White Paper Analysis: Comparing the Effects of Various Demographic, Socioeconomic, and Health Disparity Metrics on Stomach Cancer Mortality Rates in 2019 Across U.S. County Subpopulations

By: Roger Qiu and Shailja Somani

Abstract

Stomach cancer is an ongoing public health issue in the United States, with approximately 11,000 Americans expected to die of the disease in 2024. Past research has shown that oncology patient outcomes are greatly influenced by demographic, social determinants of health (SDOH), and health equity factors, but past research has often only looked at a subset of factors rather than considering a wide cross-section. The goal of this work was to consider 353 such factors to: (a) predict counties with high expected stomach cancer mortality rates to allocate higher medical resources to and (b) determine which factors are most strongly correlated with mortality rates to invest further research and program funding into.

We evaluated various models and hyperparameters, with the optimal model being a Light Gradient Boosting Machine Regressor trained on all demographic and health equity factors, as well as 90 components derived from Principal Components Analysis (PCA) on the SDOH features. The final model achieved an adjusted R2 value of 0.9607. Given that the model's performance declined at higher mortality rates, we recommend public health officials set a cutoff based on resources available and then allocate additional resources to all populations with a predicted mortality above that cutoff rather than looking to predict exact mortality rate values. SHapley Additive exPlanations (SHAP) values were used to improve model explainability. Top predictive factors in the model include age, race, preventative care (such as annual checkups and mammograms), and comorbidities (such as binge drinking, depression, and current asthma). We recommend public health officials and hospital networks invest in research, programs, and case manager support in these areas.

Table of Contents

1. Introduction

Stomach cancer remains a significant public health concern in the United States, with thousands of new cases and deaths each year. According to the American Cancer Society, an estimated 27,000 Americans will be diagnosed with stomach cancer in 2024, and about 11,000 of these cases will be fatal. This accounts for 1.5% of new cancer diagnoses annually. Understanding the factors contributing to stomach cancer mortality is crucial for developing targeted interventions and policies that can help reduce these rates and also to address health disparities. By using comprehensive data from the National Center for Health Statistics (NCHS) and a few other sources, this study aims to analyze and predict the impact of various demographic, socioeconomic, and health disparity metrics on stomach cancer mortality rates across U.S. counties.

2. Business Background

The National Center for Health Statistics (NCHS) provides extensive data on stomach cancer cases at the county level, aggregated by demographic factors such as age, sex, and race. By combining this data with county-level socioeconomic and health metrics from the Agency of Healthcare Research and Quality and CDC, we can analyze the distribution of stomach cancer mortality across multiple different influences. Past research has considered certain socioeconomic factors, but largely looked at a few in isolation, failing to capture the larger picture and feature interactions. Our comprehensive approach allows for a better understanding of the socioeconomic and health equity factors that influence stomach cancer mortality rates. Such an understanding will be crucial for developing effective public health strategies and interventions that can address the underlying causes of these disparities.

3. Problem Statement

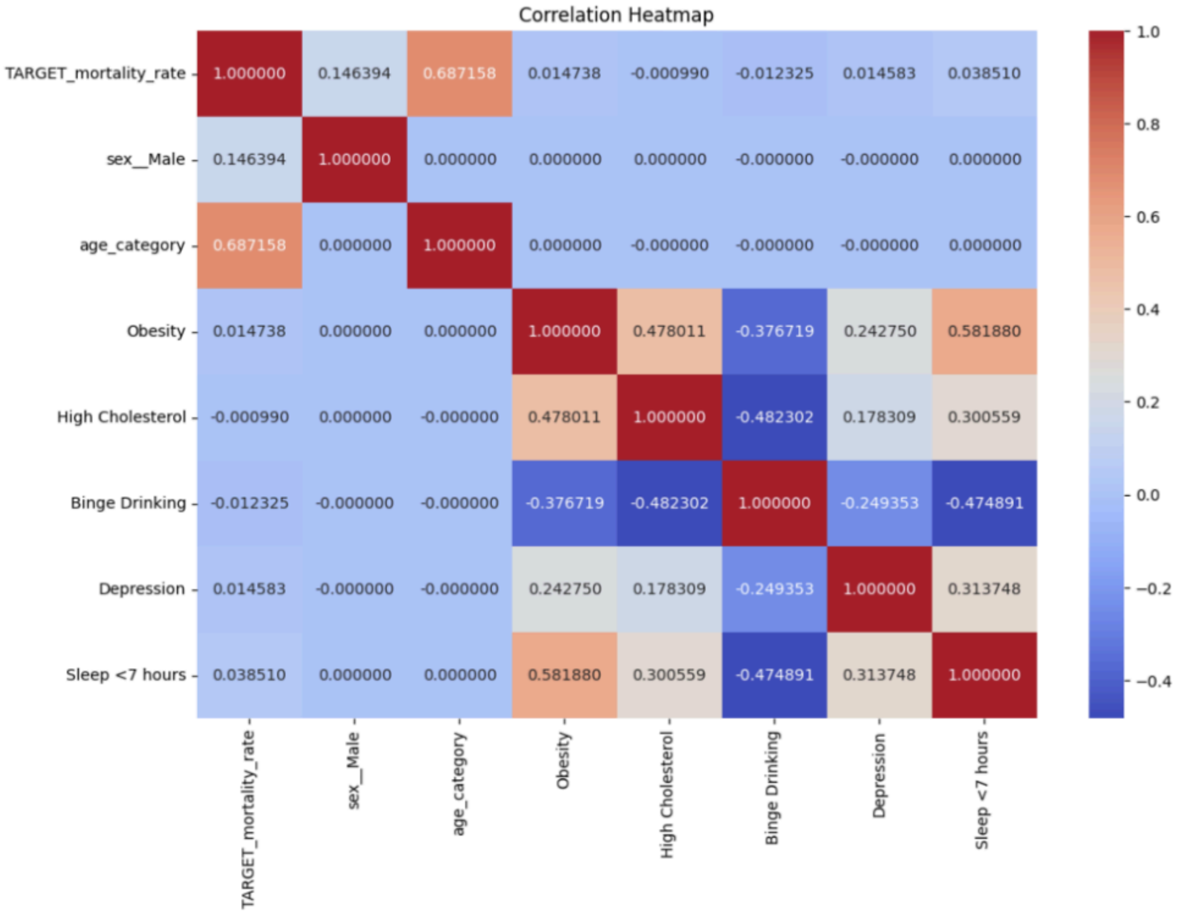
Stomach cancer mortality rates show large disparities across different U.S. counties, influenced by various demographic, socioeconomic, and other health disparity metrics. These disparities highlight the need for a comprehensive analysis to find and understand the key factors contributing to stomach cancer mortality. Existing literature often focuses on a limited number of factors or small subsets of the population, leaving a gap in understanding the larger influences on stomach cancer outcomes. By addressing this gap, this study aims to provide valuable insights that can help create the development of targeted interventions and policies designed to reduce stomach cancer mortality rates and improve health equity.

4. Summary of EDA Findings

The Exploratory Data Analysis (EDA) included a comprehensive review of the consolidated dataset through different statistical and graphical methods. This included univariate analysis which revealed a right-skewed distribution with a skewness of 2.92 and a kurtosis of 11.14, indicating a distribution with a heavy tail towards higher values. Visual tools like boxplots and histograms showed us the presence of outliers and the variability in mortality rates across counties, with a large number of counties exhibiting higher-than-average rates. Also, multivariate analysis was used to find correlations between mortality rates and other variables, with age showing a strong positive correlation of 0.69 (as demonstrated below in Figure 4.1), indicating its important role in predicting stomach cancer outcomes.

The EDA also included a distribution analysis of data available by race, sex, and age group, with the highest representation from White populations, followed by Black and Latino populations. Also, the county-level analysis showed specific regions, particularly in Alaska, with the highest stomach cancer mortality rates. Other health metrics such as obesity, high cholesterol, binge drinking, and depression were also analyzed at the county level, showing us areas with significant health challenges.

Figure 4.1
Correlation Heatmap of Target Variable (Stomach Cancer Mortality Rate) and Select County-Level Demographic and CDC Health Measures Features



5. Business Questions

1. Targeted Health Interventions: For the age groups and racial groups are most at risk, and how can targeted interventions be created to reduce mortality rates in these demographics?
2. Resource Allocation: What regions or county subpopulations should receive increased healthcare funding, staffing and resources based on the identified key features of stomach cancer mortality?
3. Long-Term Health Trends: How might future demographic shifts such as aging populations or changing racial compositions influence stomach cancer mortality rates, and what proactive measures can be taken?
4. Healthcare Access and Equity: Are there disparities in access to preventative care across different county subpopulations, and how can these be addressed?
5. Comorbidities: Which factors interact to exacerbate stomach cancer mortality risk? What public health programs can be enacted to specifically target these combinations of risk factors?

6. Scope of Analysis

This analysis includes a wide set of demographic, socioeconomic, and health disparity metrics at the county level within the United States. After the data sources used in this study are consolidated, they provide 343 total variables, including age, sex, race, income, education, access to healthcare, transportation access, mental health issues within a county, preventative care metrics, and various other health metrics. However, the analysis does not include individual-level health data due to data availability and privacy concerns.

7. Approach

Data from multiple sources were consolidated and preprocessed for analysis. The main datasets used include stomach cancer mortality rates from the NCHS, SDOH data from the American Community Survey (ACS), and health equity measures from the Centers for Disease Control and Prevention (CDC). The data is cleaned, transformed, and merged to create a comprehensive dataset for pre-processing.

In order to minimize multicollinearity issues and computing power required, we test two dimensionality reduction approaches to handle the 309 SDOH features: (1) using sex, age, race, and all 36 CDC health measures features along with 15 key SDOH features, and (2) using sex, age, race, and all 36 CDC health measures features along with the components resulting from conducting Principal Component Analysis (PCA) on the SDOH features.

For the first modeling approach, the pre-processing steps included feature selection, dropping rows with null values (only 3.82% of rows), doing a 75%-25% train-test split, and feature scaling using the StandardScaler. This resulted in a training dataset with 200,469 rows and 58 columns. In the second approach, null values were initially handled by dropping rows and columns with excessive null values. Following that, a 75%-25% train-test split was done and features were scaled using StandardScaler. After that, KNNImputer was used to impute the remaining nulls in the SDOH columns, then PCA was used on the SDOH features, reducing them to 90 principal components. This reduced the dimensionality of the SDOH data while retaining 95% of the variance from the original features. The final training dataset for this approach had 200,583 rows and 133 columns.

Following the pre-processing, modeling was conducted by evaluating hyperparameters for seven regression models for both our approaches. Full results are included in Table 7.1 below. The LightGBM model resulting from Approach 2 had the lowest RMSE. Adjusted R^2 values were also compared for the best model from Approach 1 versus Approach 2 to identify if it was worth adding in all the additional features in Approach 2. Approach 2 did have a higher adjusted R^2 value (0.9607), so the LightGBM model resulting from Approach 2 is our final model.

Table 7.1

Results from Grid Search through Hyperparameters using Five-Fold Cross-Validation. For each Modeling Approach's Training Dataset, the Optimal Hyperparameters and RMSE Values are Shown.

Model	**Approach 1 Optimal Hyperparam-eters**	**Approach 1 Training RMSE**	**Approach 1 Test RMSE**	**Approach 2 Optimal Hyperparam-eters**	**Approach 2 Training RMSE**	**Approach 2 Test RMSE**
Linear Regression	N/A	0.0001238	0.0001238	N/A	0.0001237	0.0001238
Ridge	alpha: 10.0	0.0001238	0.0001238	alpha: 10.0	0.0001237	0.0001238
Lasso	alpha: 0.1	0.0001778	0.0001782	alpha: 0.1	0.0001779	0.0001783
ElasticNet	alpha: 0.1, l1_ratio: 0.1	0.0001778	0.0001782	alpha: 0.1, l1_ratio: 0.1	0.0001779	0.0001783
XGB Regressor	learning_rate: 0.1, max_depth: 20, n_estimators: 20	0.0000480	0.0000484	learning_rate: 0.1, max_depth: 20, n_estimators: 20	0.0000478	0.0000491
LightGBM Regressor	learning_rate: 0.1, max_depth: 20, n_estimators: 20	0.0000311	0.0000359	learning_rate: 0.1, max_depth: 20, n_estimators: 20	0.0000300	0.0000355
Support Vector	C: 0.1, epsilon: 1	0.0011862	0.0011862	C: 0.1, epsilon: 1	0.0011862	0.0011863

8. Limitations

The analysis is limited by the availability and granularity of county-level data. Some counties may have missing or incomplete data for certain metrics, which could impact the accuracy and generalizability of the findings. Additionally, the study relies on aggregated data, which may not capture individual patient-level variations. Although this study provides valuable insights on how to allocate resources at a county-level, we recommend a subset of future work be done at the patient-level if data is available to.

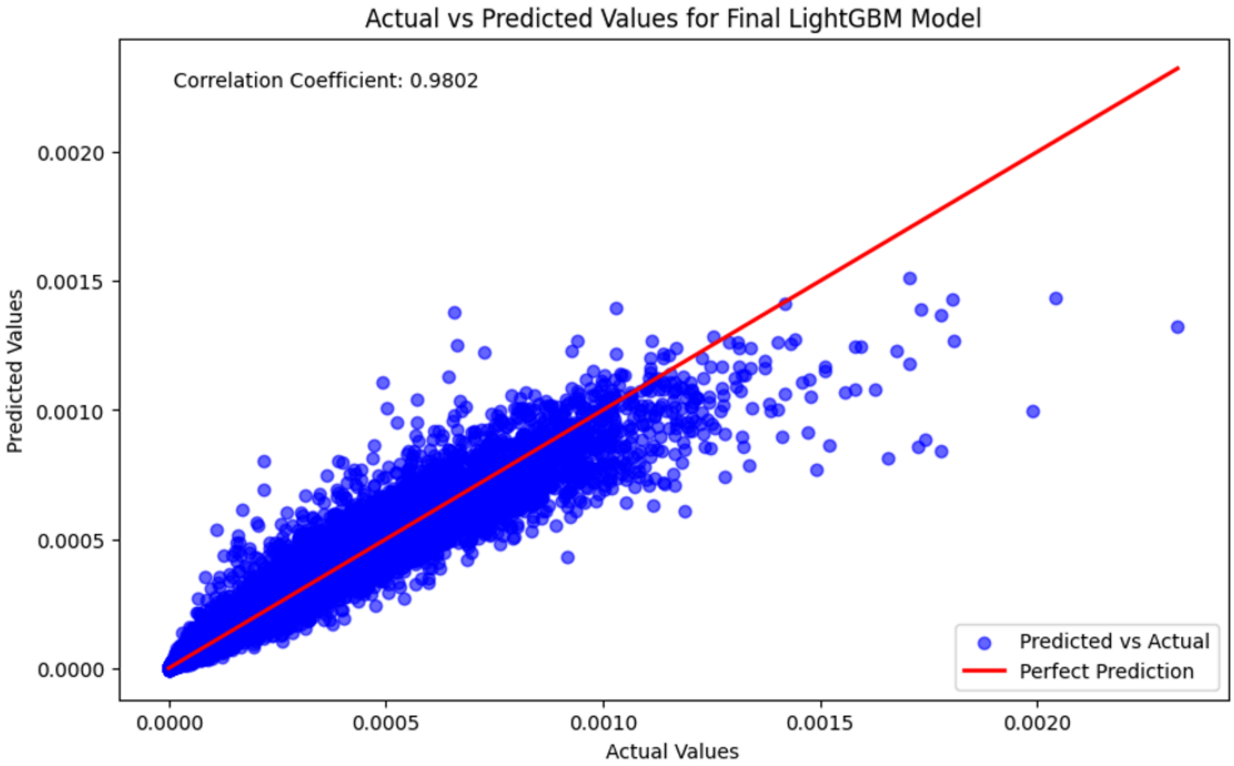
Another significant challenge encountered in this study was the lack of computational power as this research was done entirely locally due to it being an academic project with limited resources available to the authors. Given the large number of features included in the analysis, modeling and tuning hyperparameters can be computationally intensive and time-consuming even after using dimensionality reduction methods such as PCA. If a formal research organization or company was able to invest in furthering this research, we recommend further investigation of deep learning methods and additional hyperparameter tuning to ensure the resulting model is as robust as possible.

9. Solution Details

With an adjusted R^2 value of 0.9607, our final LightGBM model is an excellent tool for understanding stomach cancer mortality rates as 96.07% of the variation within the mortality rates is explainable by the features in our model. When the model was tested on the test dataset, its predicted values were also extremely close to the actual, observed values. Figure 9.1 below demonstrates that the model generally does quite well at predicting mortality rates, with a very high correlation coefficient of 0.9802 between the actual and predicted values. The only area in which the model struggles is at predicting particularly high mortality rates due to them being high outliers. For that reason, we do not recommend using the model to predict exact mortality rates, but rather to identify the top 10-15% of county subpopulations with the highest predicted stomach cancer mortality rates to invest further resources into.

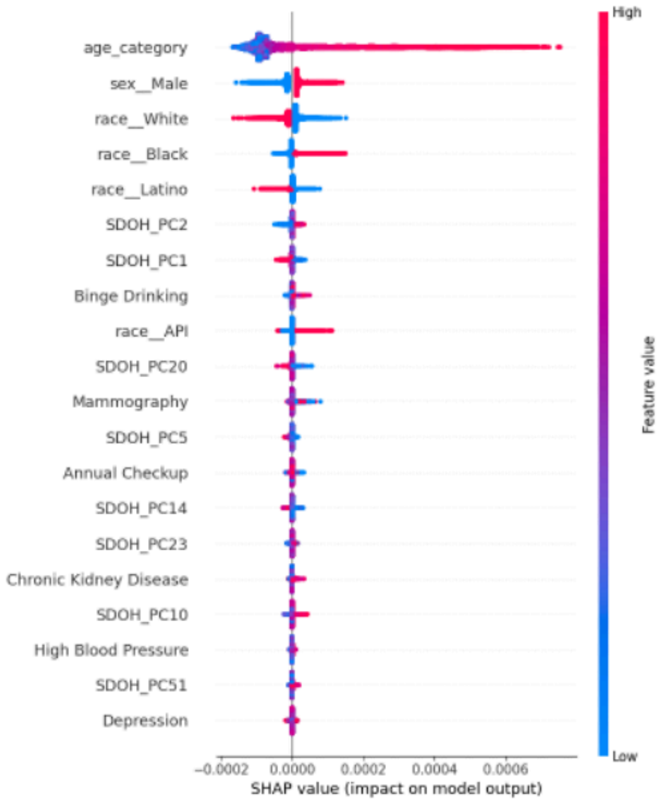
Figure 9.1

Comparing the Actual versus Predicted Values in the Test Dataset for the Final LightGBM Model



It is also crucial to mention that, even if our model only performs decently rather than exceptionally, its feature importance is very insightful, especially given the wide array of features we are considering rather than a small subset, as previous studies have done. Figure 9.2 below shows us the global SHAP values for the model, highlighting both magnitude and direction of feature importance. Age and race are key predictive features for stomach cancer mortality. The percentage of a county that receives preventative care such as annual checkups and mammography are also significant features, as well as the percentage of a county with comorbidities, such as binge drinking, depression, and chronic kidney disease. Examining global and local, prediction-level SHAP values generated from our model will allow public health officials to carefully tailor intervention programs in their specific regions.

Figure 9.2
Global SHAP Values for Final LightGBM Model



10. Concluding Summary

The final model developed in this research was a Light Gradient Boosting Machine (LightGBM) Regressor trained on 133 features, including demographic, health equity data, and 90 principal components from PCA on SDOH features. The model achieved an RMSE of 0.0000355. The model performed very well in terms of adjusted R², scoring 0.9607, which means that 96.07% of the variability in stomach cancer mortality rates can be explained by the model's features. Also, the correlation coefficient of 0.9802 between actual and predicted values confirms the model's effectiveness. Aside from the performance metrics, the model's feature importance analysis gave us the key predictors of stomach cancer mortality, such as age, race, preventative care rates, and comorbidities like binge drinking and depression. SHAP values can further be used to drill down into features driving up a given predicted mortality rate most, allowing local public health officials to tailor their interventions to issues facing their specific populations.

11. Call to Action

For action steps to reduce stomach cancer mortality rates, it is important to implement targeted resource allocation strategies based on the model's predictions. By identifying the top 10-15% of counties with the highest predicted mortality rates, public health officials can prioritize these areas for additional resources and staffing. This can involve segmenting the counties where predicted rates are greater than a certain threshold and focusing funding, medical personnel, and outreach programs to these high-risk areas. Continuous monitoring of the impact of these interventions will allow us to make adjustments to ensure their effectiveness.

Furthermore, enhancing preventative care programs in these high-risk counties is also recommended, as preventative care was identified as a key factor in reducing stomach cancer mortality. Proactive outreach efforts, such as reminding residents to attend annual checkups and screenings should also be implemented. These efforts can be supported by partnerships with local health organizations and community leaders to raise awareness about the importance of early detection. Finally, creating incentives for regular healthcare participation, such as discounted services or rewards, could also encourage residents to engage in preventative care, which should contribute to lower mortality rates. We look forward to collaborating with healthcare subject matter experts and public health officials to enact such programs.

References

Agency for Healthcare Research and Quality (AHRQ). (2023, Jun). Social Determinants of Health Database. Retrieved Jun 30, 2024, from https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html#download.

American Cancer Society. (2024, Jan 19). Key statistics for stomach cancer. Retrieved June 30, 2024, from https://www.cancer.org/cancer/types/stomach-cancer/about/key-statistics.html.

Centers for Disease Control and Prevention (CDC). (2023, Aug 25). PLACES: Local Data for Better Health, ZCTA Data 2023 release. Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health. Retrieved Jun 30, 2024, from https://data.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-ZCTA-Data-2023/qnzd-25i4/about_data.

Ilic, M., & Ilic, I. (2022). Epidemiology of stomach cancer. World Journal of Gastroenterology, 28(12), 1187–1203. Retrieved Jul 7, 2024, from https://doi.org/10.3748/wjg.v28.i12.1187.

Institute for Health Metrics and Evaluation (IHME) Global Health Data Exchange. (2024, Feb 28). United States Stomach Cancer Mortality Rates by County, Race, and Ethnicity 2000-2019. Retrieved Jun 30, 2024, from https://ghdx.healthdata.org/record/ihme-data/united-states-stomach-cancer-mortality-by-county-race-ethnicity-2000-2019.

Katella, K. (2024, February 7). Stomach cancer is still a risk for many people. Yale Medicine. Retrieved Jul 21, 2024, from https://www.yalemedicine.org/news/stomach-cancer-gastric-cancer.

Lundberg, S. (2018). An introduction to explainable AI with Shapley values. SHAP Documentation. Retrieved Aug 9, 2024, from https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html.

Rippner, N. (2018, Jul 30). County FIPS to zip code crosswalk. DataWorld. Retrieved Jun 30, 2024, from https://data.world/nrippner/fips-to-zip-code-crosswalk.

Sarkar, S., Dauer, M. J., & In, H. (2021). Socioeconomic disparities in gastric cancer and identification of a single SES variable for predicting risk. Journal of Gastrointestinal Cancer, 53(1), 170–178. Retrieved Jul 1, 2024, from https://doi.org/10.1007/s12029-020-00564-z.

Santellano, B., Agrawal, R., Duchesne, G., Sharara, M., Balas, E. A., Agrawal, G., Tsai, M.-H., Nayak-Kapoor, A., & Cortes, J. E. (2024a). EPR24-113: The role of Social Determinants of health in gastrointestinal cancer outcomes in the United States context: A systematic review. Journal of the National Comprehensive Cancer Network, 22(2.5). Retrieved Jul 1, 2024, from https://doi.org/10.6004/jnccn.2023.7217.

SEER. (2021). Cancer of the stomach - cancer stat facts. Retrieved Jul 21, 2024, from https://seer.cancer.gov/statfacts/html/stomach.html.

United States Census Bureau. (2024, Jun 18). American Community Survey Data. Retrieved Jul 22, 2024, from https://www.census.gov/programs-surveys/acs/data.html.

United States Census Bureau. (2021, Oct 8). 2010 ZIP Code Tabulation Area (ZCTA) Relationship File Record Layouts. Retrieved Jun 30, 2024, from https://www.census.gov/programs-surveys/geography/technical-documentation/records-layout/2010-zcta-record-layout.html.