors into five stages: in situ, local, regional, distant, and unknown stages (excluding in situ and unknown stage cases from our analysis). We collapsed the stage variable to “early stage” (which includes local stage only), and “late stage” (which includes regional and distant stages). To mitigate stochastic variations in percentages of LSBC, we excluded counties with fewer than 16 late-stage cases over the study period, a strategy also adopted by SEER in displaying cancer statistics, resulting in a total of 680 counties in the study. For individuals with multiple tumor records, we selected only the first record.

County-level characteristics were harvested from the Census-American Community Survey (ACS) (17), County Health Rankings & Roadmaps (CHR) (18), Area Health Resources Files (AHRF) (19), Behavioral Risk Factor Surveillance System (BRFSS) (20), and U.S. Food and Drug Administration (FDA) (21), as well as from SEER*Stat. In total, 53 variables were included in our models (Table 1). These variables were selected based on several domains of healthcare resources, behavioral risk factors, population health status, demographic compositions, and other measures of social determinants of health including income, education, occupation, housing, transportation and neighborhood safety. Their relationships

with diagnosis and treatment of breast cancer were explained under the Conceptual Reason for Inclusion column in Table 1. We temporally harmonized the data sources by selecting the years of these variables that overlapped with or were closest to the mid-years of the breast cancer data, with the assumption that no significant secular trends would substantively change any of the factors described in Table 1.

Given the nature of the study, all measures were aggregated at the county level; therefore, we did not account for variables at the individual level.

Statistical Analysis

As described in detail below, machine learning methods (including CART and random forest) and geographic information systems were used to accomplish the objectives of this study. The outcome for all models was county-level percentage of LSBC and all variables in Table 1 were included as candidate predictors.

CART uses conditional inference that recursively partitions data into smaller, more homogeneous groups characterized by combinations of predictors (53, 54). At each split, the data are divided into two homogeneous groups according to a threshold value of one of the predictors, a predictor that results in the two groups with greatest difference in the outcome (54). The splitting procedure is repeatedly applied for each of the split groups by selecting one of the predictors that holds the lowest p-value based on Pearson's correlation test if the predictor variable is numeric, or Kruskal–Wallis test if the predictor variable is categorical (53). This procedure continues until all possible splits are exhausted or until some stopping criteria is met. In this study, we set the following stopping criteria: a maximum tree depth of 6 splits, a minimum number of 80 counties in a terminal node, and lack of statistical significance for variable splits ($\alpha > 0.05$). We also conducted a sensitivity analysis of the CART model with a minimum number of 20 counties in a terminal node.

CART was used to identify phenotypes associated with differing levels of LSBC. We defined a phenotype of LSBC as the combination of characteristics along a top-down path of a tree to a terminal node (a node without any further split) which includes a group of homogenous counties with similar percentages of LSBC. Conceptually, this results in the identification of the combinations of county characteristics that predict the percentage of women with breast cancer presenting with late-stage disease.

Next, we visualized the identified phenotypes using geographic information systems and examined their distribution among regions. To optimize interpretability, the minimum number of counties in a terminal node of the CART model was set to 80 to limit the number of phenotypes presented on the map.

Random forest analysis was used to determine whether our CART model captured the most important variables in predicting the percentage of LSBC. While CART has advantages in variable identification and group classification, a major disadvantage of this single-tree model is that it is sensitive to changes in the data. Hence, the entire tree could be altered if, for example, additional counties are included in the model. In contrast, random forest analysis is more “stable” (55). It uses the same algorithm as CART, but instead of relying

on only one tree, the algorithm creates and aggregates an ensemble of trees using random variable selection and bootstrap sampling (55). It then takes an overall average of these tree models' outputs as a prediction. Next, the mean decrease in accuracy was used to calculate variables' relative importance in predicting the outcome. We created 200,000 trees with all predictor variables included in the analysis. The number of variables randomly sampled as candidates at each split was set equal to the number of splits in the results of the CART model.

Due to the nature of CART and random forest, in that both algorithms select only one of many variables at each split of their trees, they can handle highly correlated variables. However, the associations among candidate predictors remain unknown. For highly correlated predictors, while CART selects only the one that most significantly splits the group, it does not suggest that the rest of the predictors are not predictive of the outcome. Random forest partially addresses this issue with the rankings of variable importance. To further explore the associations among candidate predictors, we conducted a correlation analysis using Pearson's correlation coefficients among all splitting variables in the CART model and top 10 variables in the variable importance plot of the random forest, as well as the variable representing the proportion of women in race/ethnic minorities.

SAS v9.4 and R v3.6.1 were used for the analyses, and ArcGIS Pro v2.7.0 and Tableau v2021.1 were used for mapping and visualization. R packages "rpart", "partykit", and "randomForest" were used for conducting machine learning analyses.

Results

Within the 680 included counties between 2009 and 2018, there were 812,048 women diagnosed with invasive breast cancer, among whom 276,305 (34.0%) were diagnosed at a late stage. The median percentage of LSBC among the counties was 35.4%. The geographic distribution of counties by percentage of LSBC is presented in Figure 1.

We observed that counties in the Northeast states (New York, Massachusetts, Connecticut, and New Jersey) had lower percentages of LSBC compared to the majority of those in the south and west. The box plots showing the county distribution of LSBC by region are included in Supplemental Figure S2.

The results of the CART analysis are shown in Figure 2. Each path down to a terminal node of the tree represents a phenotype of LSBC with corresponding characteristics. P-values on the splitting nodes (< 0.05) suggest that the groups of counties split by the thresholds of the variables are statistically significantly different from each other in terms of percentage of LSBC. Counties within the same terminal node have similar percentages of LSBC and belong to the same phenotype. Phenotypes are classified into Low-Risk (LR), Medium-Risk (MR), and High-Risk (HR) by their median percentage of LSBC among counties, with LR 1 having the lowest percentage and HR 3 having the highest percentage.

The results show that among all candidate predictors, CART selected 5 variables as splitting nodes, with *percentage of uninsured women aged 18-64* on the top, followed sequentially by *percentage of mammography use among women aged 67-69 enrolled in Medicare,*

Area Deprivation Index (ADI), Urban Influence Code (UIC), percentage of people under poverty, and per capita income. The ADI is an index of social deprivation calculated from census variables, which incorporates 17 separate factors covering domains of education, employment, income, housing (costs and crowding), and transportation access (56). The ADI ranges from 30.5 to 154.5 among all counties in the study area (mean: 98.8, median: 98.5). A higher ADI indicates that the county is more deprived. Counties with an ADI greater than 99.7, as shown in the CART output, means that they were more deprived than 52.2% of all counties. Counties with a higher UIC were more rural. Specifically, counties with UIC less than or equal to 3 were metropolitan counties or micropolitan counties adjacent to a large metropolitan county.

HR 3 is the phenotype associated with the highest median percentage of LSBC (40.1%). It includes 89 counties and is characterized as having a higher percentage of uninsured middle-aged women (> 11.6%), greater area deprivation (ADI > 99.7), and more people under poverty (> 26.1%). HR 2 is the phenotype associated with the second highest median percentage of LSBC (38.4%). HR 2 has the same levels of percentage of uninsured middle-aged women and area deprivation compared to HR 1 but has a lower poverty rate (≤ 26.1%) and higher per capita income (> 32,946 US dollars). HR 1 has a slightly lower median percentage of LSBC than HR 2 (37.0% vs 38.4%), and their only difference is in per capita income.

MR 1 has the largest group of counties (n=139) among all phenotypes and has a median percentage of LSBC that is close to that of the overall study area (35.5% vs 35.4%). Counties of MR 1 has greater rates of uninsured middle-aged women (> 11.6%) but lower area deprivation (≤ 99.7).

Phenotypes LRs 1, 2, and 3, have better outcomes of LSBC compared to HRs 1, 2, 3, and MR 1. The key difference between LR phenotypes and MR or HR phenotypes is the top splitting variable (uninsured middle-aged women), with a threshold of 11.6% that separates the tree into two large branches. The variable that separates LR 1 from LRs 2 and 3 is mammography use among Medicare beneficiaries aged 67-69 years (> 68.1%). When mammography use is at a lower level (≤ 68.1%), UIC come into play and differentiates LR 2 (urban) with LR 3 (rural).

The sensitivity analysis of CART (with a minimum of 20 counties in a terminal node) presents additional splits compared to the main model, which include availability of obstetrics and gynecologists, access to exercise opportunities, breast cancer incidence rate, availability of primary care physicians, women in professional occupations, and Medicare eligibility (Supplemental Figure S3). Note that UIC no longer appears in this model, suggesting potential correlations between this and other variables.

Figure 3 shows the geographic distribution of the phenotypes identified in the main CART model. We observed strong regional variations in the composition of phenotypes. For example, Massachusetts is the only region that has only one phenotype, which is LR 1, the most favorable phenotype of LSBC, while California, Kentucky, and Georgia tend to have the greatest variability of phenotypes. LR 1 also covers a large number of counties in

Upstate New York and Iowa. LR 2 is common in metropolitan areas of San Francisco Bay Area (CA), New York (NY, NJ, and CT), Seattle, Detroit, Louisville (KY), Lexington (KY), metropolitan areas in upstate New York, and urban areas in Hawaii and Iowa. LR 3 is mostly found in rural areas in Iowa, Kentucky, and Upstate New York. No counties of any LR phenotypes appears in Georgia, Louisiana, and Idaho, and only one county in New Mexico (Los Alamos County). MR 1 covers most counties in Utah, Idaho, metropolitan Atlanta in Georgia, and Coastal California. The three HR phenotypes (with higher rates of uninsured women, greater area deprivation, and worst outcome of LSBC) are found mostly in rural counties of the southern and western states of Georgia, Louisiana, New Mexico, Idaho, and California. We also summarized these descriptions by phenotype in Table 2 and visualized the distribution of phenotypes by region in Supplemental Figure S4.

The ranking of variables by their importance for predicting percentages of LSBC based on the random forest analysis is shown in Figure 4. This dot plot ranks the variables in descending order relative to the most important predictor. Among all variables, uninsured rate among middle-aged women and mammography use among Medicare beneficiaries aged 67-69 are the two most important predictors of percentage of LSBC. Other important predictors are adult obesity, percentage of children in poverty, ADI, percentage of people under 200% of poverty, percentage of female-headed households, and percentage of people with poor or fair health. Detailed information regarding the calculation of variable importance with intermediate results, as well as an alternative measure of mean decrease in node impurity of the random forest analysis are included in Supplemental Table S1.

The correlation matrix in Supplemental Table S2 suggests that some of the candidate predictors were highly correlated with absolute values of correlation coefficient (cc) greater than 0.7. Notably, ADI was highly correlated with multiple variables, including women with high school degree or above, per capita income, teen birth, percent people with poor of fair health, poverty, child poverty, and population below 200% of the poverty level (cc: −0.77, −0.74, 0.81, 0.85, 0.88, 0.89, and 0.90, respectively). The correlation between the percentage of women in race/ethnic minority groups and other variables were moderate, with the largest cc at 0.65 with female-headed households.

Discussion

Using CART analysis, our study identified low, medium, and high risk phenotypes of LSBC consisting of county-level characteristics that were predictive of LSBC. These phenotypes were defined by combinations of indicators of uninsured rate, mammography use, area deprivation, urban-rural status, poverty rate, and per capita income. Among those, the importance of uninsured rate and mammography use was further evidenced by their top rankings in the random forest analysis.

When a smaller number of counties was allowed in a terminal node, additional characteristics came into play in the CART model. Surprisingly, the percentage of racial and ethnic minorities, a factor frequently emphasized in previous studies of LSBC (3, 4, 7, 10-13, 25, 38, 43), did not appear in the results of the CART models and was only ranked

13th in the random forest plot. Our study suggests that other contextual factors might have played more critical roles than the constructs of race and ethnicity themselves.

We also recognized the correlations among variables. Some predictors, although did not appear in CART, could still have important implications due to their strong correlations with the outcome and splitting variables. For example, several measures of population socioeconomic status were highly correlated with ADI, most of which also had relatively high rankings in the random forest plot. In addition, despite the absence of adult obesity as a splitting variable in CART, it was identified as a top-ranking variable in predicting LSBC by the random forest analysis, which is notable since it was the only physiological factor associated with LSBC.

The spatial distribution of phenotypes shows that LR phenotypes, or those comprised of counties with a relatively lower percentage of LSBC, were prevalent in northeastern states, Iowa, and select metropolitan areas, whereas HR phenotypes were mostly observed in the southern states of Georgia, Louisiana, New Mexico, and some rural areas in other states. The unbalanced distribution of phenotypes suggests that there were geospatial disparities in LSBC, and these disparities were strongly associated with population characteristics along multiple dimensions.

The geographic clustering of phenotypes suggests that the association between LSBC and various socioeconomic characteristics may be mediated by geographic-related factors. We observed strong differences in the phenotype composition of states, such as northeastern states versus southern and western states. These state-level disparities could be related to the differences in state-specific culture and policies. For example, states with stringent eligibility criteria for Medicaid enrollment may observe higher rates of uninsured among people with low socioeconomic status compared to states that do not. This was especially the case after 2014 when some states expanded their Medicaid programs under the Patient Protection and Affordable Care Act, while others had not by the end of the study period (57).

We also noted within-state variations, especially in California, Kentucky, and Georgia, where counties in large metro areas generally had more favorable phenotypes than their rural neighbors. This may be due to the large differences in demographic and socioeconomic characteristics and availability of resources between rural and urban areas as indicated in earlier studies (9, 13). However, not all urban areas outperformed rural areas in LSBC. For example, several rural counties in northern California had LR phenotypes, while Los Angeles County and San Diego County, the state's two most populous counties, were of the MR phenotype. This suggests that there was a complex relationship between LSBC and urbanicity, which could involve associations with other variables, such as uninsured rates and mammography use as shown in the CART models.

To our knowledge, this is the first study that combines machine learning and geographic methods to explore the association between LSBC and population characteristics using cancer registry data. Prior studies have investigated associations between LSBC and various demographic, socioeconomic, and behavioral factors using geospatial analyses and parametric regression models (2-8, 10-12, 46, 58, 59). However, these studies did not

evaluate the roles of these factors in specific subgroups of the population. For example, one indicator may be important to a certain group of a population but less important to other subgroups of the population. As indicated in Figure 2, when mammography use was less than or equal to 68.1%, UIC was able to significantly distinguish LR 2 and LR 3 based on the percentage of LSBC. However, it was no longer used as a splitting node when mammography use was higher. Similarly, when uninsured rate was greater than 11.6%, neither mammography use nor rural-urban status was as important as the other variables that appeared in the right branch of the tree. By using the CART-defined phenotypes, our study identifies correlates that are specific to various subgroups of the population that share common characteristics.

There are several methodological strengths that lend confidence to the study results. First, compared to parametric regression models (such as logistic regression), CART and random forest methods can deal with a greater number of predictor variables simultaneously without concerns about outliers, multicollinearity, heteroscedasticity, or distributional error structures that affect parametric procedures. Second, both CART and random forest methods are able to handle highly correlated data due to their variable selection and bootstrap sampling strategies. Finally, the identified phenotypes capture both the outcome and top predictors variables, allowing the examination of geographic patterns of LSBC from multiple aspects.

The main limitation of the current study is that individual regions in the study area were disconnected in geography, preventing the use of geospatial analyses designed for contiguous regions, such as the spatial scan statistics (60), the Local Indicators of Spatial Association (LISA) (61), and the recently developed geographically weighted random forest (62, 63). A geospatial analysis on a contiguous study area, such as the contiguous United States, could help us understand a broad scope of the disparities in LSBC and provide insights into why neighboring areas present similar or different patterns. However, this limitation was tempered by the wide distribution of the study area across the country which covered diverse populations that were comparable to the overall United States. Another limitation is that we were not able to incorporate additional individual-level characteristics in the analysis. Nevertheless, the findings of the study are still valid as the nature of our study was to discover community-level drivers of LSBC disparities.

In summary, our study shows that the use of machine learning and geographic methods is a promising avenue for future disparities research. The findings of our study suggest that the disparities of LSBC are associated with multiple characteristics of the population, and these associations vary greatly across geographies. Local interventions to reduce late-stage diagnosis of breast cancer, such as community education and outreach programs, should consider the characteristics of their communities; thus translational and implementation researchers should consider phenotype-tailored interventions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Dr. Will Penman of Composition Coaching for his advice and assistance in the earlier planning of this manuscript. This study was funded by a grant from the National Cancer Institute, Case Comprehensive Cancer Center (P30 CA043703) to S. Koroukian, J. Rose, and W. Dong. W. Dong and S. Koroukian are supported by a grant from the American Cancer Society (132678-RSGI-19-213-01-CPHPS) and by contracts from Cleveland Clinic Foundation, including a subcontract from Celgene Corporation. S. Koroukian is also supported by grants from the National Institutes of Health (R15 NR017792, and UH3-DE025487) and the American Cancer Society (RWIA-20-111-02 RWIA). W. Bensken is supported by a grant from the National Institute on Minority Health and Health Disparities (F31MD015681). U. Kim is supported by grants from the National Institute of General Medical Sciences (5T32GM007250), National Center for Advancing Translational Sciences (5TL1TR002549), and the PhRMA Foundation (PDHO18). J. Rose is supported by grants from the National Institute of Dental and Craniofacial Research (1UH2DE025487-01), the National Heart Lung and Blood Institute (R01 HL153175), and the American Cancer Society (RWIA-20-111-02 RWIA). N. Berger is supported by grants from the National Cancer Institute (2P50 CA150964, 2U54CA163060, P20CA233216, R25CA221718, and R25CA225461).

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A (2021). Cancer statistics, 2021. CA: a cancer journal for clinicians, 71(1), 7–33. [PubMed: 33433946]
2. MacKinnon JA, Duncan RC, Huang Y, Lee DJ, Fleming LE, Voti L, et al. (2007). Detecting an association between socioeconomic status and late stage breast cancer using spatial analysis and area-based measures. Cancer Epidemiology and Prevention Biomarkers, 16(4), 756–762.
3. Wang F, Luo L, McLafferty S (2010). Healthcare access, socioeconomic factors and late-stage cancer diagnosis: an exploratory spatial analysis and public policy implication. International journal of public policy, 5(2-3), 237–258. [PubMed: 23316251]
4. Spada NG, Geramita EM, Zamanian M, van Londen GJ, Sun Z, Sabik LM (2021). Changes in Disparities in Stage of Breast Cancer Diagnosis in Pennsylvania After the Affordable Care Act. Journal of Women's Health, 30(3), 324–331.
5. Anderson RT, Yang TC, Matthews SA, Camacho F, Kern T, Mackley HB, et al. (2014). Breast cancer screening, area deprivation, and later-stage breast cancer in Appalachia: does geography matter?. Health Services Research, 49(2), 546–567. [PubMed: 24117371]
6. Huang B, Dignan M, Han D, Johnson O (2009). Does distance matter? Distance to mammography facilities and stage at diagnosis of breast cancer in Kentucky. The Journal of Rural Health, 25(4), 366–371. [PubMed: 19780916]
7. Henry KA, Boscoe FP, Johnson CJ, Goldberg DW, Sherman R, Cockburn M (2011). Breast cancer stage at diagnosis: is travel time important?. Journal of community health, 36(6), 933. [PubMed: 21461957]
8. Onitilo AA, Liang H, Stankowski RV, Engel JM, Broton M, Doi SA, et al. (2014). Geographical and seasonal barriers to mammography services and breast cancer stage at diagnosis. Rural and remote health, 14(3), 180.
9. Chandak A, Nayar P, Lin G (2019). Rural-urban disparities in access to breast cancer screening: a spatial clustering analysis. The Journal of Rural Health, 35(2), 229–235. [PubMed: 29888497]
10. Wang F, McLafferty S, Escamilla V, Luo L (2008). Late-stage breast cancer diagnosis and health care access in Illinois. The Professional Geographer, 60(1), 54–69. [PubMed: 18458760]
11. Kuo TM, Mobley LR, Anselin L (2011). Geographic disparities in late-stage breast cancer diagnosis in California. Health & place, 17(1), 327–334. [PubMed: 21144791]
12. Barry J, Breen N, Barrett M (2012). Significance of increasing poverty levels for determining late-stage breast cancer diagnosis in 1990 and 2000. Journal of Urban Health, 89(4), 614–627. [PubMed: 22322332]
13. McLafferty S, Wang F (2009). Rural reversal? Rural-urban disparities in late-stage cancer risk in Illinois. Cancer, 115(12), 2755–2764. [PubMed: 19434667]
14. Surveillance, Epidemiology, and End Results (SEER) Program SEER*Stat Database: Incidence - SEER Research Plus Limited-Field Data, 21 Registries, Nov 2020 Sub (2000-2018) - Linked To County Attributes - Total U.S., 1969-2019 Counties, National Cancer Institute, DCCPS,

Surveillance Research Program, released April 2021, based on the November 2020 submission. Accessed April 25, 2021

15. Surveillance, Epidemiology, and End Results (SEER) Program. Characteristics of the SEER Population Compared with the Total United States Population. Available from: https://seer.cancer.gov/registries/characteristics.html. Accessed April 25, 2021
16. Surveillance, Epidemiology, and End Results (SEER) Program. About the SEER Registries. Available from: https://seer.cancer.gov/registries/. Accessed April 25, 2021
17. American Community Survey (ACS. Available from: https://www.census.gov/programs-surveys/acs. Accessed April 25, 2021
18. County Health Rankings & Roadmaps (CHR). Available from: https://www.countyhealthrankings.org/. Accessed April 25, 2021
19. Area Health Resources Files (AHRF). Available from: https://data.hrsa.gov/topics/health-workforce/ahrf. Accessed April 25, 2021
20. Behavioral Risk Factor Surveillance System (BRFSS). Available from: https://www.cdc.gov/brfss/index.html. Accessed April 25, 2021
21. U.S. Food and Drug Administration (FDA). Available from: https://www.fda.gov/. Accessed April 25, 2021
22. Gusberg SB (1981). The gynecologist and breast cancer. Israel journal of medical sciences, 17(9-10), 843–846. [PubMed: 7309470]
23. Frank E, Rimer BK, Brogan D, Elon L (2000). US women physicians' personal and clinical breast cancer screening practices. Journal of women's health & gender-based medicine, 9(7), 791–801.
24. Nguyen-Pham S, Leung J, McLaughlin D (2014). Disparities in breast cancer stage at diagnosis in urban and rural adult women: a systematic review and meta-analysis. Annals of epidemiology, 24(3), 228–235. [PubMed: 24462273]
25. Lannin DR, Mathews HF, Mitchell J, Swanson MS, Swanson FH, Edwards MS (1998). Influence of socioeconomic and cultural factors on racial differences in late-stage presentation of breast cancer. Jama, 279(22), 1801–1807. [PubMed: 9628711]
26. Meyer CP, Allard CB, Sammon JD, Hanske J, McNabb-Baltar J, Goldberg JE, et al. (2016). The impact of Medicare eligibility on cancer screening behaviors. Preventive medicine, 85, 47–52. [PubMed: 26763164]
27. Coughlin SS (2019). Social determinants of breast cancer risk, stage, and survival. Breast cancer research and treatment, 177(3), 537–548. [PubMed: 31270761]
28. Nguyen BC, Alawadi ZM, Roife D, Kao LS, Ko TC, Wray CJ (2016). Do socioeconomic factors and race determine the likelihood of breast-conserving surgery?. Clinical breast cancer, 16(4), e93–e97. [PubMed: 27297238]
29. Berrian JL, Liu Y, Lian M, Schmaltz CL, Colditz GA (2021). Relationship between insurance status and outcomes for patients with breast cancer in Missouri. Cancer, 127(6), 931–937. [PubMed: 33201532]
30. Zha N, Alabousi M, Patel BK, Patlas MN (2019). Beyond universal health care: barriers to breast cancer screening participation in Canada. Journal of the American College of Radiology, 16(4), 570–579. [PubMed: 30947889]
31. Shapiro JA, Seeff LC, Nadel MR (2001). Colorectal cancer-screening tests and associated health behaviors. American journal of preventive medicine, 21(2), 132–137. [PubMed: 11457633]
32. Mu L, Mukamal KJ (2016). Alcohol consumption and rates of cancer screening: Is cancer risk overestimated?. Cancer Causes & Control, 27(2), 281–289. [PubMed: 26590914]
33. Hoffman SD, Maynard RA (Eds.). (2008). Kids having kids: economic costs and social consequences of teen pregnancy (2nd ed.). Washington, DC: Urban Institute Press.
34. SmithBattle L, Freed P "Teen mothers' mental health." MCN: The American Journal of Maternal/Child Nursing 41.1. 2016: 31–36.
35. Griggs MJ, Walker R (2008). The costs of child poverty for individuals and society: a literature review.

36. Bailey BA, Byrom AR (2007). Factors predicting birth weight in a low-risk sample: the role of modifiable pregnancy health behaviors. Maternal and Child Health Journal, 11(2), 173–179. [PubMed: 17091398]
37. Cui Y, Whiteman MK, Flaws JA, Langenberg P, Tkaczuk KH, Bush TL (2002). Body mass and stage of breast cancer at diagnosis. International journal of cancer, 98(2), 279–283. [PubMed: 11857420]
38. Jones BA, Kasl SV, Curnen MGM, Owens PH, Dubrow R (1997). Severe obesity as an explanatory factor for the black/white difference in stage at diagnosis of breast cancer. American journal of epidemiology, 146(5), 394–404. [PubMed: 9290499]
39. Mahmood A, Kedia S, Dillon P, Arshad H, Ray M (2021). Food security status and breast cancer screening among women in the United States: Evidence from Health and Retirement Study and Health Care and Nutrition Study. Research Square, Preprint posted online February 18, 2021
40. Gage-Bouchard EA (2017). Social support, flexible resources, and health care navigation. Social science & medicine, 190, 111–118. [PubMed: 28858696]
41. Tarlov E, Zenk SN, Campbell RT, Warnecke RB, Block R (2009). Characteristics of mammography facility locations and stage of breast cancer at diagnosis in Chicago. Journal of Urban Health, 86(2), 196–213. [PubMed: 18972211]
42. Peek ME, Sayad JV, Markwardt R (2008). Fear, fatalism and breast cancer screening in low-income African-American women: the role of clinicians and the health care system. Journal of general internal medicine, 23(11), 1847–1853. [PubMed: 18751758]
43. Smigal C, Jemal A, Ward E, Cokkinides V, Smith R, Howe HL, et al. (2006). Trends in breast cancer by race and ethnicity: update 2006. CA: a cancer journal for clinicians, 56(3), 168–183. [PubMed: 16737949]
44. Weinmann S, Taplin SH, Gilbert J, Beverly RK, Geiger AM, Yood MU, et al. (2005). Characteristics of women refusing follow-up for tests or symptoms suggestive of breast cancer. JNCI Monographs, 2005(35), 33–38.
45. Davidson PL, Bastani R, Nakazono TT, Carreon DC (2005). Role of community risk factors and resources on breast carcinoma stage at diagnosis. Cancer: Interdisciplinary International Journal of the American Cancer Society, 103(5), 922–930.
46. Tatalovich Z, Zhu L, Rolin A, Lewis DR, Harlan LC, Winn DM (2015). Geographic disparities in late stage breast cancer incidence: results from eight states in the United States. International journal of health geographics, 14(1), 1–11. [PubMed: 25563056]
47. Crawford J, Ahmad F, Beaton D, Bierman AS (2016). Cancer screening behaviours among South Asian immigrants in the UK, US and Canada: a scoping study. Health & social care in the community, 24(2), 123–153. [PubMed: 25721339]
48. Roche LM, Niu X, Stroup AM, Henry KA (2017). Research Full Report: Disparities in Female Breast Cancer Stage at Diagnosis in New Jersey: A Spatial-Temporal Analysis. Journal of Public Health Management and Practice, 23(5), 477. [PubMed: 28430705]
49. Liu Y, Zhang J, Huang R, Feng WL, Kong YN, Xu F, et al. (2017). Influence of occupation and education level on breast cancer stage at diagnosis, and treatment options in China: A nationwide, multicenter 10-year epidemiological study. Medicine, 96(15).
50. Pudrovska T, Carr D, McFarland M, Collins C (2013). Higher-status occupations and breast cancer: a life-course stress approach. Social Science & Medicine, 89, 53–61. [PubMed: 23726216]
51. Jelleyman T, Spencer N (2008). Residential mobility in childhood and health outcomes: a systematic review. Journal of Epidemiology & Community Health, 62(7), 584–592. [PubMed: 18559440]
52. Wohlfahrt J, Andersen PK, Mouridsen HT, Melbye M (2001). Risk of late-stage breast cancer after a childbirth. American journal of epidemiology, 153(11), 1079–1084. [PubMed: 11390326]
53. Hothorn T, Hornik K, Zeileis A (2006). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical statistics, 15(3), 651–674.
54. Ryo M, Rillig MC (2017). Statistically reinforced machine learning for nonlinear patterns and variable interactions. Ecosphere, 8(11).
55. Breiman L (2001). Random forests. Machine learning, 45(1), 5–32.

56. Singh GK. Area deprivation and widening inequalities in US mortality, 1969–1998. Am J Public Health. 2003;93(7):1137–1143. [PubMed: 12835199]

57. Le Blanc JM, Heller DR, Friedrich A, Lannin DR, Park TS (2020). Association of Medicaid expansion under the Affordable Care Act with breast cancer stage at diagnosis. JAMA surgery, 155(8), 752–758. [PubMed: 32609338]

58. McElroy JA, Remington PL, Gangnon RE, Hariharan L, Andersen LD (2006). Identifying Geographic Disparities in the Early Detection of Breast Cancer Using a Geographic Information System. Preventing chronic disease, 3(1).

59. de Oliveira NPD, de Camargo Cancela M, Martins LFL, de Souza DLB (2021). A multilevel assessment of the social determinants associated with the late stage diagnosis of breast cancer. Scientific Reports, 11(1), 1–9. [PubMed: 33414495]

60. Kulldorff M (1997). A spatial scan statistic. Communications in Statistics-Theory and methods, 26(6), 1481–1496.

61. Anselin L (1995). Local indicators of spatial association—LISA. Geographical analysis, 27(2), 93–115.

62. Georganos S, Grippa T, Gadiaga AN, Linard C, Lennert M, Vanhuysse S, et al. (2019). Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto International, 36(2), 121–136.

63. Luo Y, Yan J, McClure S (2021). Distribution of the environmental and socioeconomic risk factors on COVID-19 death rate across continental USA: a spatial nonlinear analysis. Environmental Science and Pollution Research, 28(6), 6587–6599. [PubMed: 33001396]

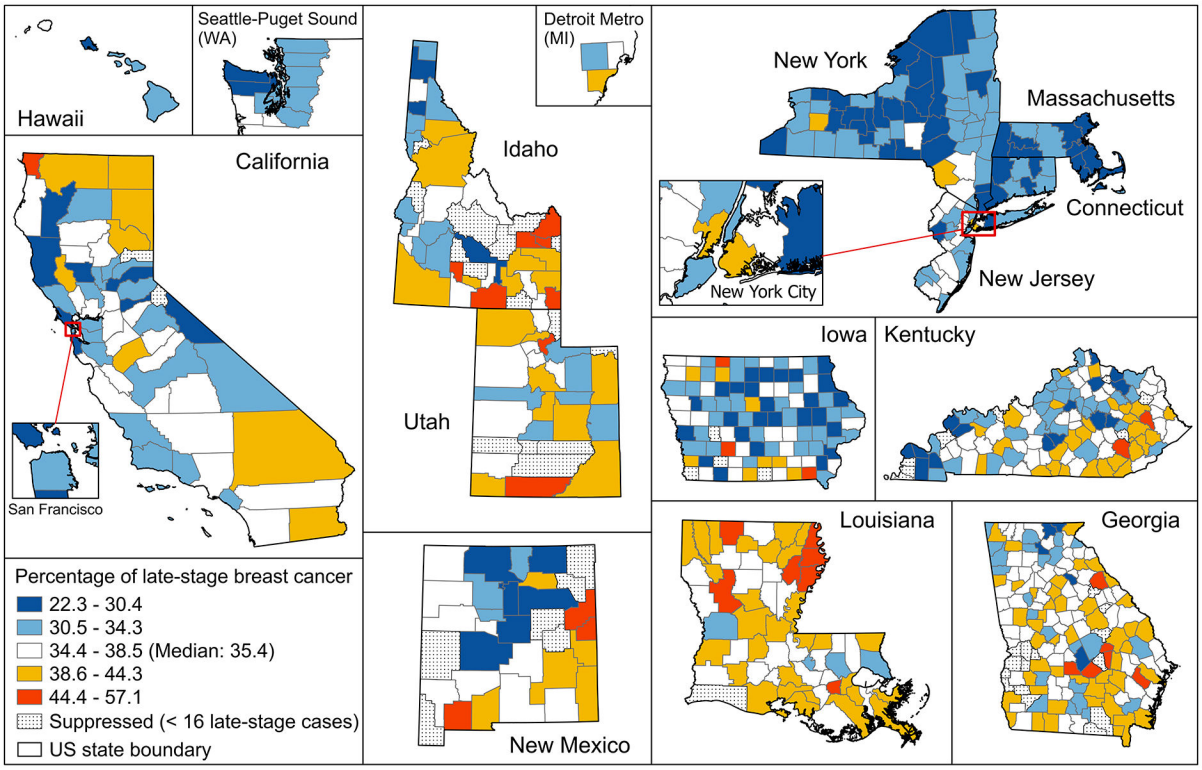

Figure 1.

Percentage of late-stage breast cancer at the time of diagnosis among SEER counties during years 2009-2018 (2009-2017 for Massachusetts). Maps are not on the same scale. Percentages were classified by Jenks natural breaks classification method.

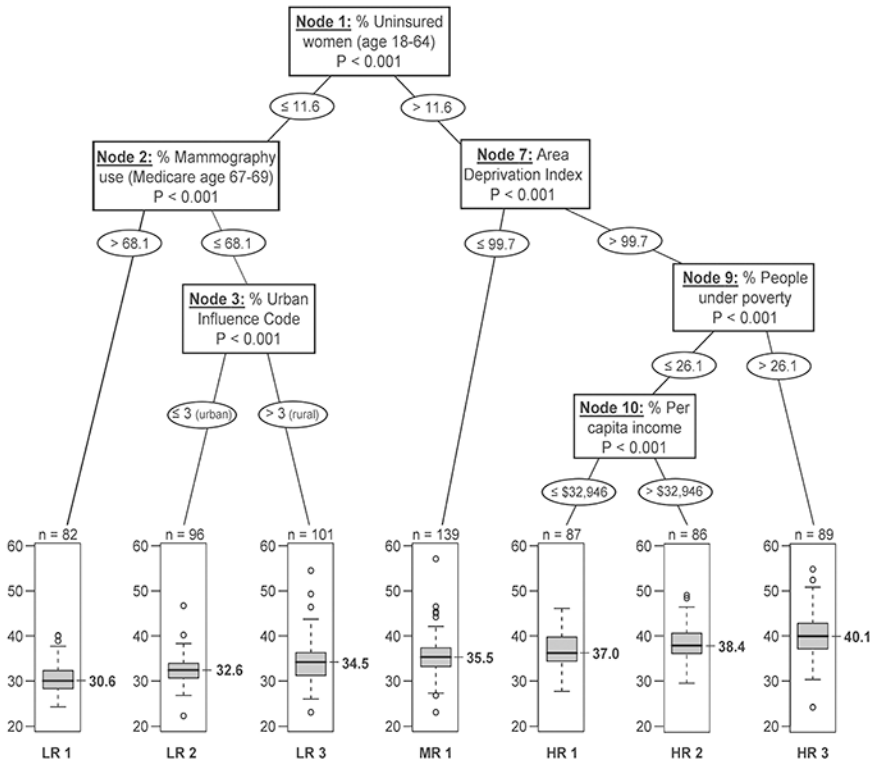

Figure 2.

Classification and regression tree analysis in predicting percentage of LSBC. Each path down to a terminal node represents a phenotype of LSBC. Box plots in the terminal nodes represent the percentages of LSBC among counties. Minimum number of counties in a terminal node was set to 80. Counties with an Area Deprivation Index greater than 99.7 were more deprived than 52.2% of all 680 counties in the study. Counties with an Urban Influence Code (UIC) smaller or equal to 3 were Metropolitan counties or Micropolitan counties adjacent to a large Metropolitan county; Counties with a higher UIC were more rural.

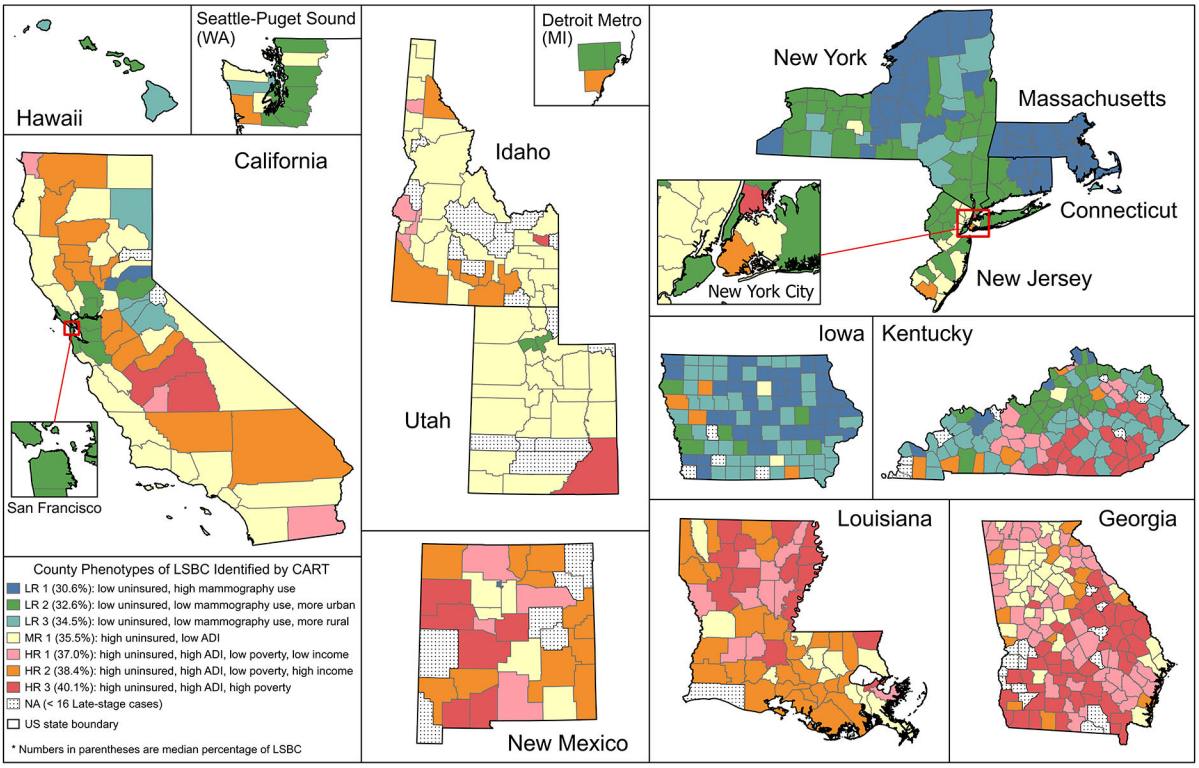

Figure 3.

Geographic distribution of county phenotypes of late-stage breast cancer (LSBC) identified by the classification and regression tree (CART). Maps are not on the same scale.

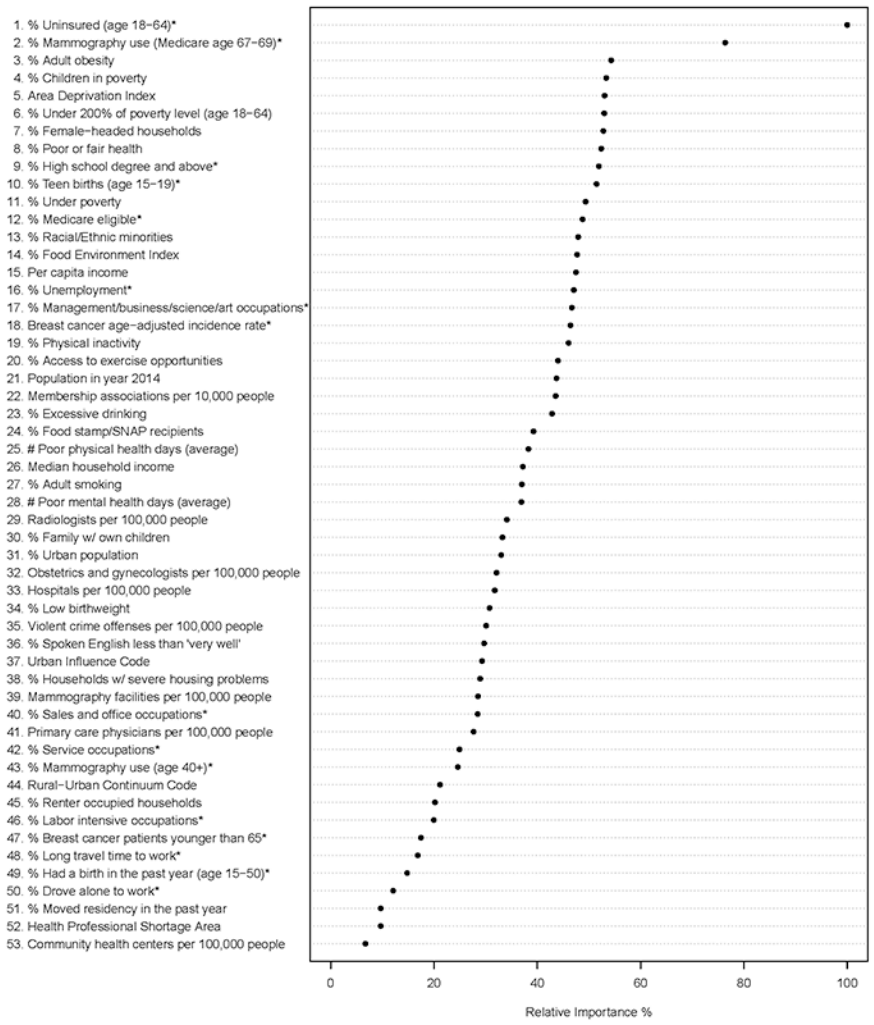

Figure 4.

Dot chart of random forest analysis showing variable importance for predicting counties with high proportion of LSBC. The most important variable is at the top and scaled to 100%. The importance of the rest of the variables is shown relative to the top one. The star sign (*) at the end of a variable indicates the variable is specific to females.

Table 1.

Definitions of predictor variables in CART and random forest analyses

Variable	Year(s)	Source	Conceptual Reason for Inclusion
% Patients younger than 65	2009-2018	SEER	Women aged 65 years and older may have better access to health insurance as they are mostly eligible for Medicare.
Incidence rate of breast cancer (age-adjusted)			Incidence rate is a measure of the breast cancer burden in the population.
Mammography use (age 40+)	2010	BRFSS	Mammography screening helps detect breast cancer at an early stage (1).
Mammography facilities per 100,000 population	2016	FDA	The availability of mammography could impact the uptake of screening in an area (5-9).
Hospitals per 100,000 population	2015	AHRF	Hospitals could recommend and provide screening services for women.
Community Health Centers per 100,000 population	2014		Community Health Centers could provide screening services or facilitate referrals for screening.
Primary care physicians per 100,000 population	2014		Sufficient primary care physicians are essential for preventive cancer care, and referrals for diagnostic services when necessary (10-12).
Obstetricians-gynecologists per 100,000 population	2015		Obstetrician and gynecologists are more likely to discuss and perform breast cancer screening and more likely to recognize breast cancer than other physicians (22, 23).
Radiologists per 100,000 population	2015		Radiologists are essential for the diagnosis and staging of breast cancer.
Population estimate	2014		An indicator of urbanicity, which may be associated with screening uptake (24).
% Urban Population	2010		An indicator of urbanicity, which may be associated with screening uptake (24).
Per Capita Income	2014		Lower income is associated with late-stage diagnosis of breast cancer (10, 25).
Rural-Urban Continuum Code	2013		An indicator of urbanicity, which may be associated with screening uptake (24)
Urban Influence Code	2013		An indicator of urbanicity, which may be associated with screening uptake (24)
Health Professional Shortage Area - Primary Care	2015		Sufficient availability of primary care physicians is essential for preventive cancer care, and referrals for diagnostic services when necessary (10-12).
% Eligible for Medicare	2014		Medicare-eligible individuals are more likely to undergo all cancer preventive services (26).
Median Household Income	2014		Lower income has been associated with late-stage diagnosis of breast cancer (10, 25).
% People in Poverty	2014		Poverty has been associated with late-stage diagnosis of breast cancer (27).
% Food Stamp or SNAP Recipients	2014		Receipt of SNAP benefits may be predictive of breast cancer tumor size (28).
% Uninsured women (age 18-64)	2014		Women without health insurance are more likely to be diagnosed at late stage for breast cancer compared to those with insurance (29).
% People under 200% of Poverty (age 18-64)	2014		Lower income and poverty are associated with late stage at breast cancer diagnosis (10, 25, 27).
% People with poor or fair health	2014	CHR	Overall poor or fair health may be positively or negatively associated with mammography screening based on either increased healthcare contacts or competing health priorities (30).
Poor physical health days	2014		More poor physical health days may be positively or negatively associated with mammography screening based on either increased healthcare contacts or competing health priorities (30).

Variable	Year(s)	Source	Conceptual Reason for Inclusion
Poor mental health days	2014		People with more poor mental health days may have lower priority of screening in the context of managing other medical and life issues (30).
Adult smoking	2014		Smoking may be inversely associated with use of colorectal cancer-screening tests. (31).
Excessive drinking	2014		Alcohol consumption may be associated with breast cancer screening rates (32).
Teen births	2007-2013		Teenage women who bear a child are more likely to have lower socioeconomic status and psychological distress in their later lives (33, 34).
Children in poverty	2014		Child poverty could reflect long term negative consequences of the population along various aspects of social determinants health (35).
Low birthweight	2008-2014		Low birthweight may indicate maternal exposure to various health risks (36).
Adult obesity	2013		High body mass is associated with late-stage breast cancer at diagnosis (37, 38).
Food environment index	2010&2014		The urgency of food insecurity may deprioritize the receipt of preventive screening services (39).
Physical Inactivity	2013		Physical activity may be associated with use of colorectal cancer screening tests (31).
Access to exercise opportunities	2010&2014		A study found that physical activity was also associated with use of colorectal cancer-screening tests (31).
Mammography use (Medicare age 67-69)	2014		Mammography screening helps detect breast cancer at an early stage (1).
Social associations	2014		People with adequate social support had more healthcare access and fostered more productive relationships with their healthcare providers (40).
Violent crime	2012-2014		Homicide rate in the neighborhoods of women's nearest screening facility is associated with breast cancer late-stage diagnosis (41).
Severe housing problems	2009-2013		People with severe housing problems might have lower priority of screening in the context of managing other acute issues (42).
% Racial/Ethnic Minorities	2012-2016	ACS	Race and ethnicity have been strong predictors of late-stage breast cancer (25, 43).
% Family with own children (age < 18)			Women with more children are less likely to receive follow-up of tests or seek care for symptoms suggestive of breast cancer (44).
% Female-headed households			Women from neighborhoods with greater percentages of female-headed households may be at higher risk of LSBC (45).
% Women with high school degree or higher (age 25+)			Women from neighborhoods with less educated people may be at higher risk of LSBC (45, 46).
% People spoke English less than "very well" (age 18+)			Language may be a barrier to breast cancer screening (46-48).
% Women in management/ business/science/arts occupations (age 16+)			Occupation categories are associated with breast cancer stage at diagnosis (49, 50).
% Women in service occupations (age 16+)			
% Women in sales and office occupations (age 16+)			
% Women in labor intensive occupations (age 16+)			
% Renter occupied households			Women living in area with higher rates of renter occupied households are more likely to be diagnosed with LSBC (48).

Variable	Year(s)	Source	Conceptual Reason for Inclusion
% People moved residency in the past year			High frequency residential change is potentially a marker for the clinical risk of behavioral and emotional problems (51).
% Women worker drove alone to work (age 16+)			Percentage of women driving alone to work is an indicator of vehicle availability, which is an indicator of spatial access to screening services (3).
% Women worker with >= 30 min travel time to work (age 16+)			Travel time to work may be an indicator of proximity to urban centers where most screening services are located, which in turn may be associated with cancer stage at diagnosis (6-9).
% Women (age 15-50) had a birth in the past 12 months			After a childbirth, mothers experience a transient increased risk of late-stage breast cancer (52).
% Women unemployed among those in labor force (age 16+)			Women living in area with higher rates of unemployment were more likely to be diagnosed with LSBC (48).
Area Deprivation Index	2014	R Package ‘Sociome’	Neighborhood deprivation along various aspects of social determinants of health may be associated with LSBC (4, 5).

SEER = Surveillance, Epidemiology, and End Results; BRFSS = Behavioral Risk Factor Surveillance System; FDA = Food and Drug Administration; AHRF = Area Health Resources Files; CHR: County Health Rankings & Roadmaps; ACS = American Community Survey

Table 2.

Characteristics of phenotypes and prevalent regions. Under Prevalent Regions, the listing of states in multiple phenotypes refers to different counties within state, not overlapping areas.

Phenotype (median % LSBC)	Characteristics associated with LSBC	Prevalent Regions
LR 1 (30.6%)	Lower uninsured (≤11.6%), higher use of mammography (>68.1%)	Massachusetts, New York, Connecticut, Iowa
LR 2 (32.6%)	Lower uninsured (≤11.6%), lower use of mammography (≤68.1%), urban area	New York, New Jersey, Connecticut, Kentucky, Hawaii, California (San Francisco Bay), Iowa, Utah, Seattle Puget Sound, Detroit
LR 3 (34.5%)	Lower uninsured (≤11.6%), lower use of mammography (≤68.1%), rural area	Kentucky, Iowa, New York, Hawaii
MR 1 (35.5%)	Higher uninsured (>11.6%), lower area deprivation (≤99.7)	Utah, Idaho, California, New Jersey, Georgia, Louisiana, New Mexico
HR 1 (37.0%)	Higher uninsured (>11.6%), higher area deprivation (>99.7), lower poverty (≤26.1%), lower per capita income (≤ $32,946)	Georgia, Kentucky, Louisiana, New Mexico
HR 2 (38.4%)	Higher uninsured (>11.6%), higher area deprivation (>99.7), lower poverty (≤26.1%), higher per capita income (> $32,946)	Louisiana, California, New Mexico, Idaho, Georgia, New York (Kings County)
HR 3 (40.1%)	Higher uninsured (>11.6%), higher area deprivation (>99.7), higher poverty (>26.1%)	Georgia, Kentucky, Louisiana, New Mexico, New York (Bronx County)