COMMUNITY WELLNESS INITIATIVES FOR CHRONIC DISEASE PREVENTION

SOURABH KAR

Contents

1. INTRODUCTION

Chronic diseases such as diabetes, cardiovascular disease, and obesity have become leading public health concerns worldwide, accounting for roughly three-quarters of all deaths (WHO, 2024). Chronic diseases have also been the leading causes of death in the United States for some time (CDC, 1999). The prevalence of diabetes is expected to roughly double worldwide by 2050 (Klein, 2023) forecasting a stark rise in the chronic disease burden. In response, the healthcare industry and public health systems are increasingly turning to data-driven solutions. As healthcare systems become data-driven ecosystems, the integration of artificial intelligence (AI), predictive analytics, and real-time surveillance is transforming the monitoring, understanding, and management of chronic diseases. For example, most healthcare providers now use remote patient monitoring systems for conditions like diabetes or heart failure, where patients' blood sugar or blood pressure readings are transmitted to clinicians, and AI algorithms detect any worrisome changes. These innovations are consistent with RQ4 in that they create new preventive care delivery channels that can be measured and optimized to improve chronic disease outcomes. Recent studies have demonstrated strong correlations between telehealth-enabled preventive care metrics (such as virtual visit adherence rates and remote monitoring compliance) and improved chronic disease control in previously underserved populations.

This literature review explores how these technological advancements are reshaping chronic disease management while also examining the challenges that accompany them. It critically analyzes key advancements in healthcare analytics (e.g., AI-driven prediction models and big data surveillance), investigates persistent difficulties in data-driven chronic disease management (such as data interoperability issues, health disparities, and ethical concerns), and identifies opportunities for improving intervention strategies. By synthesizing understanding from recent research and

industry innovations, the review illustrates how modern technological breakthroughs are influencing chronic disease surveillance, healthcare policy, and patient-centered care in practice.

1.1 RESEARCH QUESTIONS

The research questions guide the review:

1. How do combinations of preventive care access, behavioral risk factors, and population characteristics influence chronic disease prevalence patterns across urban areas?
2. What is the relationship between population size and healthcare resource distribution in determining chronic disease outcomes?
3. How do behavioral risk factors cluster geographically, and what is their collective impact on chronic disease outcomes?
4. How can understanding the link between preventive care metrics and chronic disease outcomes inform the development of targeted healthcare interventions?

2. LITERATURE REVIEW

DATA ANALYTICS IN CHRONIC DISEASE SURVEILLANCE

In response to RQ1, the detection and surveillance of disease were enhanced greatly through AI and machine learning. Public health is accelerated by the use of AI as it accelerates diagnosis and intervention, as described by Singareddy et al. (2023). AI models examine large EHR data and detect subtle patterns to enhance risk estimation and targeted prevention (Jiang, Y., 2024). Pan et al. (2024) discuss new applications primarily addressing diabetes, hypertension, and lung disease, enhancing risk stratification (Jiang, Y., 2024).

These are complemented by infrastructural activities such as the modernization of the data on chronic disease by the CDC. Nonetheless, there are shortcomings mainly data standardization and

system interoperability (Qi et al., 2023). The use of standard FHIR standards is recommended by WHO (2024) to standardize the surveillance. We also expect to integrate disparate datasets and address such interoperability issues in our work.

BEHAVIORAL RISK FACTORS AND GEOGRAPHIC DISPARITIES

In response to RQ2, studies demonstrate that behavioral risk factors (for example, smoking and physical inactivity) disproportionately impact poorer populations and drive rates of chronic disease (Rahelić et al., 2024; CDC BRFSS, 2024). Preventive services screening and vaccinations, for instance, can lower these risks (Singareddy et al., 2023), and community-level interventions are advocated for by WHO (2024). These results indicate the necessity of geographically focused prevention programs.

SOCIOECONOMIC DISPARITIES IN OUTCOMES

For RQ3, Qi et al. (2023) emphasize the increased rates of chronic disease faced by lower-income populations because of lack of insurance, low literacy, and access barriers. Turner and Hohman (2024) advocate for models of equity such as the use of mobile clinics and care subsidies to enhance outreach. Outcomes such as screening and visitation rates are useful to determine the intervention points. Although the ACA expanded access, there are still disparities between rural and underserved populations (CDC, 2024). We incorporate SES indicators to enhance predictive models and resource allocation.

INDUSTRY TRENDS IN AI & CHRONIC DISEASE MANAGEMENT

The market for managing chronic diseases is anticipated to grow from $5.7B in 2024 to reach $18.4B in 2033 (Dimension Market Research, 2024). AI-powered tools—such as wearable devices, chatbots, and remote monitoring—improve patient care and the detection of diseases at

an early stage (Jafleh, 2024; Singareddy et al., 2023). These technologies facilitate RQ1 and RQ4 through the extension of preventive care and personalization.

IMPLEMENTATION CHALLENGES & LIMITATIONS

Even with advantages, resistance to change, infrastructure expenditure, and labor shortages are obstacles to AI implementation (Wang et al., 2024; Aldossari et al., 2024). Ethical issues and interoperability are still key (Qi et al., 2024; Panch et al., 2018).

SUCCESSFUL CASES & FUTURE DIRECTIONS

Lark Health, Seha Virtual Hospital, and Sword Health demonstrate success in real-world applications in the area of AI-based chronic care (Singareddy et al., 2023; Jafleh et al., 2024; Smith et al., 2023). Use within elderly and low-income populations is low because of cost and digital impediments (Jafleh et al., 2024). It is to be overcome through training programs, subsidies, and accessibility design (WHO, 2024).

3. METHODOLOGY

3.1 DATA SOURCE AND COLLECTION

This research employs a quantitative, secondary data analysis approach investigating how preventive care utilization, behavioral risk factors, population characteristics, and healthcare resource distribution contribute to chronic disease prevalence across urban U.S. communities. Our primary data source was the "500 Cities: Census Tract-level Data (GIS Friendly Format), 2018 release" dataset, which includes data from 2015 and 2016. This dataset was provided by the Centers for Disease Control and Prevention (CDC), Division of Population Health, Epidemiology and Surveillance Branch, with funding from the Robert Wood Johnson Foundation (RWJF) in conjunction with the CDC Foundation. We utilized the GIS-friendly format that can be joined with census tract spatial data to produce maps of 27 health measures at the census tract level. The dataset

includes four key measures from the 2015 BRFSS that remained consistent with the 2017 release: high blood pressure, high blood pressure medication usage, high cholesterol, and cholesterol screening.

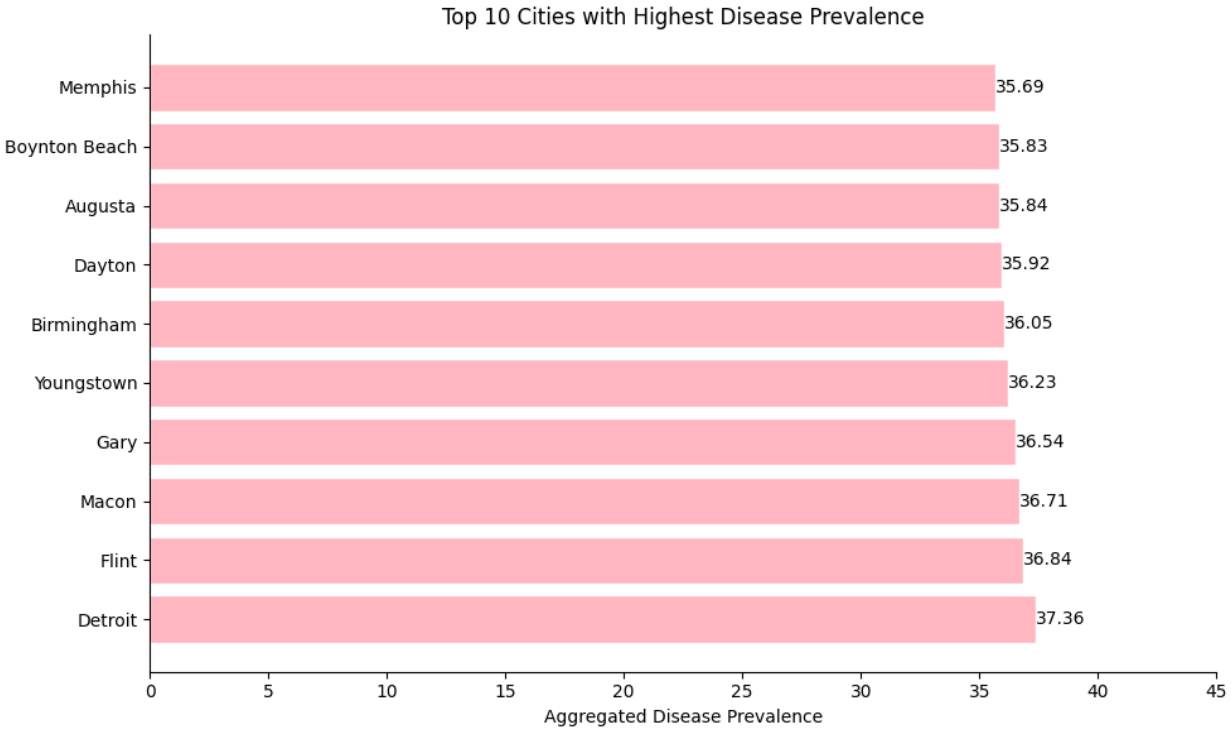

Figure 1: the top 10 U.S. cities with the highest aggregated chronic disease prevalence

Figure 1 displays the top 10 U.S. cities with the highest aggregated chronic disease prevalence based on tract-level data. These cities, highlight geographic hotspots of health risk.

3.2 UNIT OF ANALYSIS AND DATA PREPARATION

Census tracts were the unit of analysis, which allowed for enough detail to detect significant patterns in preventive service access and the burden of chronic diseases. Data preparation included the merging of several datasets on the basis of standard geographic identifiers, cleaning inconsistencies with Microsoft Excel and Python, filling in missing values with statistical

imputation, and transformation of the variables to satisfy statistical requirements for regression modeling.

Target Variable: Chronic Disease Index

The primary outcome variable was a composite Chronic Disease Index based on eight of the foremost diseases in the CDC data, weighted for public health burden.

CrudePrev is short for "Crude Prevalence," meaning the crude or unadjusted prevalence rate of a health condition within a population. It is the number of all cases divided by the overall population, presented as a percentage, without adjustment for demographic characteristics such as age, sex, or race. Variables ending with the suffix "_CrudePrev" in the CDC's 500 Cities dataset represent such crude prevalence estimates of health conditions at the census tract level.

Chronic Disease Index = (0.3 × DIABETES_CrudePrev) + (0.25 × BPHIGH_CrudePrev) +
(0.2 × OBESITY_CrudePrev) + (0.1 × CANCER_CrudePrev) +
(0.05 × COPD_CrudePrev) + (0.05 × KIDNEY_CrudePrev) +
(0.025 × MHLTH_CrudePrev) + (0.025 × PHLTH_CrudePrev)

The weighting strategy is based on disease prevalence, impact on deaths, and cost burden. The greatest weights were assigned to diabetes, hypertension, and obesity because of their causal and comorbid contributions to all other diseases. This aggregate index directly addresses all four research questions in the sense that it allows for examination of how access to preventive care, risk behavior, and population characteristics together determine disease burden, permits evaluation of nonlinear associations between population characteristics and health conditions, enhances analysis of behavioral risk clusters, and provides an objective measure of outcome by which to compare how preventive care utilization affects health disparities in various contexts.

This aligns with the contemporary demands for holistic health outcomes measurement (WHO, 2024; CDC, 2024; Pan et al., 2024; Singareddy et al., 2023) and offers a framework of methodological integrity to examine the intricate relationship between access to preventive services and the burden of chronic disease at the census tract level.

Prior to modeling, we performed standard data preprocessing. Missing values for health measure variables were imputed at the regional level using medians to preserve local variation but avoid bias. Population size attributes were log-transformed to reduce skewness. Categorical attributes were label-encoded as needed, and all continuous attributes were Min-Max normalized to impart comparability to the predictors. Further, extreme outliers were visually scanned and capped using domain knowledge to prevent model distortion during training.

3.3 TOOLS USED

This project employed a combination of tools and Python libraries to support the entire data science pipeline—from data preparation to predictive modeling and visualization. Microsoft Excel was utilized at initial stages to organize and format raw datasets, such as to remove extraneous columns and normalize column names. The primary analysis and modeling were done using Python, and the Jupyter Notebook environment was the main interface to develop the code, document it, and visualize the data.

Data cleaning and data manipulation were accomplished using the pandas library to conduct join on geographic IDs, filter data, group data, and manage missing values. Numerical computations were aided by the numpy library and statistical computations and data transformation by scipy, including log scaling and normalization. Simple imputation (mean and mode) and regression-

based imputation were carried out on missing data using the scikit-learn modules to promote data consistency.

Exploratory data analysis (EDA) was carried out using libraries including matplotlib and seaborn to produce visualizations including correlation heatmaps, scatterplots, histograms, and boxplots that yielded initial insights into trends and associations within the dataset.4.EDA BASED ON RESEARCH QUESTIONS

4.1 RESEARCH QUESTION 1

How do combinations of preventive care access, behavioral risk factors, and population characteristics influence chronic disease prevalence patterns across urban areas?

Figure 1 examined the relationship between physical inactivity with the Chronic Disease Index (CDI). The scatterplot indicated a strong positive linear relationship, which indicated that with higher levels of physical inactivity, chronic disease levels also increased at the census tract level. This is a strong confirmation of the reality that behavioral risk factors—specifically sedentary behaviors—are strongly correlated with urban chronic disease patterns. These findings confirm the work of Rahelić et al. (2024), which indicated that physical inactivity is a strong predictor of chronic disease, particularly among socioeconomically disadvantaged populations. The CDC's Behavioral Risk Factor Surveillance System (2024) also indicated higher levels of disease with higher levels of inactivity. Thus, this chart is a strong data-driven confirmation that physical inactivity is a high-impact predictor of chronic disease burden, supporting RQ1's focus on the role of behaviors.

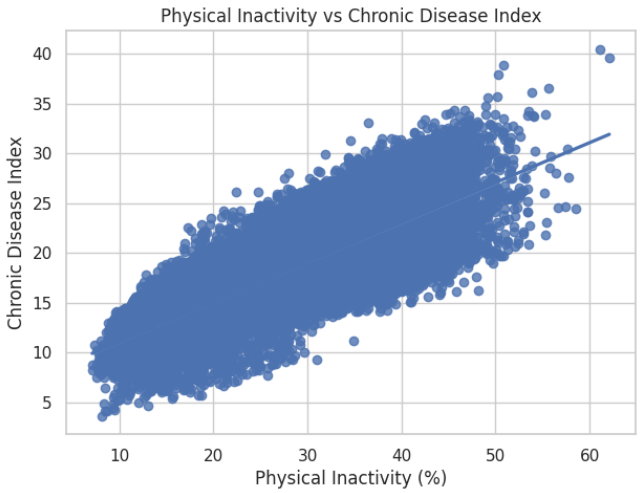

Figure 2: PHYSICAL INACTIVITY VS CHRONIC DISEASE INDEX

The second visualization compared the CDI across quartiles of preventive checkup rates. Contrary to expectations, tracts with the highest checkup rates (Q4) exhibited the highest median CDI. While preventive care is generally assumed to reduce disease burden, this counterintuitive finding suggests that areas with high checkup rates may be responding to already elevated health risks. It reveals a critical insight for RQ1: preventive care must be analyzed in conjunction with other contextual factors—such as socioeconomic status and health behaviors—to interpret its effectiveness. This observation is supported by Qi et al. (2023) and Singareddy et al. (2023), who noted that despite increased access to preventive care, outcomes remain poor in high-risk populations unless behavioral and environmental barriers are simultaneously addressed. Therefore, this chart emphasizes that preventive care alone is insufficient without considering the full ecosystem of contributing variables.

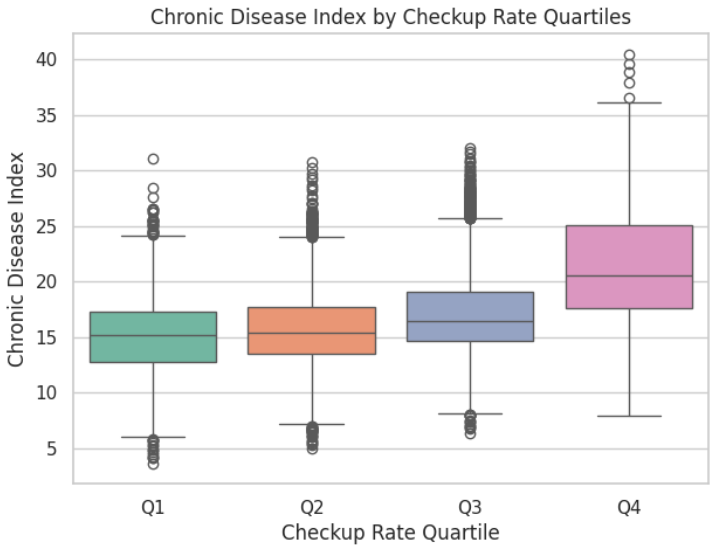

Figure 3: BOXPLOT – CDI BY CHECKUP RATE QUARTILES

Boxplot compared CDI for census tracts with low access to preventive care and high levels of smoking with all other tracts. The results were conclusive: the combined high-risk group had a much higher median CDI. This solidly supports the hypothesis for RQ1 that combinations of access and risk factors compound chronic disease outcomes. The finding also supports Gatzweiler et al. (2023), and Ashburner et al. (2017), whose work advocated for intervention models that account for the overlap of vulnerabilities inherent in urban settings. Isolating the high-risk subgroup, the analysis indicates the way two disadvantages limit access to checkups and high levels of tobacco use compound health disparities for urban residents. This suggests that future predictive modeling and public health interventions not only should account for individual risk factors but also analyze their interactions to better target resource distribution.

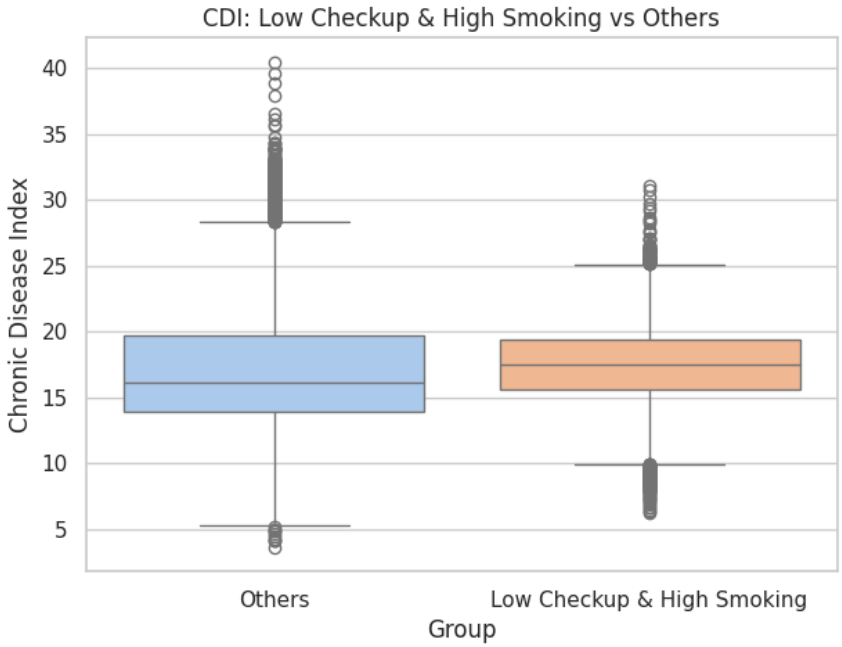

Figure 4: CDI – LOW CHECKUP & HIGH SMOKING VS OTHERS

CONCLUSION FOR RQ1 EDA

The exploration data analysis of RQ1 produced strong evidence that the prevalence of chronic disease in urban census tracts is heavily influenced by access to preventive care, as well as by behavioral risk factors. Inactivity was strongly and positively correlated with the Chronic Disease Index, further solidifying its position as the leading behavioral determinant of poor health. In a surprising result, higher levels of preventive checkups were correlated with higher chronic disease burden, suggesting that access to preventive care is responsive rather than preventive, especially for high-risk groups. Perhaps most saliently, census tracts with low levels of preventive care and high smoking levels had the highest levels of chronic disease, suggesting that combinations of access barriers and risk multiply health disparities. These findings affirm the core hypothesis of RQ1: chronic disease patterns derive from interacting factors, rather than from individual

variables. This finding highlights the significance of multidimensional modeling approaches that consider individual predictors as well as their joint effects for the project's next stage.

4.2 RESEARCH QUESTION 2

What is the relationship between population size and healthcare resource distribution in determining chronic disease outcomes?

To examine the relationship of population size with chronic disease outcomes (RQ2), two plots were considered. The first was a boxplot of the Chronic Disease Index (CDI) for three population ranges: tracts with less than 10,000 residents, 10,000–20,000, and 20,000–30,000 residents. The plot showed that tracts with fewer than 10,000 residents had consistently higher median CDI values, with greater variation and more extreme outliers. This is consistent with the hypothesis that small urban neighborhoods may carry a higher burden of chronic illness, perhaps due to limited access to healthcare services, transportation barriers, or the disproportionate distribution of preventive services.

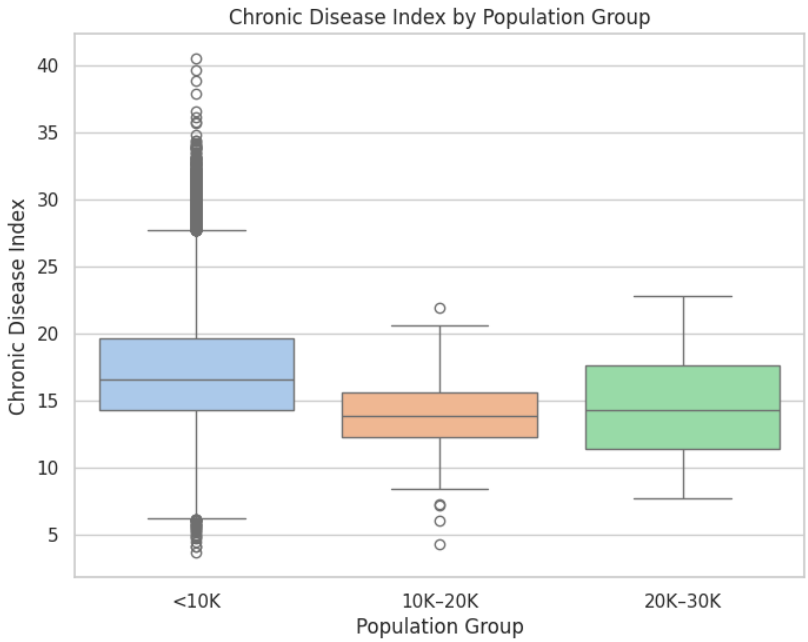

Figure 5 – CHRONIC DISEASE INDEX BY POPULATION BIN

The plot of Population of census data against the Chronic Disease Index, with a LOWESS smoothing curve for the purpose of identifying non-linear relationships. The curve had a definite falling slope, which suggests that with a growing population, the chronic disease rate is likely to be lower. Of particular importance is the steepness of the fall, most precipitous below the 10,000-population point, beyond which the curve slopes down less steeply. The curve supports the hypothesis of RQ2 that population thresholds affect the outcome of the disease. The LOWESS trend also indicates that census tracts with less than 10,000 residents represent a tipping point, beyond which accessibility and effectiveness of healthcare could be quite different from those of denser tracts.

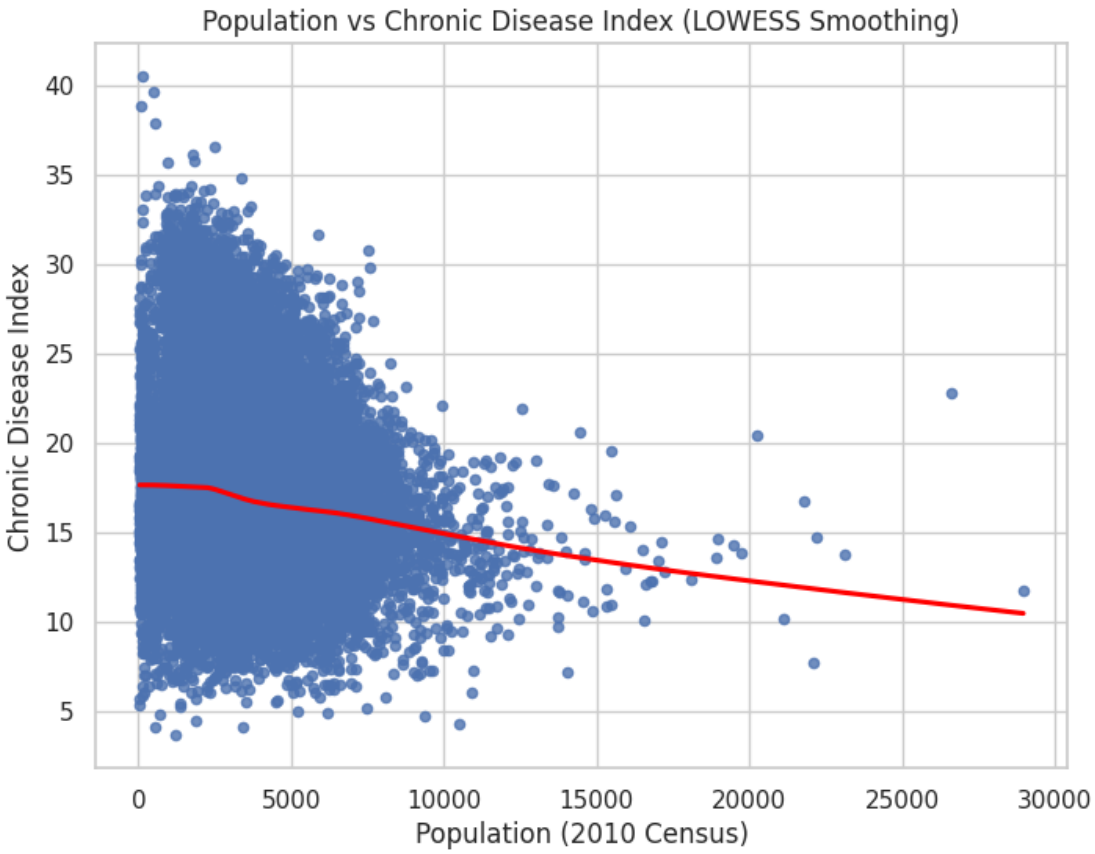

Figure 6 – POPULATION VS CDI (WITH LOWESS TRENDLINE)

CONCLUSION FOR RQ2 EDA

The boxplot analysis shows that census tracts with less than 10,000 residents had consistently higher median Chronic Disease Index (CDI) scores, with increased variation and outliers. These results suggest that low-population communities may bear a disproportionate chronic disease burden. These results agree with those of Benavidez et al. (2024), which revealed that low-density neighborhoods experience longer travel distances to healthcare providers, resulting in delayed diagnosis and limited access to preventive care. The CDC (2024) also shows that the availability of healthcare resources is typically low for low-population communities, further exacerbating

health disparities. Together, these results suggest the significance of population size as a contextual determinant of chronic disease, and they justify the application of population-stratified methods for predictive modeling and planning for achieving greater health equity.

4.3 RESEARCH QUESTION 3

How do behavioral risk factors cluster geographically, and what is their collective impact on chronic disease outcomes?

The health risk factor correlation heatmap shows the most serious health risks in physical inactivity, smoking, obesity, and diabetes clustering together across census tracts, with potential common socioeconomic or environmental factors. Consistent with the observations of Qi et al. (2023) and Rahelić et al. (2024), chronic disease risks tend to accumulate in communities with systemic prevention barriers. Binge drinking was the lone risk with negative correlations with the other risks, possibly due to a unique demographic or spatial structure, as noted by Gatzweiler et al. (2023), which correlated it with urban, younger populations. These comorbidity patterns affirm the crux of RQ3: that chronic disease burden is the product of the interaction of a number of behaviors and risks rather than singular factors. Based on this clustering, the next step toward segmenting tracts based on their common behavior profile and measuring their combined impact on chronic disease outcomes is the application of K-Means Clustering.

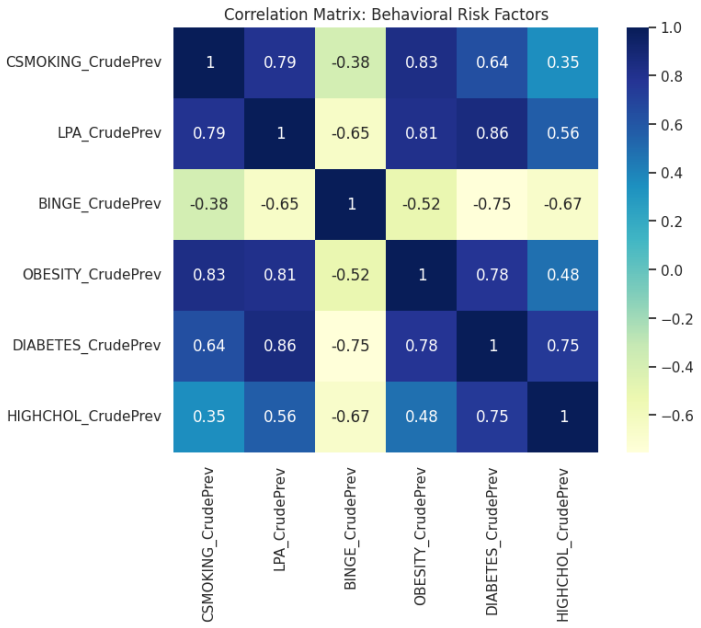

Figure 7 CORRELATION HEATMAP OF BEHAVIORAL RISK FACTORS

K-MEANS CLUSTERING

Clustering the behavioral risk factors showed that certain neighborhoods share similar patterns of unhealthy behaviors, and these areas tend to have higher chronic disease rates. This supports Research Question 3 by confirming that risks like smoking, inactivity, and obesity often occur together and have a stronger impact when combined. These findings align with Qi et al. (2023) and Rahelić et al. (2024), who noted that multiple health risks often cluster in disadvantaged communities. It also reflects Gatzweiler et al.'s (2023) observation that some behaviors, like binge drinking, may follow different patterns. Overall, the results show that looking at combined risks gives a clearer picture of where disease burdens are highest.

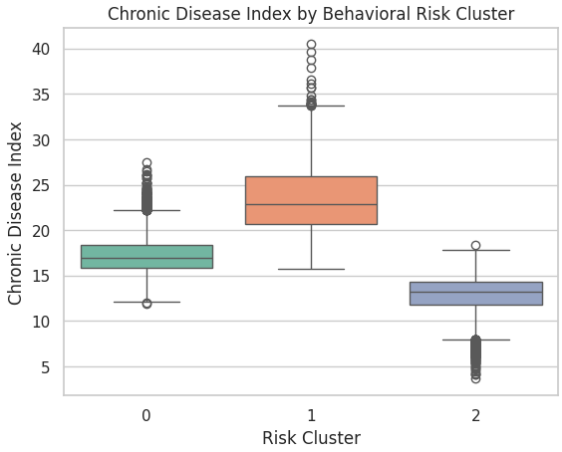

Figure 8: Chronic Disease Index by Behavioral Risk Cluster

CONCLUSION FOR RQ3 EDA

The exploratory analysis for RQ3 confirms that the smoking, exercise, obesity, and diabetes behavioral risk factors cluster together across the same census tracts, forming common health risk profiles. K-Means clustering revealed that neighborhoods with several simultaneous behavioral risks also carry a significantly higher chronic disease burden, affirming the hypothesis directly that such risks, collectively, but not individually, play a role in health outcomes. Consistent with the work of Qi et al. (2023) and Rahelić et al. (2024), chronic disease is often the cumulative effect of several unhealthy behaviors among socioeconomically disadvantaged populations. The distinctive binge drinking pattern supports Gatzweiler et al.'s (2023) suggestion that some behaviors have unique demographic or geographic patterns. Overall, the analysis affirms RQ3 by demonstrating that chronic disease understanding is reliant on the understanding of the interaction of the behavioral risks at the community level, further confirming the applicability of segmentation tools such as clustering for public health analysis and planning.

4.4 RESEARCH QUESTION 4:

How can understanding the relationship between preventive care metrics and chronic disease outcomes inform the development of targeted healthcare interventions?

The boxplot and scatterplots for RQ4 show that preventive care activities such as checkup rates, cholesterol testing, and pap testing don't necessarily have a straightforward inverse relationship with chronic disease burden. In fact, the highest checkup rate locations (Q4) have higher levels of Chronic Disease Index (CDI). This suggests that preventive services may be more available in high-disease neighborhoods, possibly as a response to prevalent existing diseases, rather than as a prevention success. Weak, even positive, relationships for all three preventive markers strengthen this interpretation. These findings reinforce RQ4 directly by showing that preventive care measurements can't be utilized independently to define disease outcomes and must be interpreted contextually. This is congruent with research by Qi et al. (2023) and Singareddy et al. (2023), which determined that preventive access tends to increase in communities already plagued with poor health, especially with urban underserved populations. Therefore, understanding where preventive care and disease burden overlap can be utilized to define key intervention zones where care access is present but not yet effective, guiding more focused public health interventions.

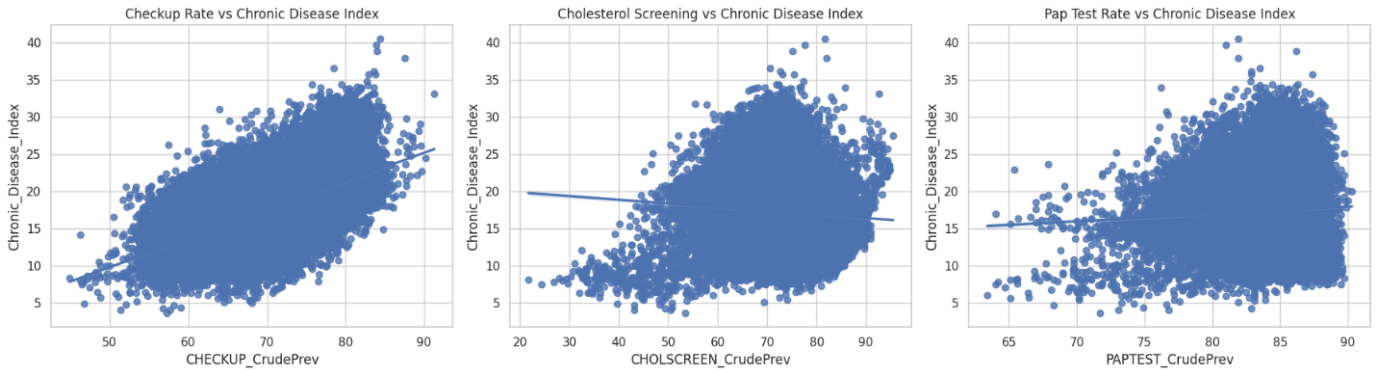

Figure 9: **Scatterplots (Preventive Metrics Vs CDI)**

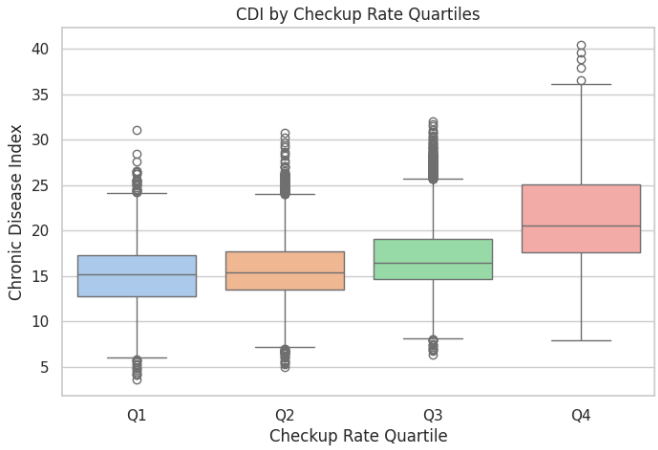

Figure 10 Quartile Boxplots (Checkup Rate Vs CDI)

The checkup rate vs. Chronic Disease Index quadrant analysis reveals four census tract groups with disparate levels of access to care and health outcomes. Of particular importance is the "Low Care / High Disease" quadrant, which reflects the highest chronic disease burden with low preventive care utilization—ideal for targeted intervention. Of most significance is the fact that the largest number of tracts fall into the "High Care / High Disease" quadrant, suggesting that, while care exists, it may be reactive, not preventive of disease. This addresses Research Question 4 by showing that the balance of care access and outcomes can be used to inform the location of the most needed intervention. These findings complement those of Singareddy et al. (2023) and Qi et al. (2023), who emphasized that healthcare access is insufficient—without behavior change and equitable distribution, chronic disease outcomes will be elevated. The quadrant strategy is useful for determining where care exists but is not health-promoting, and where care is lacking and disease is on the rise, offering a valuable public health planning tool.

Figure 11 Scatterplot with Color-Coded Quadrants

CONCLUSION FOR RQ4 EDA

The exploratory analysis of RQ4 reveals that although access to preventive care is beneficial, it does not necessarily result in lower chronic disease outcomes. Weak, even positive, relationships were observed between checkup rates and chronic disease through the analysis of scatterplots and quartiles, indicating that preventive services may be available in neighborhoods that already have poor health. The quadrant analysis supported this by showing census tracts with high care and high disease, and low care and high disease ideal targets for intervention. These findings reinforce the underlying premise of RQ4: that the interaction of care access and disease is crucial for the development of targeted healthcare strategies. In line with the work of Qi et al. (2023) and Singareddy et al. (2023), these results suggest that access must be supplemented with effective delivery and behavior change if the burden of the disease is to be reduced.

5. RESEARCH QUESTIONS AND HYPOTHESIS

5.1 RESEARCH QUESTION 1

How do combinations of preventive care access, behavioral risk factors, and population characteristics together influence chronic disease prevalence patterns across urban areas?

5.1.1HYPOTHESIS

Census tracts with lower preventive care utilization and higher risk behaviors will show significantly higher rates of chronic diseases .

5.1.2 LITERATURE INSIGHT:

This is corroborated by existing research. Rahelić et al. (2024) and the CDC (2024) reinforce the fact that smoking and lack of exercise are leading causes of chronic sickness, especially among socioeconomically disadvantaged groups. Furthermore, according to Qi et al. (2023), preventive care is not sufficient to achieve better results if the issue of risk behaviors and environmental obstacles is not solved concurrently.

5.1.3 WHAT OUR DATA SHOWED:

We used 3 visualizations in our exploration data analysis to test this hypothesis.

First, we created a scatterplot of chronic disease index vs. physical inactivity (LPA_CrudePrev), and we saw a strong linear relationship, which confirmed that with higher levels of inactivity in a census tract, the burden of chronic disease was also higher. That is strong evidence that physical inactivity is a strong predictor of chronic disease, which is further supported by the work of Rahelić et al. (2024).

Then, we compared CDI with checkup rate quartiles using a boxplot. In a surprising finding, the top quartile of the checkup rate had the highest median CDI, suggesting that preventive care is of a reactive rather than preventive kind. In agreement with the results of Qi et al. (2023) and Singareddy et al. (2023), we observed that preventive care services are concentrated in neighborhoods with already elevated health risks, and structural factors limit their long-term effect. Finally, a boxplot of tracts with low checkup rates and high smoking rates relative to all other tracts exhibited the highest CDI for the combined high-risk group, further indicating that interacting factors compound chronic disease outcomes. This finding is congruent with the suggestion of Gatzweiler et al. (2023) that overlapping vulnerabilities should be incorporated into public health strategy formulation.

5.1.4 INTERPRETATION & RECOMMENDATION:

Together, these EDA results confirm our hypothesis that chronic disease prevalence is shaped not just by individual factors but by their combined effects. The fact that preventive care access did not always correspond with lower disease levels indicates that care delivery must be contextualized within behavior and population characteristics. We recommend integrated, community-level interventions that target both access improvements and behavioral risk reduction, as supported by the literature. These findings also lay a solid foundation for predictive modeling to identify which combinations of variables most strongly predict chronic disease burden across urban neighborhoods.

5.2 RESEARCH QUESTION 2

What is the relationship between population size and healthcare resource distribution in determining chronic disease outcomes?

5.2.1 HYPOTHESIS

The effectiveness of healthcare resource allocation varies non-linearly with population size with areas above and below 10,000 residents showing distinct patterns in the relationship between healthcare access and disease prevalence.

5.2.2 WHAT OUR DATA SHOWED

To used two visualizations in EDA for RQ2 to come this result.

First, a boxplot of the Chronic Disease Index (CDI) across three ranges of population (<10K, 10K–20K, 20K–30K) revealed that tracts with fewer than 10,000 residents had the highest median CDI with the most spread. That suggests that small communities carry a heavier and more varied burden of illness, perhaps due to resource limitations or barriers to access.

Second, a population size vs. CDI scatterplot with a LOWESS smooth curve had a definite non-linear decreasing trend. The steepest decline of the disease burden was below the population size of 10,000, with the curve then flattening. This is consistent with the hypothesis that 10,000 residents are a population tipping point below which healthcare access and delivery become less effective or increasingly fragmented.

5.2.3 INTERPRETATION & RECOMMENDATION:

The results firmly confirm Hypothesis 2, with nonlinear population size association with health outcomes, with the most vulnerable being those of the smallest tracts. The results confirm the

literature and offer new insights into the importance of population-sensitive health resource planning. Predictive modeling and policy should be population thresholds-sensitive when planning for the coverage of interventions and services, especially for resource-deprived neighborhoods with fewer than 10,000 residents.

5.3 RESEARCH QUESTION 3

How do behavioral risk factors cluster geographically, and what is their collective impact on chronic disease outcomes?

5.3.1 HYPOTHESIS

Census tracts will show distinct patterns of behavioral risk factor clustering with areas of multiple high-risk behaviors demonstrating excessively higher chronic disease prevalence.

5.3.2 LITERATURE INSIGHT

This is confirmed by Rahelić et al. (2024) and Qi et al. (2023), who noted that unhealthy behaviors such as smoking, lack of exercise, and obesity cluster together in socioeconomically deprived neighborhoods, leading to much-elevated disease burden. Gatzweiler et al. (2023) also observed that some of the risk behaviors, for example, binge drinking, may have other demographic or geographic patterns, and hence clustering of behavior may be population context-specific.

5.3.3 WHAT OUR DATA SHOWED

To identify which of the behavioral risk factors are associated with chronic disease outcomes, the individual correlation of their presence with the Chronic Disease Index (CDI) was first computed. Figure 12 indicates that diabetes, poor physical health, obesity, inactivity, elevated cholesterol, and

smoking showed the strongest correlation. These six characteristics were used for K-Means clustering in order to detect important community-based risk patterns.

We also generated a correlation heatmap, which further revealed strong interrelationships among these variables—particularly between smoking, physical inactivity, obesity, diabetes, and high cholesterol. These patterns confirm that behavioral risks tend to co-occur within the same census tracts. Interestingly, binge drinking exhibited weak or negative correlations with the others, suggesting a different spatial or demographic pattern, as also noted by Gatzweiler et al. (2023).

Using the selected variables, we performed K-Means clustering to segment census tracts based on behavioral profiles. The model identified three distinct clusters with varying levels of cumulative risk. A subsequent boxplot of CDI by cluster showed that the highest-risk group exhibited the greatest chronic disease burden, whereas the lowest-risk group had the smallest. This strongly supports our hypothesis that a combination of co-occurring behavioral risks compounds chronic disease outcomes.

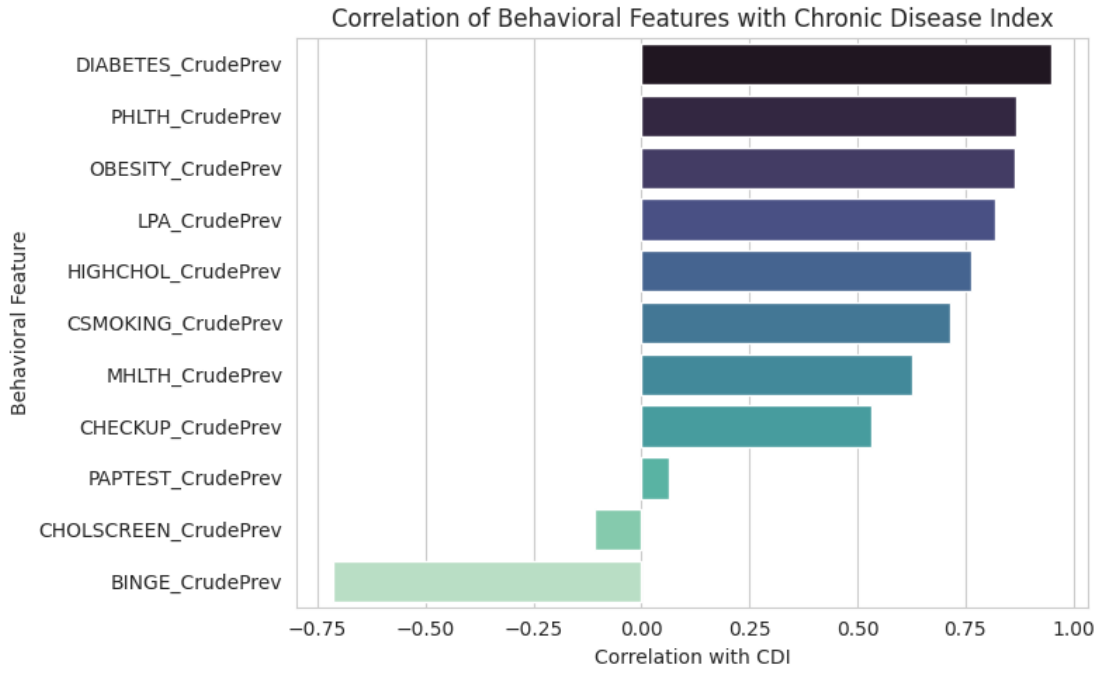

Figure 12: Correlation of Behavioral Features with Chronic Disease Index

5.4 INTERPRETATION & RECOMMENDATION

Our findings fully support Hypothesis 3 and align with the literature review. Behavioral risks clearly cluster within certain communities and have a compounded effect on chronic disease outcomes. These insights validate the use of clustering methods in chronic disease analysis and support public health strategies that target neighborhoods not just with individual risks but with multiple overlapping vulnerabilities. Including cluster labels in predictive modeling will further enhance the targeting and effectiveness of disease prevention efforts.

5.4 RESEARCH QUESTION 4

How can understanding the relationship between preventive care metrics and chronic disease outcomes inform the development of targeted healthcare interventions?

5.4.1 HYPOTHESIS

Areas with similar patterns of preventive care use, but different disease outcomes will reveal opportunities for targeted intervention strategies and resource optimization.

5.4.2 LITERATURE INSIGHT:

Singareddy et al. (2023) and Qi et al. (2023) elucidate that even though preventive care is beneficial, it will not necessarily lead to better health if other matters—like poverty, low education levels, or poor housing are not addressed. Their research shows that even if people from underserved populations seek healthcare services, their health will not be better if these other matters are not addressed.

5.4.3 WHAT OUR DATA SHOWED

Scatterplots of checkup rates, cholesterol screening, and pap tests against CDI had weak or positive correlations, indicating that increased care is not necessarily a predictor of less illness. The boxplot

of CDI by checkup rate quartiles also supported this, with the highest checkup group unexpectedly having the highest median CDI.

Quadrant plot with four quadrants based on checkup rate and CDI. The plot easily identified tracts with high care and high disease, and those with low care and high disease as the intervention priority zones. The fact that high-care zones consistently had high diseases supports the hypothesis that efficiency gaps exist in the translation of care into outcomes and that the same levels of care can produce differential health outcomes.

5.4.4 INTERPRETATION AND RECOMMENDATION:

These results support Hypothesis 4 and are consistent with previous research. The significance of context-specific interventions is further supported by the differences in outcomes by zones with the same use of preventive care. Our quadrant analysis is a useful technique for public health planning that may be used to rank and identify areas with high disease and no care, as well as areas with available care but limited access. With this knowledge, targeted intervention measures that save resources can be planned.

6. MODELING AND PREDICTION

6.1 DATA PREPARATION

The modeling phase employed the processed and cleansed dataset developed during the methodology stage. The Chronic Disease Index (CDI) was the response variable, indicating total chronic disease burden at the census tract level. Predictor variables were selected based on their relevance to the research questions and results of the exploratory data analysis. These included behavioral risk factors (like smoking, inactivity, obesity, and diabetes), preventive care practices (like checkups and screening), and population factors. A K-Means cluster label was also introduced as a categorical feature that indicates combined behavioral risk patterns identified in RQ3. The

dataset was split into training and testing subsets with an 80/20 split, with both models (Random Forest and Support Vector Regression) being trained and tested on the same partitions for the purpose of direct comparison of performance.

6.2 MODEL SELECTION STRATEGY

In observing the impact of population attributes, preventive care, and behavioral risks on rates of disease labeled as chronic, we used two machine learning techniques: Support Vector Regression and Random Forest. These were chosen since they are recognized to offer comprehensible results with good predictive performance.

We employed Random Forest, which best suits the detection of intricate associations and interactions between the variables. The method is in conformity with that of Rahelić et al. (2024), who established that the utilization of tree-based models is suitable for the detection of intricate health factors. Random Forest provides insights of significance through feature importance analysis, which makes it possible for us to establish what factors most contribute to the outcomes of the chronic disease.

We also employed K-Means Clustering to aid Research Question 3. We employed it to cluster census tracts into similar patterns of behavior, including smoking and physical activity. These groupings were then employed within the models to predict the impact of group-level risk patterns on chronic disease. We extended the research of Gatzweiler et al. (2023), which demonstrate the utility of clusters for the examination of community health risks.

In addition, Support Vector Regression (SVR) was also employed to further refine prediction performance. Due to the potential of SVR to model non-linear associations through the application of kernel functions, the ability of an SVR was examined in a trial to determine whether including a more flexible margin-based estimator would lead to better predictive performance. The

application of SVR is ensured to consider intricate non-linear structures, as suggested in recent models of chronic disease (Qi et al., 2023; Pan et al., 2024).

The models were all trained on 80% of the data and then tested on 20%. We compared the performance using the same metrics — R^2, RMSE, MAE, and MAPE — to equitably compare and determine which model best predicted the burden of chronic disease.

6.3 MODEL TRAINING & EVALUATION

As we prepared the data and selected the models, we trained and tested each of the three with an 80/20 train-test split. Random Forest, and Support Vector Regression were tested on the following metrics:

1. R^2 Score (the measure of the model's explanatory power for chronic disease variance).
2. Root Mean Squared Error (RMSE),
3. Mean Absolute Error (MAE).
4. Mean Absolute Percentage Error (MAPE)

These measurements helped us to better understand the accuracy of the predictions, as well as the closeness of the model results with actual Chronic Disease Index (CDI) values.

6.3.1 RANDOM FOREST REGRESSION

Random Forest worked well to predict the outcomes of chronic disease. It had an R^2 of 0.9901, which implied that it accounted for about 99% of the variance in the Chronic Disease Index. It also had an RMSE of 0.4612 and an MAE of 0.3429, along with having just a low MAPE of 2.05%, which shows precise prediction for all census tracts. This shows that the Random Forest performed extremely well to model the non-linear relationship and interaction amongst the variables in the data.

The scatter plot displays the predicted values plotted against actual Chronic Disease Index values from the Random Forest model. The data points are tightly grouped along the diagonal indicating good consistency and predictive precision in the model.

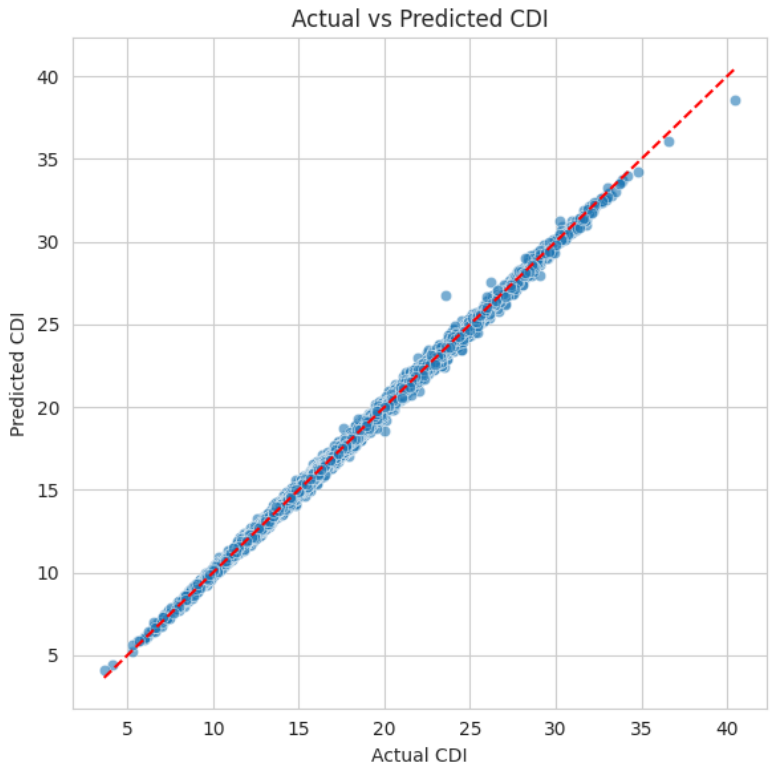

Figure 13: Actual Chronic Disease Index values with the predicted values from the Random Forest model

6.3.2 SUPPORT VECTOR REGRESSION

Support Vector Regression also performed well, with an R^2 of 0.9941, RMSE of 0.3563, and MAE of 0.2435, performing better than all of the other models in this study. The MAPE was 1.52%, the lowest among all models, indicating highly accurate predictions relative to actual values. SVR is ideally capable of modeling complex, non-linear patterns in the data, so its strong performance here is evidence that nuance in the variables was well captured.

6.4 MODEL COMPARISON

Below is a chart comparing each model based on four major metrics: R^2, RMSE, MAE, and MAPE. Of these two models, Support Vector Regression (SVR) performed best. It produced the highest R^2 (0.9941), i.e., it accounted for maximum variance in levels of chronic disease, and lowest error values in RMSE (0.3563), MAE (0.2435), and MAPE (1.52%). This means that SVR was the most accurate and consistent model overall.

Based upon this comparison, SVR is deemed to be the optimal model for chronic disease outcome prediction. The SVR prediction scatterplot is consistent with this result: points fall precisely along the diagonal, reflecting minimal error with high prediction accuracy. Indeed, Support Vector Regression (SVR) produced the highest R^2 (0.9941), with the lowest values for all metrics, which only serves to point to its ability to identify complex, non-linear relationships in the data. Thus, SVR is the most accurate model in this study.

Random Forest, though not achieving the absolute highest performance, still demonstrated excellent predictive capability with an $R^2 = 0.9901$, RMSE = 0.4612, MAE = 0.3429, and MAPE = 2.05%. The Random Forest model provided the additional advantage of feature importance analysis, offering valuable insights into which factors most strongly influence chronic disease outcomes.

In short, while SVR produced the most effective predictions, Random Forest was helpful too since it brought transparency, stability, and applicability to the real-world for planning in health-related contexts in which understanding contributing factors could be equally important as raw prediction accuracy.

MODEL	R^2 SCORE	RMSE	MAE	MAPE
Random Forest	0.9901	0.4612	0.3429	2.05
Support Vector Regression	0.9941	0.3563	0.2435	1.52

Table 6.4.1

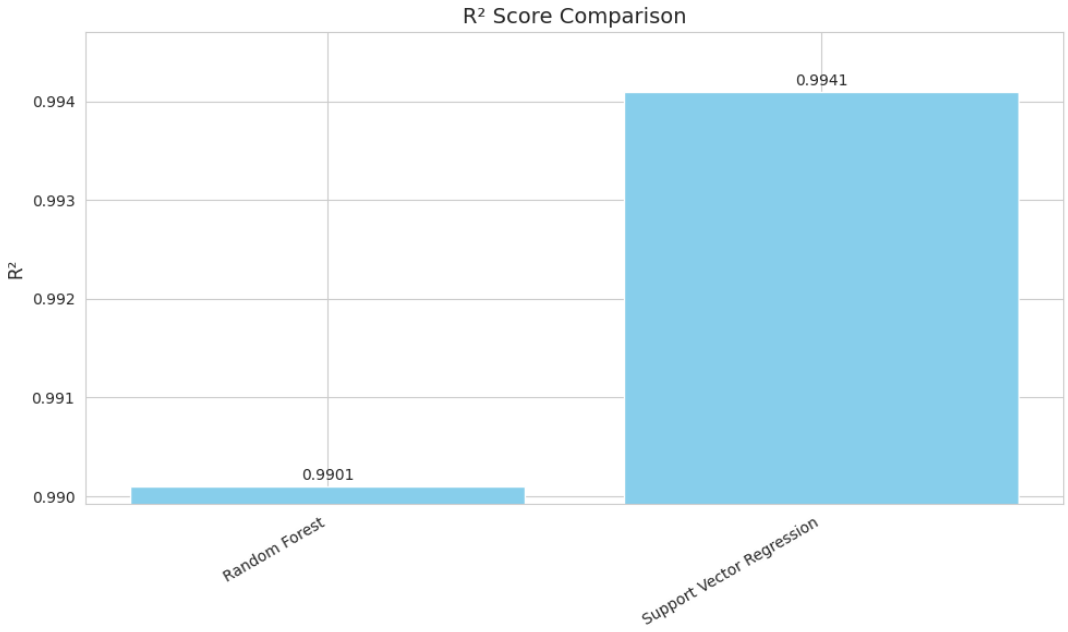

Figure 14: R^2 Comparison

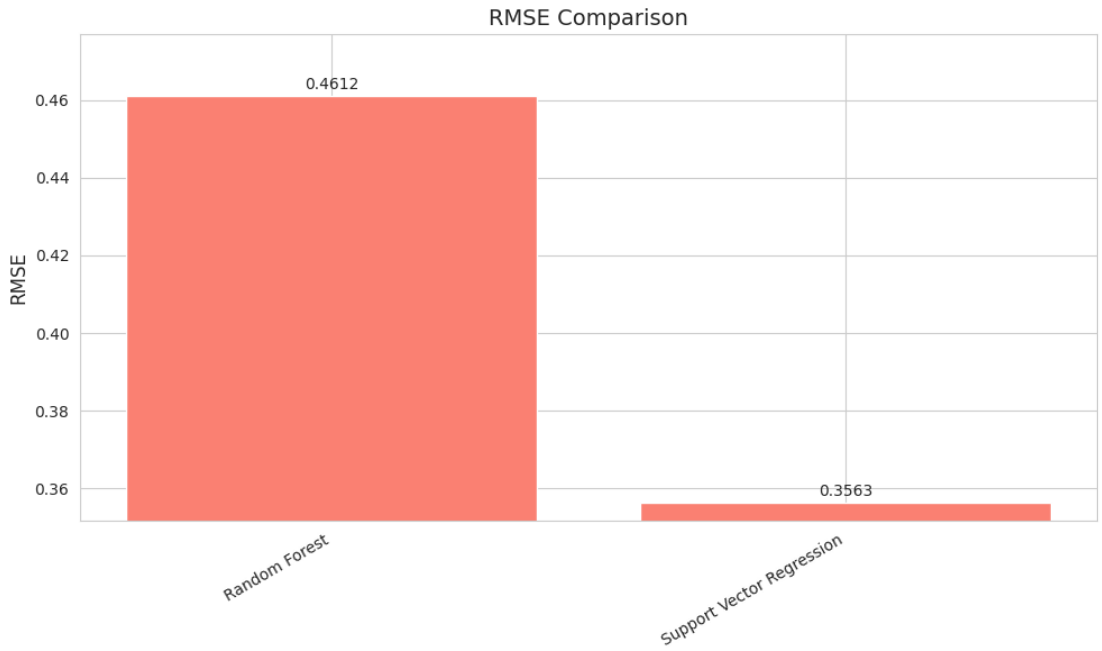

Figure 15: RMSE Comparison

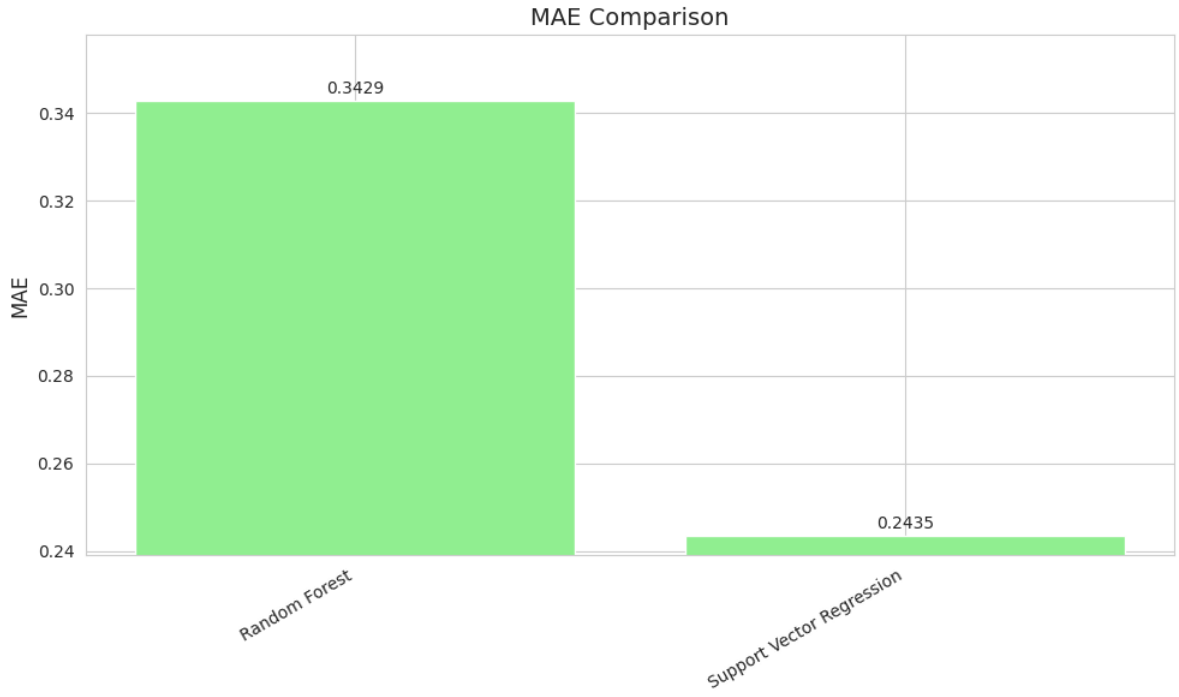

Figure 16: MAE Comparison

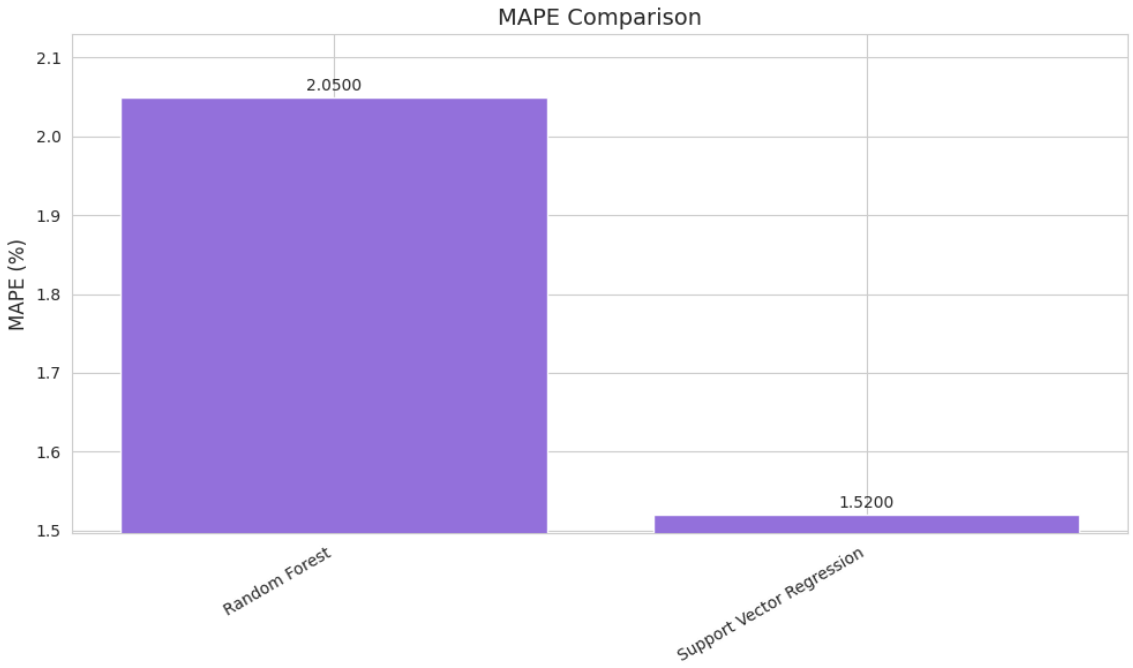

Figure 17: MAPE Comparison

7. K-MEANS CLUSTERING FOR BEHAVIORAL RISK PROFILES

To see which of the behavior risk factors cluster with the urban census tracts, we used K-Means Clustering with six behavior features:

Smoking (CSMOKING), Physical Inactivity (LPA), Obesity (OBESITY), Diabetes (DIABETES), High Cholesterol (HIGHCHOL), and Binge Drinking (BINGE).

These variables were chosen due to their strong correlations.

We used the Elbow Method for the calculation of the number of clusters. The elbow point we looked at was for k = 3, which indicated three distinct behavior patterns with census tracts. Each cluster was a neighborhoods with their respective patterns of risks—some with high smoking and inactivity, some with combined or lower risks.

To find the connection of these clusters with chronic disease, we produced a boxplot of the Chronic Disease Index (CDI) for the three groups. The plot indicated that the CDI values of the clusters with elevated behavior risks were much higher, supporting the hypothesis for RQ3.

We also included the cluster labels as categorical predictors into the supervised models to estimate whether being a member of a high-risk group increased the predicted burden of disease. We did this to take the combined effect of patterns of behavior into consideration, rather than analyzing individual variables.

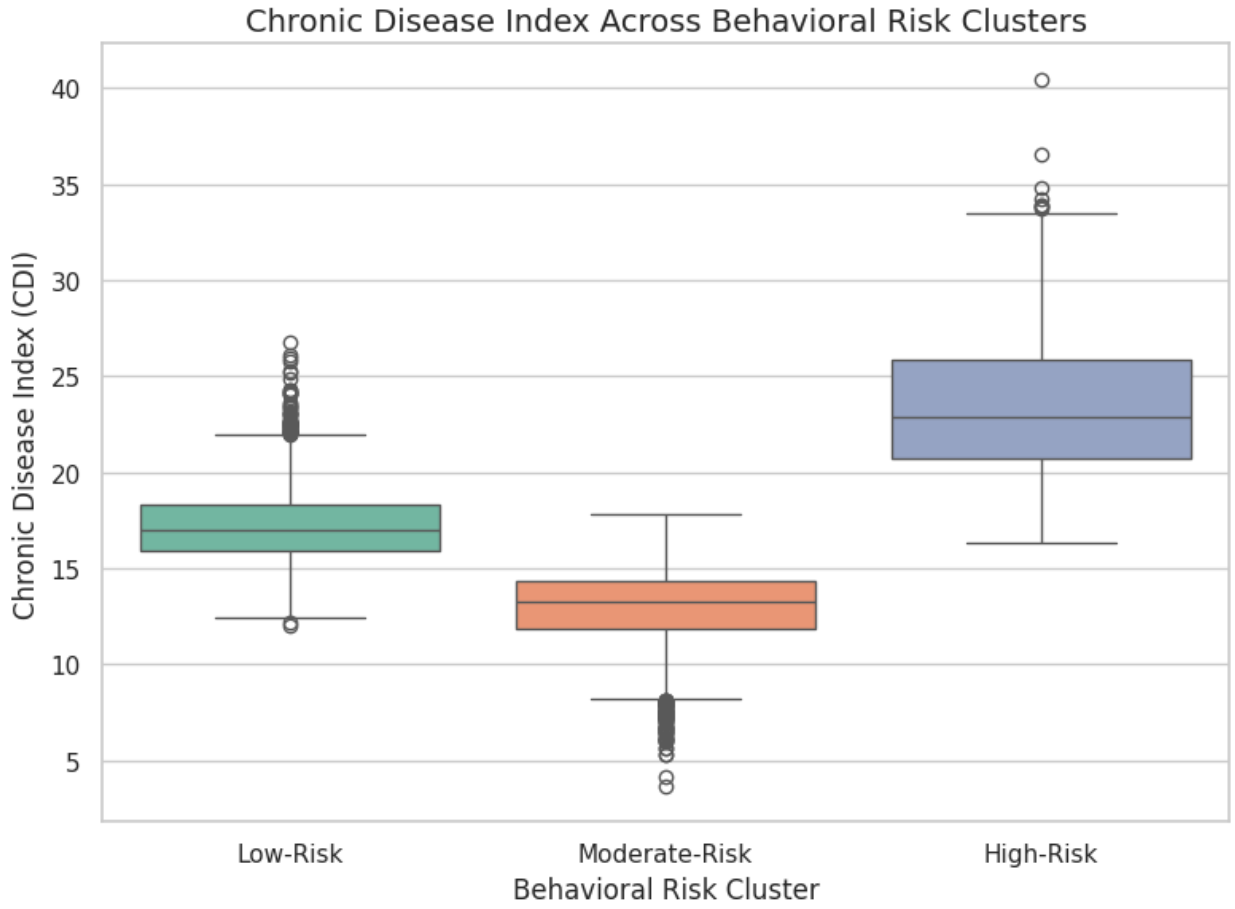

Figure 18: Boxplot Of Chronic Disease Index By Cluster

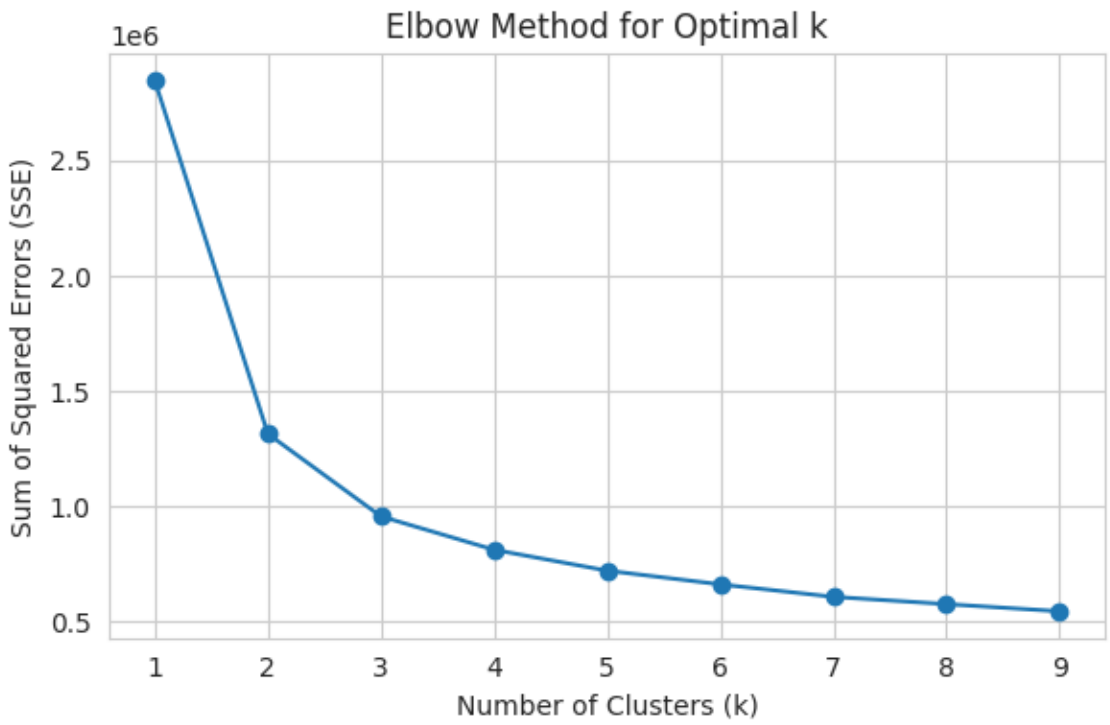

Figure 18: Elbow Method for Optimal K

To further show how behavioral risk patterns vary across clusters, we used a radar chart to display the average values of each risk factor for the three groups. This visualization highlights how Cluster 3, for instance, exhibits elevated levels across multiple dimensions, while other clusters show partial or low-risk profiles.

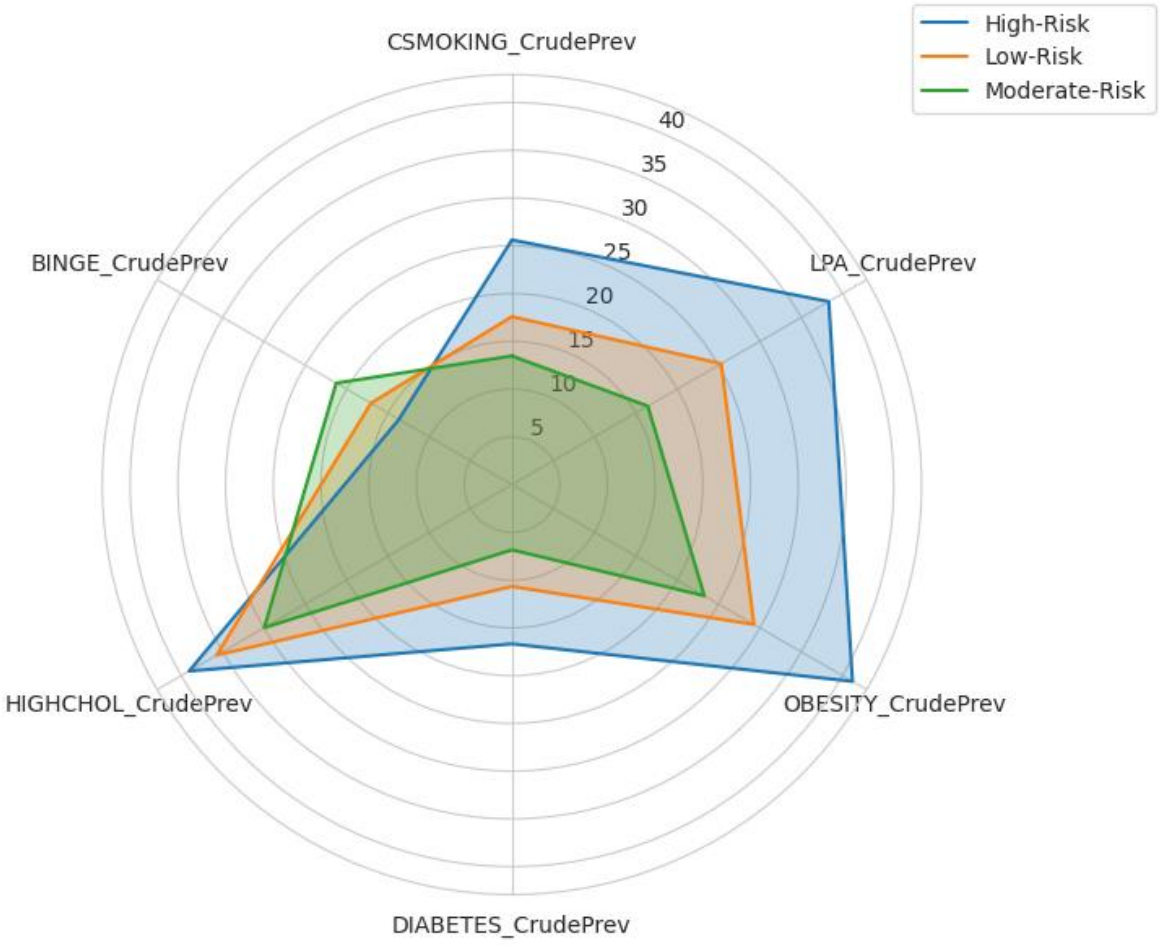

Figure 19: Radar (Spider) Chart Behavioral Profile of Each Cluster Across All 6 Features

8. FEATURE IMPORTANCE ANALYSIS (RANDOM FOREST)

The results show that being physically inactive and smoking were the strongest predictors of chronic disease. Check-ups, obesity, and diabetes also played a strong role. These results aligned with what we saw during EDA previously and with previous work by Qi et al. (2023) and Rahelić et al. (2024), which concluded these behaviors were the root causes of illness.

These aspects hold strong applicability to our research questions. In the instance of RQ1 and RQ3, the results show that higher individual risks and combined behavior patterns can increase levels of

illness. RQ4 is also supported, since checkup frequencies ranked as one of the strongest influences—although we can see from analysis that for most areas, checkups take place with increased frequency because people already are sick, not for the prevention of illness.

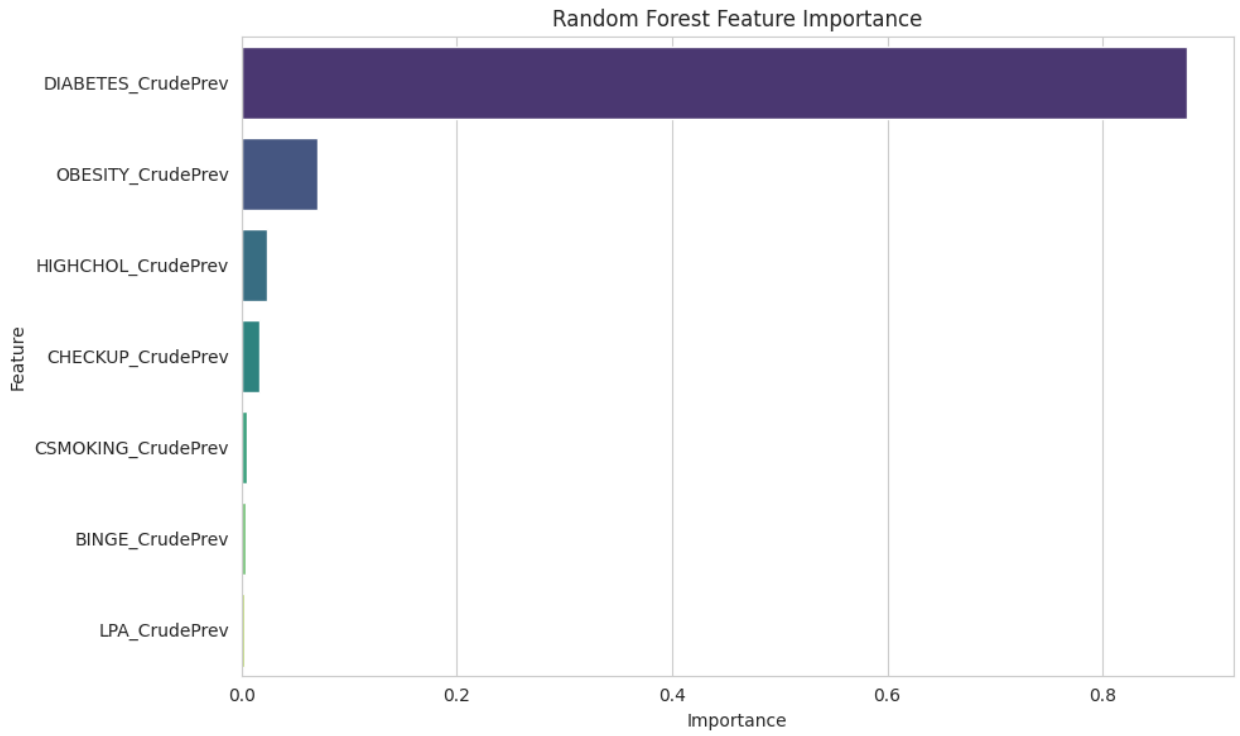

Figure 19: Random Forest Feature Importance

9. LIMITATIONS

Despite such insightful results produced by this analysis, several limitations impact its generalizability. First, the analysis used 2018 CDC 500 Cities cross-sectional data, such that causal or temporal associations cannot be supported through inferences. It would have tracked behavioral risks as well as preventive care's dynamics over time through longitudinal data.

Secondly, though tract-level census data is useful for local investigation, tracts can have within-tract variation like household variation. The variables of highest relevance, for example, the presence of care or socioeconomic stressors, were not included directly, potentially limiting the models' explanatory power.

Third, although both Random Forest and Support Vector Regression performed well, both have challenges in explainability. There are no intrinsic feature importance measures in SVR, and while Random Forest is transparent, it can be tricky for non-technical stakeholders to explain.

Fourth, the K-Means clustering method presumed uniformly shaped clusters and did not consider spatial associations between tracts. This could constrain the reliability of interpretations within geographically near regions.

Fifth, although spatial identifiers such as FIPS codes were available, this study did not conduct geographic mapping or spatial autocorrelation analysis due to scope and resource limitations.

Finally, key health determinants like diet, air quality, systemic inequality, and mental health were not included in the dataset. Having such variables could make subsequent studies even more robust as well as enable more effective targeting of public health interventions.

These restrains present avenues of future work in the direction of employability of spatiotemporal data, richer community measurements, as well as explainable AI for enhanced prediction as well as policy applicability.

10. CONCLUSION

The project analyzed the joint impact of health behaviors, access to preventive care, and the characteristics of the population on chronic disease outcomes in urban census tracts in the United

States. The results clearly found that smoking, obesity, lack of physical activity, and diabetes were all consistent and strong predictors of chronic disease burden. Preventive care measures such as checkup rates were also an important factor, though our results note that they tended to be more reflective of reactive care—in which individuals receive medical attention once symptoms have developed—than of proactive, community-based disease prevention.

Exploratory analyses uncovered that risk drivers tend to cluster in understandable geographic and demographic patterns. Using K-Means clustering, we discovered that numerous neighborhoods have shared behavioral profiles, while those with intersecting high-risk patterns had consistently the highest rates of chronic disease. This reinforces the position that targeting individual behavior in silos could be less efficient than working with the combined risk environment within communities.

We used two strong machine learning models for prediction purposes—Random Forest and Support Vector Regression (SVR). Random Forest performed exceedingly well in determining complex interactions as well as variable importance, providing performance as well as explainability. SVR showed the highest predictive power with an R^2 value of 0.9941 and the lowest error rates in all measures. This indicates that patterns in the underlying data were complex and subtle, for which SVR excelled in capturing.

These findings directly answer all our research questions and corroborate major public health studies (i.e., Qi et al., 2023; Rahelić et al., 2024), supporting the notion that chronic disease is not caused by an individual variable, but an interdependent combination of behavioral, socioeconomic,

as well as healthcare system variables. Additionally, our application of clustering and predictive modeling offers an actionable set of guidelines for health agencies to identify priority areas of highest risk, distribute resources in an efficient manner, and target community-level interventions. From public health and policy considerations, the effort highlights the significance of holistic, evidence-based strategies that impact individual actions as well as the structural health barriers. The results of the work reinforce the importance of enhancing preventive care not only in terms of frequency but also in terms of equity and access.

On the basis of these projections, subsequent studies can build on this work through the use of longitudinal data, spatial autocorrelation models, or other social and environmental drivers such as housing stability, air quality, or access to healthy foods. The addition of these variables could make projections of chronic disease both more precise and useful in real-world contexts.

As such, what is demonstrated through the project is that by bringing health behavior data, preventive care patterns, and characteristics of populations together with advanced modeling, the actionable insights generated make chronic disease prevention both more precise and effective amongst urban populations.

11. REFERENCES

Centers for Disease Control and Prevention. (1999). *Chronic diseases and their risk factors: The nation's leading causes of death.* https://www.cdc.gov/chronic-disease/data-surveillance/index.html

Centers for Disease Control and Prevention. (2024). *Chronic disease data and surveillance.* https://www.cdc.gov/chronic-disease/data-surveillance/index.html

Jafleh, E. A., Alnaqbi, F. A., Almaeeni, H. A., Faqeeh, S., Alzaabi, M. A., & Al Zaman, K. (2024). *The role of wearable devices in chronic disease monitoring and patient care: A comprehensive review. Frontiers in Public Health, 12*, Article 1347561. https://doi.org/10.3389/fpubh.2024.1347561

Klein, H. E. (2023, June 22). *Diabetes prevalence expected to double globally by 2050. The American Journal of Managed Care.* https://www.ajmc.com/view/diabetes-prevalence-expected-to-double-globally-by-2050

Pan, M., Li, R., Wei, J., Peng, H., Hu, Z., Xiong, Y., Li, N., Guo, Y., Gu, W., & Liu, H. (2024). *Application of artificial intelligence in the health management of chronic disease: Bibliometric analysis. Frontiers in Medicine, 11*, 1506641. https://doi.org/10.3389/fmed.2024.1506641

Panch, T., Szolovits, P., & Atun, R. (2018). *Artificial intelligence, machine learning, and health systems. Journal of Global Health, 8(2),* 020303. https://doi.org/10.7189/jogh.08.020303

Qi, M., Santos, H., Pinheiro, P., McGuinness, D. L., & Bennett, K. P. (2024). *Demographic and socioeconomic determinants of access to care: A subgroup disparity analysis using new equity-focused measurements. National Library of Medicine.* https://pmc.ncbi.nlm.nih.gov/articles/PMC10653411/

Rahelić, V., Perković, T., Romić, L., Perković, P., Klobučar, S., Pavić, E., & Rahelić, D. (2024). *The role of behavioral factors on chronic diseases—Practice and knowledge gaps. Healthcare, 12(24),* 2520. https://doi.org/10.3390/healthcare12242520

Singareddy, S., Prabhu, V. S. N., Jaramillo, A. P., Yasir, M., Iyer, N., Hussein, S., & Nath, T. S. (2023). *Artificial intelligence and its role in the management of chronic medical conditions: A systematic review. Cureus, 15(9),* e46066. https://pmc.ncbi.nlm.nih.gov/articles/PMC10607642/

Smith, J. P., & Harvey, P. J. (2011). *Chronic disease and infant nutrition: Is it significant to public health? Public Health Nutrition, 14(2),* 279–289. https://doi.org/10.1017/S1368980010001953

Turner, K., & Hohman, K. H. (2024). *Demonstrated progress and future promise of chronic disease data modernization. Preventing Chronic Disease, 21*, E240396. https://doi.org/10.5888/pcd21.240396

Wiemken, T. L., & Kelley, R. R. (2020). *Machine learning in epidemiology and health outcomes research. Annual Review of Public Health, 41*, 21–36. https://doi.org/10.1146/annurev-publhealth-040119-094437

World Health Organization. (2024). *Noncommunicable diseases.* [Fact sheet]. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

Gatzweiler, F. W., Jayasinghe, S., Siri, J. G., & Corburn, J. (2023). *Towards a new urban health science*. *International Science Council*. https://council.science/blog/towards-a-new-urban-health-science/

Ashburner, J. M., Horn, D. M., et al. (2017). *Chronic disease outcomes from primary care population health program implementation*. *The American Journal of Managed Care, 23*(12). https://www.ajmc.com/view/chronic-disease-outcomes-from-primary-care-population-health-program-implementation