A Fine-grain Geospatial and Demographic Analysis of Breast Cancer Patterns in New York State

Fiona Murphy
Dept. of Biomedical Informatics
Stony Brook University
Stony Brook, NY, USA
fiona.murphy@stonybrook.edu

Kayley Abell-Hart
Dept. of Biomedical Informatics
Stony Brook University
Stony Brook, NY, USA
kayley.abell-hart@stonybrook.edu

Fusheng Wang
Dept. of Biomedical Informatics, Dept. of Computer Science
Stony Brook University
Stony Brook, NY, USA
fusheng.wang@stonybrook.edu

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9671852

Abstract— **Cancer impacts many people across New York State, with breast cancer being one type of cancer that is especially prevalent. Breast cancer can be detected through screening procedures, and if caught early in its development, has a higher likelihood of treatability and a less severe outcome. In order to better understand existing patterns of breast cancer throughout New York State, we conducted a geospatial and demographic analysis in which we sought to identify disparities in screening, diagnosis, and mortality rates across regions and demographic groups. Statistical analyses were completed using patient-level data from the New York Statewide Planning and Research Cooperative System (SPARCS) from 2005 to 2019, where demographic data including age, race, and ethnicity, as well as location-based data such as addresses were collected. Geospatial analysis results revealed clustering for screening and diagnosis rates in certain regions of the state, as well as hot spots. Results from the demographic analysis indicated temporal trends in cancer rates for various age, race, and ethnic groups, as well as a disparity in the mortality rates of breast cancer across race and ethnic groups.**

Keywords— breast cancer, geospatial, demographic, mortality, health disparities

I. Introduction

Within New York State, cancer prevails as the second leading cause of death, with roughly 35,000 New Yorkers dying from cancer each year [1]. Among all types of cancer affecting either sex, breast cancer has the highest incidence in the state. For females in New York, breast cancer proves to be a great burden with the second highest mortality rate of any cancer type [1]. In order to reduce the severity of breast cancer's progression and increase its ability to be treated, it is recommended that women receive routine screening in the form of mammograms. Early detection of breast cancer from screening procedures has been found to decrease mortality rates, especially among older age groups in which breast cancer is more prevalent [2].

Providing individual communities with insight into local trends regarding breast cancer screening and diagnosis rates allows these communities to implement targeted intervention strategies. More impactful than informing communities about broad-scale trends at the county-level or across large regions of the state is a fine-grain geospatial analysis of breast cancer patterns, down to the census-tract level. This high spatial resolution can provide a detailed look at disparities in breast cancer rates, signifying locations that may be lacking access to screening or treatment resources, and thus are in need of greater support. Additionally, there is an urgent need for communities to understand disparities in access to screening, diagnostics, and treatment across demographic groups such as age, race, and ethnicity. Again, having a better understanding of which groups of people are disproportionately screened for or are dying from breast cancer can allow local governments and public health institutions to take more targeted actions.

Large-scale patient level data allows for this analysis of breast cancer trends across New York State. The New York State Department of Health Statewide Planning and Research Cooperative System (SPARCS) [3] provides visit-level data for patients who have used inpatient, outpatient, emergency room, and ambulatory services across the entire state. Relevant information for our analyses provided in this database include diagnoses, treatments, services, charges, patient addresses, and demographic data such as age, race, and ethnicity from 2005 through 2019. SPARCS also contains data related to patient mortality status, including patient disposition, All Patient Refined Risk of Mortality (APR ROM) values, and All Patient Refined Severity of Illness (APR SOI) values.

Combining the use of this large-scale patient-level data with our high resolution spatial and demographic analysis techniques, we were able to perform a specialized investigation into the patterns of breast cancer screening, diagnosis, and mortality rates across New York State. Our investigation included an assessment of census-tract level variation in breast cancer screening and diagnosis rates, an analysis of temporal trends for screenings and diagnoses, and an evaluation of differences in screening, diagnosis, and mortality rates across demographics. Identifying these patterns can contribute valuable information that communities can wield to eliminate the burden of breast cancer felt by thousands of women across New York State.

978-1-6654-3902-2/21/$31.00 ©2021 IEEE

II. METHODS

A. Data Source

This study utilized patient-level data from New York State's SPARCS database, an all-payer data reporting system that collects inpatient data from every Article 28 hospital, as well as outpatient data from ambulatory surgery services, emergency departments, and other outpatient clinic services [3]. All data were used with the approval of and in accordance with guidelines established by Stony Brook University's Institutional Review Board and the Office of Quality and Patient Safety, and New York State's Department of Health.

For this study, patient data were obtained from 2005 to 2019. In order to identify patients diagnosed with the target cancers or who had participated in selected screening procedures, ICD-9, ICD-10, and CPT codes were used. Data obtained from SPARCS included patient addresses, demographic characteristics, including age, race and ethnicity, and patient mortality data through values such as patient disposition, APR ROM values, and APR SOI values.

B. Demographic Statistical Analysis

The demographic statistical analysis component of our investigation was conducted using Microsoft Excel [8] and population data from the American Community Survey of the US Census Bureau [9], including the creation of plots and graphs. Linear regression tests were performed in order to evaluate linear temporal trends over the timeframe the data was collected from, with statistically significant results having a p-value less than 0.05. Chi-squared tests were performed to determine statistical differences in total rates of breast cancer diagnosis and screening among different demographics, with p-values less than 0.05 denoting a significant difference between groups.

C. Geospatial Analysis

The geospatial analysis component of our investigation was conducted using the addresses of target patients. These records had previously been aggregated and then filtered for a singular discharge record per patient for diagnoses and unique claim instances (one record per procedure) for screenings, so as to avoid over-counting records. Using Stony Brook University's EaserGeocoder [4], patient addresses were converted into a latitude and longitude coordinate, after which these coordinates were related to their corresponding census tracts using PostGIS [5]. TIGER/Line [6] and population data [9] from the US Census Bureau and ArcGIS [7] were used to produce all maps, as well as perform spatial analyses. The Global Moran's I index, which measures the spatial autocorrelation of an attribute on a scale of -1.0 to 1.0, was identified in order to determine whether breast cancer screening and diagnosis rates were clustered by similar rate. Additionally, hot spots and cold spots were identified for breast cancer screening and diagnosis rates as census tracts with significantly high or low rates, respectively, that were clustered with other census tracts that had significantly high or low rates.

III. RESULTS

Records of female patients with breast cancer diagnosis codes or breast cancer screening procedure codes were collected from SPARCS and aggregated by year, as well as by age group, race, and ethnicity. These data counts were converted into rates per 100,000 population using yearly population estimates from the US Census Bureau [10], [11]. From 2005 to 2019, there were 2,469,500 unique instances of breast cancer screenings (multiple individual screenings per patient were collected), producing a rate of 24,653.1 breast cancer screenings per 100,000 population. The yearly totals reflected an increase in screenings over this time period ($p < 0.0005$) (Fig. 1). Over the same time period of 2005 to 2019, 345,569 unique patient records were identified with having a breast cancer diagnosis, yielding a rate of 3,449.8 breast cancer diagnoses per 100,000 population, with yearly trends showing an increase in breast cancer diagnoses ($p < 0.0005$) (Fig. 2).

A. Temporal Trend Analysis

Within each of the four upper age brackets (35-44, 45-54, 55-64, 64+), there was a significant increase in breast cancer screenings from 2005 to 2019 ($p < 0.001$), while the lower three age brackets (0-14, 15-24, 25-34) showed only a slight positive trend in breast cancer screenings ($p < 0.001$) (Fig. 3). Note that there was a spike in the reporting rates of SPARCS in 2011, which may explain the jump in screening rates in 2011. Additionally, the most recent data release from SPARCS with patient records collected from 2016 to 2019 contains roughly double the amount of records as the old SPARCS data batch, which could again be a possible explanation for the jump in screening rates from 2015 to 2016. For all race and ethnic groups, there was a moderate positive trend in breast cancer screenings from 2005 to 2017 ($p < 0.05$) (Fig. 4).

With respect to breast cancer diagnoses, there was a

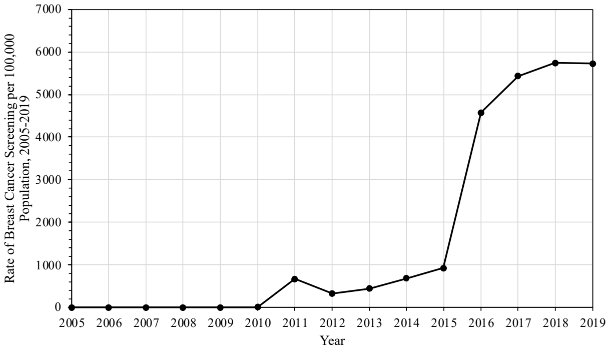


Fig. 1. Temporal trend for overall rate of breast cancer screening per 100,000 population, 2005-2019.

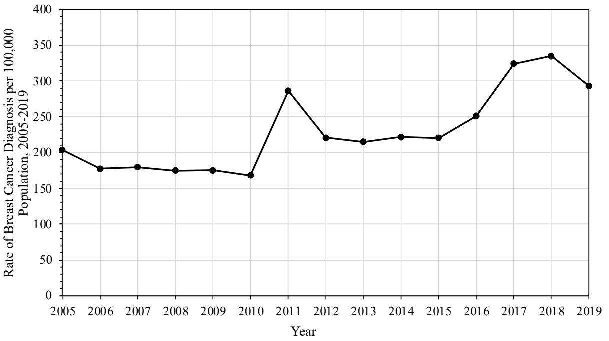


Fig. 2. Temporal trend for overall rate of breast cancer diagnosis per 100,000 population, 2005-2019.

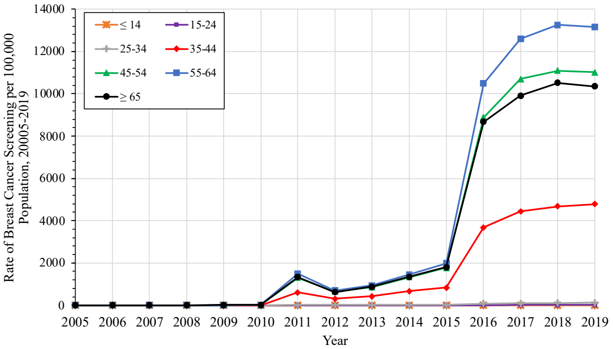


Fig. 3. Temporal trends for breast cancer screening rate per 100,000 population by age group, 2005-2019.

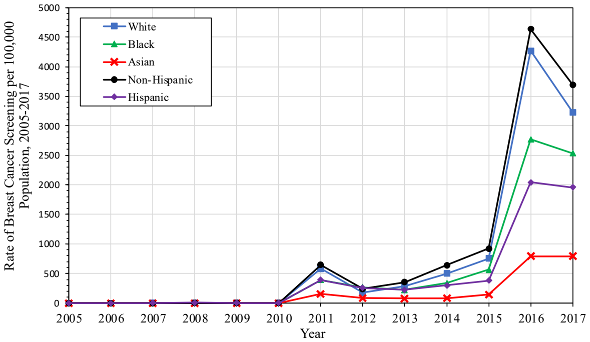


Fig. 4. Temporal trends for breast cancer screening rate per 100,000 population by race and ethnic group, 2005-2017.

significant increase in diagnosis rates for the upper three age groups (45-54, 55-64, 65+) ($p < 0.005$), and a slight, but statistically significant, increase in diagnosis rates for every other age group ($p < 0.005$) (Fig. 5). Asian women were the only racial or ethnic group to show a statistically significant increase in breast cancer diagnosis rates from 2005 to 2017 ($p < 0.05$) (Fig. 6) However, fluctuations in 2011 and/or 2017 might be unduly influencing these trends.

B. Disparities Among Race and Ethnic Groups

There was a significant difference in the rate of breast cancer screening between each race and between each ethnicity ($p < 0.0001$), with White women having the highest rate of breast cancer screening among all racial groups, and Non-Hispanic women having the highest rate of breast cancer screening among the two ethnic groups (Fig. 7). Similar findings were revealed for breast cancer diagnosis rates, where there was a significant difference in these rates between each racial group and each ethnic group ($p < 0.0001$) (Fig. 8). Also, the same pattern was found where White women and Non-Hispanic women had the highest rates of breast cancer diagnoses (Fig. 8).

Three metrics were collected from SPARCS to evaluate breast cancer mortality rates — patient disposition, APR ROM (Risk of Mortality) values, and APR SOI (Severity of Illness) values. Patient disposition reveals actual deaths, while ROM and SOI values can help to indicate a more severe progression of breast cancer in a patient. ROM and SOI values are both divided into 4 categories, where a value of “1” represents a minor level, “2” represents a moderate level, “3” a major level, and “4” an extreme level [12]. Among all races and ethnicities, Black women were found to have the highest percentage of deaths, with 6.7% of Black female breast cancer patients dying while in a SPARCS-reporting facility or service (Fig. 10). This percentage of deaths in Black female breast cancer patients is more than double that of White or Asian women, who had 3.2% and 3.1% of patients die, respectively. Hispanic women had a higher rate of deaths amongst breast cancer patients (4.4%) than Non-Hispanic women (3.7%). We also observed extremely left-shifted distributions in the ROM and SOI values of White, Asian, and Non-Hispanic women, revealing that most of these

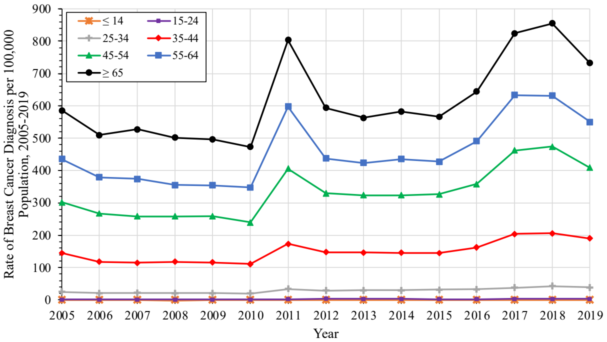


Fig. 5. Temporal trends for breast cancer diagnosis rate per 100,000 population by age group, 2005-2019.

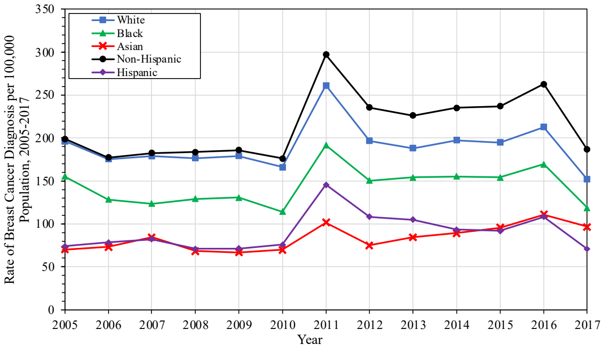


Fig. 6. Temporal trends for breast cancer diagnosis rate per 100,000 population by race and ethnic group, 2005-2017.

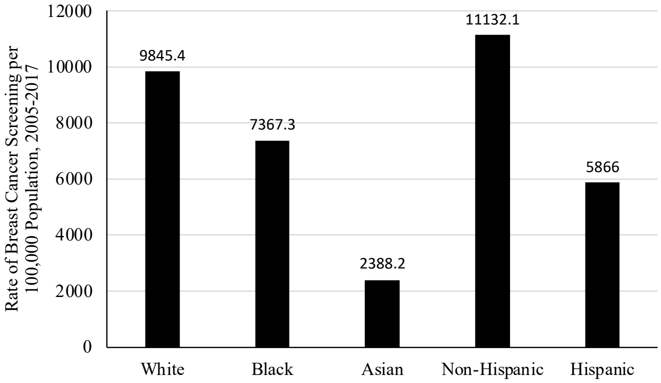


Fig. 7. Rate of breast cancer screening per 100,000 population by patient race and ethnic group, 2005-2017.

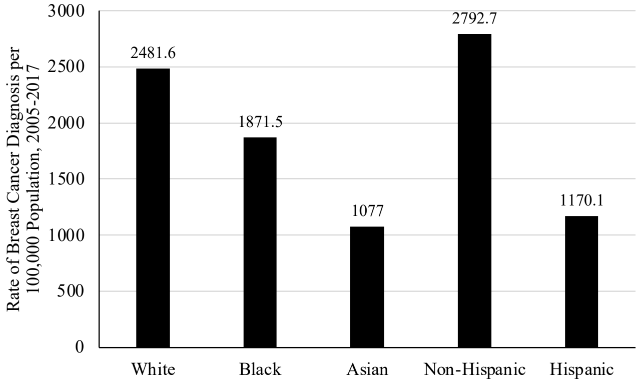


Fig. 8. Rate of breast cancer diagnosis per 100,000 population by patient race and ethnic group, 2005-2017.

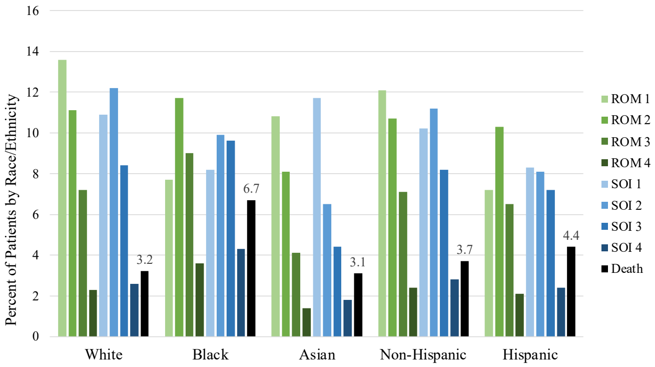


Fig. 10. Percentage of breast cancer patients by race and ethnic group with each Risk of Mortality (ROM) and Severity of Illness (SOI) value, and percentage of deaths, 2005-2019.

patients had a mild or moderate risk of mortality or severity of illness. On the other hand, Black and Hispanic women had more right-shifted distributions for these values, indicating more instances of a high risk of mortality or severity of illness among these demographic groups.

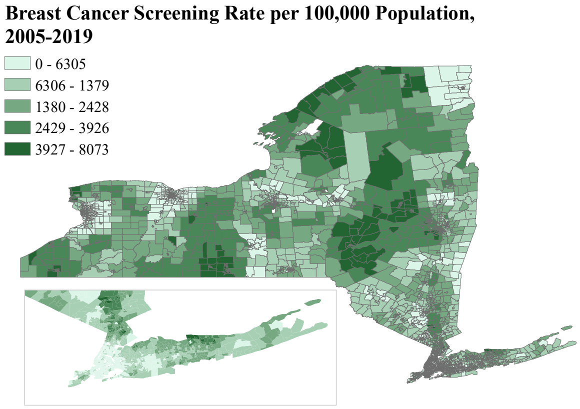
C. *Geospatial Analysis*

Along with a demographic analysis, we performed a geospatial analysis of breast cancer screening and diagnosis rates at a high spatial resolution across New York State. In the map of breast cancer screening rates, the Global Moran's I index is 0.500 ($p < 0.0001$), indicating high spatial autocorrelation. This means that census tracts with high rates of breast cancer screening are clustered together and census tracts with low screening rates are likewise clustered together (Fig. 9). We observed hot spots in central and northern New York, including many rural areas, with cold spots for screening around Buffalo and Rochester (Fig. 12). There was also a cold spot for breast cancer screening in New York City, directly adjacent to screening hot spots in Suffolk County on Long Island and the Westchester region above New York City (Fig. 12).

In the map of breast cancer diagnosis rates, the Global Moran's I Index is 0.204 ($p < 0.0001$), suggesting minimal spatial clustering (Fig. 11). Hot spots were identified around every major city in New York State except for New York City, which was identified as a cold spot (Fig. 13).

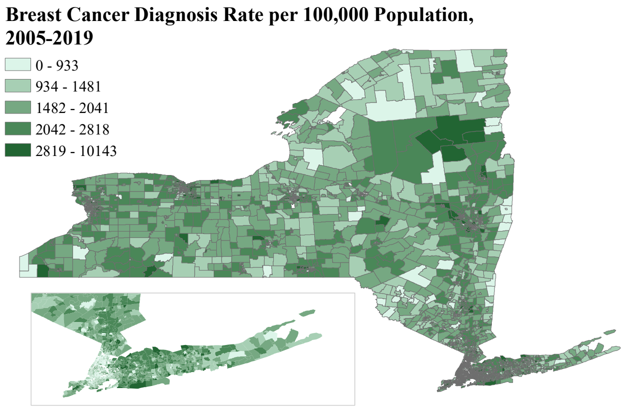
IV. Discussion

An important result from our demographic analysis is a disparity existing between diagnosis and mortality rates across different race and ethnic groups. While Black women had a lower rate of breast cancer diagnosis than White women, the percentage of deaths among Black female breast cancer patients is more than double that of White breast cancer patients. Also, with right-shifted distributions, Black and Hispanic women had greater rates of high values for risk of mortality and severity of illness. These findings support national trends in which White women have the highest breast cancer incidence rate, yet Black women have the highest rate of mortality from breast cancer [13], [14]. This information is important for local governments and health care systems, as there could be inequality in access to screening, diagnosis, or breast cancer treatments among racial and ethnic groups. The difference in screening rates among every racial and ethnic group reveals that local institutions could make targeted efforts to increase screening rates in demographic groups with lower rates of screening in order to potentially curb



Breast Cancer Screening Rate per 100,000 Population, 2005-2019
0 - 6305
6306 - 1379
1380 - 2428
2429 - 3926
3927 - 8073

Fig. 9. Map of breast cancer screening rate per 100,000 population by census tract in New York State.



Breast Cancer Diagnosis Rate per 100,000 Population, 2005-2019
0 - 933
934 - 1481
1482 - 2041
2042 - 2818
2819 - 10143

Fig. 11. Map of breast cancer diagnosis rate per 100,000 population by census tract in New York State.

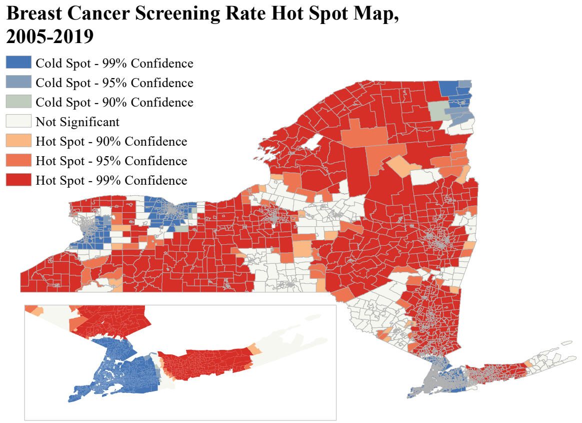

Breast Cancer Screening Rate Hot Spot Map, 2005-2019
Cold Spot - 99% Confidence
Cold Spot - 95% Confidence
Cold Spot - 90% Confidence
Not Significant
Hot Spot - 90% Confidence
Hot Spot - 95% Confidence
Hot Spot - 99% Confidence

Fig. 12. Map of breast cancer screening hot spots (and cold spots) by census tract in New York State.

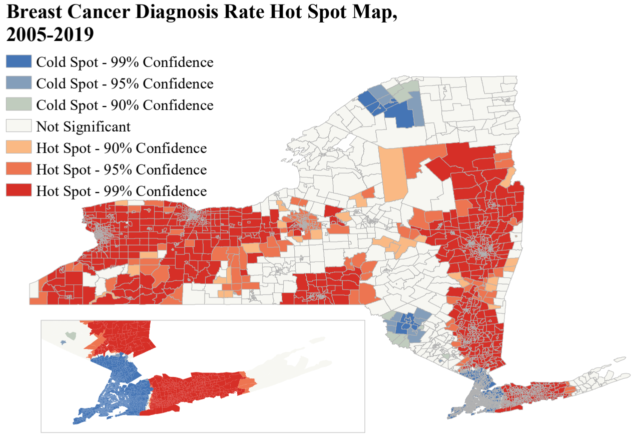

Breast Cancer Diagnosis Rate Hot Spot Map, 2005-2019
Cold Spot - 99% Confidence
Cold Spot - 95% Confidence
Cold Spot - 90% Confidence
Not Significant
Hot Spot - 90% Confidence
Hot Spot - 95% Confidence
Hot Spot - 99% Confidence

Fig. 13. Map of breast cancer diagnosis hot spots (and cold spots) by census tract in New York State.

disparities in mortality rates. Screening is important for early detection and quick treatment in order to inhibit the progression of breast cancer.

The location of hot spots (and inversely cold spots) can also be of great use to individual communities. Buffalo and Rochester were cold spots for screening rates, yet hot spots for diagnosis rates, suggesting that there may be a need to increase breast cancer screening rates in these areas. These maps are also useful to identify location-based disparities in screening and diagnosis patterns. New York City was a cold spot with significantly lower rates for both screenings and diagnoses; however, hot spots surrounded the city on both sides. This could suggest that New York City is not screening enough for breast cancer, which in turn is accounting for the low rates of diagnosis as breast cancer cases go undetected.

With the availability of a high resolution spatial analysis, local governments can identify disparities in breast cancer screening and diagnosis rates down to the census tract. In turn, governments and healthcare institutions can work together to implement highly refined community outreach that increases screening rates, with the objective of decreasing the incidence of advanced breast cancer.

V. Study Limitations

There were some limitations to this study resulting from inconsistencies in the data made available from SPARCS. Data regarding patient race and ethnicity were only available through September of 2017, preventing a more recent and complete analysis of trends regarding patient race and ethnicity. Additionally, there were some jumps in SPARCS data reporting, such as in 2011 and from 2016 onwards, which affected temporal trends.

SPARCS also only collects data on patients receiving designated services, meaning that there is likely missing data from some of our analyses, such as breast cancer patients who died at home, and thus are not counted towards mortality rates.

VI. Conclusion

In order to better address the burden of cancer across New York State, detailed analysis is required across geospatial regions as well as demographics. With recently available large-scale data from SPARCS and fine-grain geospatial analysis techniques, patterns of cancer across the state can be observed at a highly refined level. Identifying these patterns at a high spatial resolution can provide communities with greater knowledge about the specific impact that breast cancer has in their area, as opposed to a more broad-scale analysis, which may be less informative for local communities. This information can potentially make targeted interventions possible, such as outreach efforts to increase breast cancer screening in areas with low screening rates, right down to the census tract. As such, the use of high resolution analysis techniques coupled with the availability of large-scale data can lead to greater advancements in our awareness of cancer patterns across New York State at the community level, and thus greater improvements in breast cancer detection, treatment, and prevention.

Acknowledgment

This research is supported in part by grants from National Science Foundation ACI 1443054 and IIS 1541063, and National Institute of Health U01CA242936.

References

[1] *NYS Cancer Registry*. https://www.health.ny.gov/statistics/cancer/registry/. Accessed 30 July 2021.

[2] Humphrey, Linda L., et al. "Breast Cancer Screening: A Summary of the Evidence for the U.S. Preventive Services Task Force." *Annals of Internal Medicine*, vol. 137, no. 5_Part_1, American College of Physicians, Sept. 2002, pp. 347–60. *acpjournals.org (Atypon)*, https://doi.org/10.7326/0003-4819-137-5_Part_1-200209030-00012.

[3] Bureau of Health Informatics. SPARCS Operations Guide. (NYS Department of Health, 2016).

[4] Rashidian, Sina, et al. EaserGeocoder: integrative geocoding with machine learning. In Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information

Systems (SIGSPATIAL '18). ACM, New York, NY, USA, 572-575. 2018, https://doi.org/10.1145/3274895.3274929.

[5] PostGIS, https://postgis.net/.

[6] US Census Bureau. *Tiger/Line data*, https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html, 2018.

[7] ArcGIS 10.8, https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview?rsource =%2Fsoftware%2Farcgis%2Farcgis-for-desktop.

[8] Microsoft Excel 16.51, https://www.microsoft.com/en-us/microsoft-365/excel.

[9] Bureau, US Census. "American Community Survey Data." *The United States Census Bureau*, https://www.census.gov/programs-surveys/acs/data.html.

[10] Bureau, US Census. "2019 Population Estimates by Age, Sex, Race and Hispanic Origin." *The United States Census Bureau*, https://www.census.gov/newsroom/press-kits/2020/population-estimates-detailed.html.

[11] Bureau, US Census. "State Intercensal Tables: 2000-2010." *The United States Census Bureau*, https://www.census.gov/data/tables/time-series/demo/popest/intercensal-2000-2010-state.html.

[12] *SPARCS-Inpatient and Outpatient Output Data Dictionaries*. https://www.health.ny.gov/statistics/sparcs/datadic.htm.

[13] Yedjou, Clement G., et al. "Assessing the Racial and Ethnic Disparities in Breast Cancer Mortality in the United States." *International Journal of Environmental Research and Public Health*, vol. 14, no. 5, May 2017, p. 486. *PubMed Central*, https://doi.org/10.3390/ijerph14050486.

[14] Shoemaker, Meredith L., et al. "Differences in Breast Cancer Incidence among Young Women Aged 20–49 Years by Stage and Tumor Characteristics, Age, Race, and Ethnicity, 2004–2013." *Breast Cancer Research and Treatment*, vol. 169, no. 3, June 2018, pp. 595–606. *PubMed Central*, https://doi.org/10.1007/s10549-018-4699-9.