Addressing Health Disparities in Preventive Cancer Screenings In California Using Data Visualization & Clustering

Author: Winnie Taiwo

I. Introduction

Access to preventive cancer screenings is a critical component of reducing cancer-related mortality, yet significant racial and socioeconomic disparities persist, particularly in underserved communities in California. These disparities often lead to delayed diagnoses and poor health outcomes, disproportionately affecting vulnerable populations. While national datasets like BRFSS and SEER have provided valuable insights into these disparities, their lack of neighborhood-level granularity limits the ability to identify and address local disparities effectively. This project seeks to bridge this gap by integrating health, socioeconomic, and geographic data to provide a nuanced analysis of disparities in access to cancer screenings.

Leveraging machine learning and geospatial analysis, we aim to uncover complex relationships between income, geography, and healthcare access at the neighborhood level. By employing clustering algorithms, we identify underserved areas and offer data-driven recommendations for the placement of new cancer screening centers. Our methodology highlights disparities and provides actionable insights to guide healthcare policy and resource allocation. This innovative approach goes beyond traditional statistical methods, enabling a deeper understanding of the factors contributing to inequities in preventive care. We hope to facilitate the development of targeted interventions that improve access to cancer screenings, ultimately reducing health disparities and improving outcomes for underserved populations in California.

II. Problem Definition

Precise formal definition

Racial and socioeconomic disparities in California significantly impact access to preventive cancer screenings, leading to delayed diagnoses and poor health outcomes for underserved populations (Smedley, Stith, & Nelson, 2003; American Cancer Society, 2021). Existing studies utilizing national datasets like BRFSS and SEER often need more neighborhood-level granularity, which limits the understanding of local disparities (Jones, 2013; Freeman, 2004). This project aims to fill this gap by leveraging a comprehensive dataset that integrates health, socioeconomic, and geographic data to analyze how income levels and geographic location influence access to cancer screenings. By employing clustering algorithms, we will identify patterns of healthcare access disparities, pinpoint underserved areas, and provide precise, actionable recommendations for improving preventive care availability in California (Siegel, Miller, & Jemal, 2020; Dwyer-Lindgren et al., 2017).

Jargon-free definition

Preventive cancer screenings save lives, but not everyone in California has equal access. Communities with lower incomes and fewer healthcare resources often experience delayed diagnoses and worse health outcomes (Smedley, Stith, & Nelson, 2003; American Cancer

Society, 2021). While national studies like those using BRFSS and SEER data help highlight these issues, they don't provide enough detail to understand what's happening in specific neighborhoods (Jones, 2013; Freeman, 2004). This project will use health, socioeconomic status, and geography data to uncover which areas have the greatest need for cancer screening services. Using clustering methods, we will recommend where new screening centers should be placed to address these disparities (Siegel, Miller, & Jemal, 2020; Dwyer-Lindgren et al., 2017).

III. Literature Survey

Research on disparities in healthcare access has consistently highlighted the critical role of socioeconomic and racial factors in influencing outcomes, particularly in preventive care. Smedley, Stith, and Nelson (2003) highlight systemic biases and socioeconomic barriers as key drivers of racial and ethnic disparities in healthcare, while Freeman (2004) links poverty and social injustice to disparities in cancer outcomes. Siegel, Miller, and Jemal (2020) emphasize the socioeconomic divide in cancer incidence and mortality, calling for targeted interventions. Despite these insights, the studies lack actionable, localized solutions to address specific underserved areas.

Bowser et al. (2024) and Recchia et al. (2022) employ clustering methods to identify patterns in healthcare utilization, demonstrating the potential of machine learning to uncover disparities. Grote and Keeling (2022) discuss ethical considerations in applying these algorithms, emphasizing the need for equitable designs. Despite these advancements, only some studies integrate clustering with diverse datasets encompassing socioeconomic, health, and geographic factors. Lastly, Tipton et al. (2023) and Qi et al. (2023) assess healthcare algorithm impacts on disparities but focus on broad policy-level insights, leaving gaps in practical neighborhood-level interventions.

Ruggles (2018) and Codex (2024) explore the use of visualization tools such as D3.js for geospatial analysis and advocate for integrating interactive visualizations to enhance stakeholder engagement. Moujahid (2016) provides technical guidance for visualizing geospatial data but needs to connect these tools to actionable healthcare interventions.

IV. Proposed Method

A. Intuition

The existing research work which focuses on health disparities sometimes seems not to pay attention to very important spatial inequalities that exist at the neighborhood level, as most of it is based on aggregate data (Freeman, 2004; Schootman et al., 2017). There is a lack of integration of health, socioeconomic as well as geographical data which can be used for clustering and analysis and this makes it difficult to explore intricate and multidimensional patterns of inequity (Bowser et al., 2024; Grote & Keeling, 2022). Furthermore, while some literature may identify disparities, there are few such data based practical recommendations on the allocation of resources for setting up new screening sites (Chen & McBride, 2021).

As such, in the paper, we apply available datasets integrating cancer screening in relation to spatial factors to fill this gap and provide insights into factors that are the root causes of access

disparities. For example BRFSS and SEER studies employing such large datasets usually fail to pinpoint the neighborhoods having the targeted underserved populations and as such are not able to sufficiently characterize local differences in access to care among the underserved communities. These limitations make it impossible for policymakers to comprehend the complex and intricate relationship between socioeconomic and geographic factors, and the access to preventive healthcare services. Unlike the aforementioned methods, our method delivers the much needed detail in the analysis of disparities by combining health, social, economic and geographic variables sourced from the CDC and NIH.

Also, our project builds on what previous studies did not touch on since it tackles issues in oncology and utilizes the advantage of machine learning combined with geospatial tools. Doing so permits us to optimally place new cancer screening centers. Gaps in the existing techniques are addressed and a hyper-localized approach is provided. As such, our framework enhances the precision of disparity identification and equips decision-makers with practical, geography-specific solutions for communities in California.

B. Detailed Approach

Algorithms

1. *Data Preprocessing*
   The data pipeline began with filtering raw CDC datasets to focus on neoplasm-related deaths (ICD codes C00-D48) grouped by individual California counties. These datasets were then merged with socioeconomic data, such as poverty rates, uninsured populations, education levels, and median income for each county. Geographic coordinates were appended for spatial analysis. To ensure consistency, redundant columns were removed, missing values handled by removing rows or replacing missing values with a zero when applicable, and features were standardized using Min-Max scaling.
2. *Clustering*
   K-means clustering was applied to group counties with similar socioeconomic and health profiles. The initial clustering used three clusters, later optimized to eight clusters using the elbow method. Validation metrics, including a silhouette score of 0.286 and a Davies–Bouldin Index of 0.871, confirmed the clustering quality. Dimensionality reduction techniques, such as PCA and t-SNE, enabled effective visualization of cluster separations.
3. *Ranking*
   A ranking algorithm prioritized clusters based on their need for cancer screening resources. Factors like uninsured rates, poverty levels, and cancer deaths were normalized and assigned weights. A composite score calculated for each cluster identified the top three regions requiring immediate attention.

User Interfaces

Interactive Map
Using the Folium library, an interactive map was developed to visualize clusters geographically. Counties were marked with color-coded icons, each displaying cancer deaths, population, and socioeconomic indicators in a popup. The map was exported as an HTML file for accessibility and stakeholder review.

Cluster Statistics Dashboard
Dashboards created with Pandas and Matplotlib provided insights into cluster characteristics through charts and tables. Visualizations highlighted disparities in income, education, and healthcare access, making the data more comprehensible for decision-makers.

V. Experiments/Evaluation

A. Testbed Description

The experimental setup was designed to identify regions in California most in need of additional cancer screening resources, based on an analysis of cancer mortality and socioeconomic indicators. The data source was the CDC's "Underlying Cause of Death" dataset for 2018–2022, filtered for neoplasm-related deaths (ICD codes C00-D48) across all California counties. This dataset was further enriched with socioeconomic factors such as uninsured rates, income levels, poverty rates, education levels, and food quality indices, for each California county. We added geographic coordinates corresponding to each county for spatial visualization. These features were then standardized using Min-Max scaling to ensure comparability. K-means clustering was employed to group counties into clusters with similar socioeconomic and health profiles. Validation methods, including the silhouette score and the Davies–Bouldin Index, were applied to ensure the robustness of the clustering results. Overall, the experiments aimed to answer the following questions:

1.What socioeconomic and demographic factors characterize counties with high cancer mortality?	**3. What do geographic disparities reveal about cancer mortality and socioeconomic factors?**
2. How can counties be clustered by health and socioeconomic profiles?	**4. Which clusters and counties most need additional cancer screening resources?**

The experimental setup provided a systematic way to analyze complex, multidimensional data, answering critical questions about the intersection of healthcare access, socioeconomic conditions, and cancer outcomes. By integrating clustering, ranking, and visualization, the study offered actionable insights for improving cancer screening resource allocation.

B. Experiments and Observations

The experiments focused on analyzing cancer-related mortality across California counties from 2018 to 2022 to identify regions most in need of additional cancer screening resources. Data from the CDC was filtered to include only neoplasm-related deaths (ICD codes C00-D48) and further enriched with socioeconomic indicators, including population, uninsured rates, income levels, poverty rates, and education levels. Geographic coordinates were added to enable spatial analysis. After preprocessing, the data underwent extensive cleaning and normalization using Min-Max scaling, ensuring consistency across variables like cancer deaths, income, and socioeconomic factors.

Initial clustering was performed using K-means, starting with three clusters to provide a broad categorization of counties. Subsequently, the elbow method was applied to determine the optimal number of clusters, which was found to be eight. Validation metrics, including the silhouette score (0.286) and the Davies–Bouldin Index (0.871), indicated moderate cluster separation and compactness. The clustering metrics reflect acceptable but not optimal performance for the dataset. These clusters were visualized through PCA and t-SNE plots, which revealed distinct groupings based on socioeconomic and health-related attributes. Cluster summaries highlighted the unique characteristics of each group. For instance, Cluster 0 had high uninsured rates (10.78%), significant poverty levels, while Cluster 6, represented by Los Angeles County, exhibited extremely high cancer mortality (14,196 deaths) and uninsured populations. Conversely, Cluster 1 included counties with lower poverty levels (6.94%) and higher education and income levels, indicative of better access to healthcare and resources.[20]

A detailed ranking system was developed to prioritize clusters based on their need for additional cancer screening resources. The ranking algorithm assigned weights to critical factors, including poverty rates, uninsured percentages, and cancer mortality, reflecting the multifaceted nature of healthcare inequities. Clusters with higher composite scores, such as Cluster 0, Cluster 6, and Cluster 4, were identified as top priorities [20]. Cluster 0 included medium-sized counties with a combination of high uninsured rates and poverty, while Cluster 6, dominated by Los Angeles County, stood out due to its overwhelming population size and mortality burden. Cluster 4, comprising smaller counties, displayed moderate mortality but notable socioeconomic vulnerabilities.

We experimented using different weights when calculating composite scores. However, clusters containing counties with large metropolitan areas often had higher poverty rates, uninsured percentages, and generally worse socioeconomic factors. They remained the top clusters for additional cancer screening sites regardless of which factors we weighed higher than others.

The analysis also involved creating interactive maps using Folium, where markers color-coded by cluster highlighted counties' cancer death rates, population size, and socioeconomic indicators. These visualizations allowed for an intuitive understanding of geographic disparities and clusters' healthcare needs. For example, counties in Cluster 0 and Cluster 6 were prominently displayed with high cancer mortality and socioeconomic deprivation, underscoring their need for intervention.

Further analysis of the top-ranked clusters delved into specific socioeconomic and health disparities. Counties were reassessed to examine poverty rates, educational attainment, and language isolation, providing granular insights into their unique challenges. This deeper examination confirmed the priority areas and informed actionable recommendations. For instance, counties in Cluster 0 could benefit from targeted community outreach and expanded insurance coverage initiatives, while Los Angeles County might require large-scale infrastructure improvements to address its population's healthcare demands.

The study demonstrated a comprehensive approach to understanding cancer mortality through clustering, ranking, and visualization techniques. The findings underscored the critical interplay

between socioeconomic factors and healthcare access, providing a data-driven framework for prioritizing cancer screening resources in underserved regions.

VI. Conclusion and Discussion

This project highlights the critical role of data-driven approaches in addressing racial and socioeconomic disparities in access to preventive cancer screenings across California. By integrating health, socioeconomic, and geographic data, we have developed a comprehensive framework to identify and prioritize underserved areas for targeted interventions. The use of clustering algorithms allowed us to uncover nuanced patterns of healthcare inequity, grouping counties with similar socioeconomic and health profiles, and pinpointing regions most in need of additional resources.

One of the key takeaways from our project is that counties that experience high poverty, low-income, and lower educational achievements are correlated to clusters with low cancer screening sites. This is alarming and should be addressed by the leaders of these counties for the overall health of their citizens. Those highly at risk for late-stage diagnosis programs in place to assist them, starting with placing cancer screening centers in their neighborhoods. Additionally, our geospatial visualizations made it easier to see the geographic disparities clearly, providing actionable insights that can help policymakers make smarter decisions about where to allocate resources.

The project's potential impact is profound. By addressing gaps in existing studies and delivering granular, localized insights, this work provides a roadmap for improving healthcare equity. Recommendations for new screening centers, community outreach programs, and infrastructure enhancements are tailored to the specific needs of underserved populations, ultimately aiming to reduce cancer-related mortality and improve health outcomes.

This project could be improved on in a few ways. Firstly, fine-tuning the clustering algorithms would help increase the performance of the model; ideally we would have wanted an ideal score for both the Silhouette and Davies-Bouldin Index Score to be above a 0.5 which could be achieved with feature scaling and further dimensionality reduction. Another method is in using different sets of data. There are many different socioeconomic and health factors that our clustering model was not trained on. A specific feature that we sought after was the cancer incidence rate for each California county which would have been directly applicable to our project. Another feature we would have liked to include was the percentage of smokers in the population of each county, as smoking is known to be directly correlated with increased cancer risk. However, this data proved difficult to obtain access to during our data collection. There are also many different clustering methods with their own sets of strengths and tradeoffs potentially leading to different results.

VII. Team Effort Statement

All team members contributed equally to the project.

References

1. Adekugbe, A. P., & Ibeh, C. V. (2024). Tackling health disparities in the United States through data analytics: A nationwide perspective. *Health Data Journal*, 1(1), 45-67.
2. Bowser, D. M., Maurico, K., Ruscitti, B. A., & Crown, W. H. (2024). American clusters: Using machine learning to understand health and health care disparities in the United States. *Health Affairs Scholar, 2*(3), qxae017. https://doi.org/10.1093/haschl/qxae017
3. Chen, H., & McBride, M. (2021). A Geospatial Approach to Identifying Screening Gaps. *Journal of Cancer Prevention, 22*(2), 94-108. https://link.springer.com/chapter/10.1007/978-3-031-66413-7_3
4. Codex, A. C. (2024). Integrating D3.js with GIS for Geographic Data Visualizations. *Geospatial Data Science Journal, 15*(2), 152-168.
5. Cyr, M. E., Etchin, A. G., Guthrie, B. J., & Shadel, B. N. (2019). Access to specialty healthcare in urban versus rural US populations: A systematic literature review. *BMC Health Services Research, 19*, 974. https://doi.org/10.1186/s12913-019-4815-5
6. Dwyer-Lindgren, L., et al. (2017). Inequalities in Life Expectancy Among US Counties, 1980 to 2014: Temporal Trends and Key Drivers. *JAMA Internal Medicine, 177*(7), 1003-1011. https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2626194
7. Freeman, H. P. (2004). Poverty, Culture, and Social Injustice: Determinants of Cancer Disparities. *CA: A Cancer Journal for Clinicians, 54*(2), 72-77. https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/canjclin.54.2.72
8. Grote, T., & Keeling, G. (2022). Enabling Fairness in Healthcare Through Machine Learning. *Ethics in Information Technology, 24*, 39. https://doi.org/10.1007/s10676-022-09658-7
9. Jones, A. (2013). BRFSS Data Analysis on Cancer Screening Access. *Journal of Epidemiology, 29*(4), 234-247.
10. Moujahid, A. (2016). Interactive Data Visualization of Geospatial Data using D3.js, DC.js, Leaflet.js and Python. *Data Science Journal*, 3(8), 45-56.
11. Qi, M., Santos, H., Pinheiro, P., McGuinness, D. L., & Bennett, K. P. (2023). Demographic and socioeconomic determinants of access to care: A subgroup disparity analysis using new equity-focused measurements. *PLoS ONE, 18*(11), e0290692. https://doi.org/10.1371/journal.pone.0290692
12. Recchia, D. R., Cramer, H., Wardle, J., et al. (2022). Profiles and predictors of healthcare utilization: Using a cluster-analytic approach to identify typical users across conventional, allied, and complementary medicine, and self-care. *BMC Health Services Research, 22*, 29. https://doi.org/10.1186/s12913-021-07426-9
13. Ross, C., & Mirowsky, J. (2001). Neighborhood Disadvantage and Health. *Social Science & Medicine, 53*(2), 151-169. https://www.jstor.org/stable/3090214
14. Ruggles, D. (2018). Visualizing Health Data with D3.js: A Comprehensive Guide. *Journal of Visualization, 4*(2), 76-85. https://doi.org/10.1007/s12650-018-0500-0
15. Schootman, M., Gomez, S. L., Henry, K. A., Paskett, E. D., Ellison, G. L., Oh, A., Taplin, S. H., Tatalovich, Z., & Berrigan, D. A. (2017). Geospatial Approaches to Cancer Control and Population Sciences. *Cancer Epidemiology, Biomarkers & Prevention, 26*(4), 472-475. https://doi.org/10.1158/1055-9965.EPI-17-0104

16. Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer Statistics, 2020. *CA: A Cancer Journal for Clinicians, 70*(1), 7-30. https://acsjournals.onlinelibrary.wiley.com/doi/epdf/10.3322/caac.21590
17. Smedley, B. D., Stith, A. Y., & Nelson, A. R. (Eds.). (2003). *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. National Academies Press. https://nap.nationalacademies.org/catalog/12875/unequal-treatment-confronting-racial-and-ethnic-disparities-in-health-care
18. Tipton, K., Leas, B. F., Flores, E., Jepson, C., Aysola, J., Cohen, J., Harhay, M., Schmidt, H., Weissman, G., Treadwell, J., Mull, N. K., & Siddique, S. M. (2023). *Impact of Healthcare Algorithms on Racial and Ethnic Disparities in Health and Healthcare*. Comparative Effectiveness Review No. 268. Agency for Healthcare Research and Quality. https://doi.org/10.23970/AHRQEPCCER268
19. U.S. Department of Health and Human Services, Agency for Healthcare Research and Quality. (2024). *Impact of Healthcare Algorithms on Racial and Ethnic Disparities in Health and Healthcare*. Comparative Effectiveness Review No. 268. Prepared by ECRI-Penn Medicine Evidence-based Practice Center. AHRQ Publication No. 24-EHC004. https://doi.org/10.23970/AHRQEPCCER268

[20]

Clusters	**Counties**
0	Fresno County, Imperial County, Kern County, Riverside County, San Bernardino County, Tulare County
1	Amador County, El Dorado County, Napa County, Nevada County, Placer County, Sacramento County, San Benito County, Santa Cruz County, Solano County, Sonoma County, Yolo County
2	Humboldt County, Lassen County, Shasta County, Siskiyou County, Tehama County, Trinity County, Yuba County
3	Alameda County, Contra Costa County, Marin County, San Francisco County, San Mateo County, Santa Clara County
4	Butte County, Calaveras County, Glenn County, Kings County, Lake County, Mendocino County, Merced County, Monterey County, San Joaquin County, Stanislaus County, Sutter County, Tuolumne County
5	Orange County, San Diego County

6	Los Angeles County
7	San Luis Obispo County, Santa Barbara County, Ventura County