Comparing Associations Of Chronic Health Outcomes with Social Determinants of Health (SDoH) Indices Using Machine Learning

Vandana Gupta
School of Computing
University of Connecticut
Storrs, CT 06269, USA
vandana.gupta@uconn.edu

Swapna S. Gokhale
School of Computing
University of Connecticut
Storrs, CT 06269, USA
swapna.gokhale@uconn.edu

ABSTRACT

Chronic health outcomes require ongoing medical attention and have a significant impact on a person's quality of life. It is widely accepted that social determinants of health (SDoH) shape the onset and management of chronic health outcomes. Among the many composite indices that assess SDoH, there is no consensus on which index best explains these associations between health outcomes and social determinants. Furthermore, chronic outcomes may be modulated by place or geography both through cultural, social, and political forces and spatial correlations. The novelty of this paper lies in building a machine learning (ML) methodology to compare the strengths of SDoH indices in explaining the associations between chronic health outcomes and social determinants while adjusting for geography. The methodology is illustrated by studying the relative strengths of the Social Vulnerability Index (SVI) and Social Deprivation Index (SDI) in explaining age-adjusted prevalence rates of 12 chronic health outcomes obtained from the CDC PLACES project. Results suggest that the SVI is more strongly associated with all 12 chronic health outcomes, however, the increase in the strength of SVI over SDI varies across the health outcomes. For each outcome, importance scores of all SVI measures are then normalized according to its four sub themes, while introducing geography/place as the fifth sub theme. Comparing the relative importance of these five sub themes leads to a grouping of the outcomes into three clusters depending on whether geography/place, racial minority status, or socio-economic measures shows the greatest impact. The emergence of geography as a dominant sub theme alongside conventional social determinants underscores the value of our approach in providing the capability to consider the modulating effect of geography on understanding the relationships between social determinants and health.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Applied computing** → **Life and medical sciences**; • **Information systems** → *Geographic information systems.*

SpatialEpi'24, October 29-November 1 2024, Atlanta, GA, USA

ACM ISBN 979-8-4007-1153-4/24/10
https://doi.org/10.1145/3681777.3698469

KEYWORDS

Chronic Health Outcomes, Social Vulnerability Index, Social Deprivation Index, Machine Learning, Random Forests, Geography, Racial Minority

ACM Reference Format:
Vandana Gupta and Swapna S. Gokhale. 2024. Comparing Associations Of Chronic Health Outcomes with Social Determinants of Health (SDoH) Indices Using Machine Learning . In *5th ACM SIGSPATIAL International Workshop on Spatial Computing for Epidemiology (SpatialEpi'24), October 29-November 1 2024, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3681777.3698469

1 INTRODUCTION & MOTIVATION

Chronic health outcomes refer to the long-term effects and complications resulting from chronic diseases such as diabetes, heart disease, chronic respiratory diseases, and cancer [1]. These conditions often require ongoing medical attention and management, significantly impacting patients' quality of life and longevity [2]. Chronic health conditions cause 1.7 million deaths annually, accounting for more than 70% of all deaths and costing the healthcare system about $3.8 trillion each year [3, 4].

Social determinants of health (SDoH), which are non-medical factors including the conditions in which people are born, grow, work, live, worship and age affect a wide range of chronic health outcomes [5, 6]. SDoH is a broad umbrella framework that intends to consider comprehensive parameters of a person's life circumstances that can impact their health. Yet, there is no universal agreement on which measures should be included under this framework. Studies that link SDoH to health vary widely in the factors they consider. Some include only personal socio-economic status (SES), typically characterized by income, education, and occupation/unemployment [7, 8], showing that individuals with lower personal SES are more likely to report poorer health, have a shorter life expectancy, and propensity to chronic diseases. Some studies show the link between neighborhood-level conditions and chronic health outcomes [9]. Some others contend that a combination of social and economic conditions of individuals and neighborhoods better explains chronic health outcomes [10, 11].

In recent years, many composite indices have been developed to succinctly capture population demographics and socio-economic conditions in order to understand and address disparities in health outcomes. These composite indices include: Area Deprivation Index (ADI) [10], Neighborhood Deprivation Index (NDI) [12], Social Vulnerability Index (SVI) [13], and Social Deprivation Index (SDI) [14].

These indices differ in their broad goals, in the measures they include, how these measures are defined and assessed, and their approaches to compose these measures into an aggregate index. These wide differences make it difficult to determine which index holds the ability to best explain the associations between social determinants and a range of health outcomes [12]. The impact of social determinants may also be modulated by geography/place, which ties social, political and cultural forces as well spatial correlations to health outcomes [15]. The inclusion of geography complicates SDoH analysis, and hence, is mostly ignored even though it is apparent that the conditioning effect of geography must be considered to produce honest estimates of how SDoH measures impact health outcomes.

This paper presents a machine learning (ML) approach that can compare how strongly SDoH indices can explain the associations between social determinants and health outcomes while adjusting for geography. We illustrate our approach by comparing how strongly the Social Deprivation Index (SDI) [14] and the Social Vulnerability Index (SVI) [16] can explain associations between SDoH measures and age-adjusted prevalence rates for 12 chronic health outcomes from the CDC PLACES project [4]. Our methodology is based on random forests, a flexible, non-parametric, ensemble ML model that can include geo-coordinates as independent model covariates. Our results indicate that the SVI more strongly links SDoH to chronic health outcomes compared to the SDI, however, the difference in the strength of association is not uniform across all the health outcomes. We then aggregate and normalize the importance scores of the SDoH measures from the SVI into its four sub themes, introducing geography/place as the fifth sub theme. A comparison of the relative importance of these five sub themes divides the 12 health outcomes into three groups. The first group comprises of outcomes most dominantly impacted by geography/place, the second group comprises of outcomes primarily affected by racial minority status, and the third group includes outcomes tied to socio-economic status in the most dominant manner. Given that geography is as strongly tied to a subset of health outcomes alongside socio-economic and racial minority status, further underscores the importance of our approach that offers the capability to account for the moderating influence of geography in SDoH analysis.

The rest of the paper is organized as follows: Section 2 outlines our ML approach. Section 3 provides a detailed description of the three data sets. Section 4 presents results and discussion. Section 5 compares and contrasts related work. Section 6 concludes and presents future directions.

2 ML METHODOLOGY

Our methodology is based on random forests (RF) [17] as the foundational model. RF is an aggregate ML technique; it combines predictions from multiple individual decision trees to produce a more accurate result. Each tree is trained on a random subset of the training data (bootstrapping) and a random subset of the input features. Each tree estimates the outcome, and the final estimate is the average of all the estimates produced by the trees. This approach reduces overfitting and increases robustness compared to a single decision tree. Fig. 1 shows how a RF comprises a collection of decision trees.

In our approach, for n SDoH indices and m health outcomes we build a collection of $m \times n$ RF models. The RF model for each index includes the measures in that index as explanatory variables and the health outcome as the outcome variable. Additionally, we also include the latitude and longitude of the centroids of chosen geographical units (county, census tract, ZCTA) as predictor variables as this is the most effective way of considering geographical variations [18]. RFs are robust and flexible to incorporate diverse types of covariates such as SDoH measures and geo-coordinates in a single modeling framework. For each health outcome, the performance of the RF models for all SDoH indices is compared using the coefficient of determination R^2, which indicates the proportion of variance explained, with higher values suggesting better explanatory power. The index with the highest R^2 will have the strongest ability to explain the associations between social determinants and that health outcome.

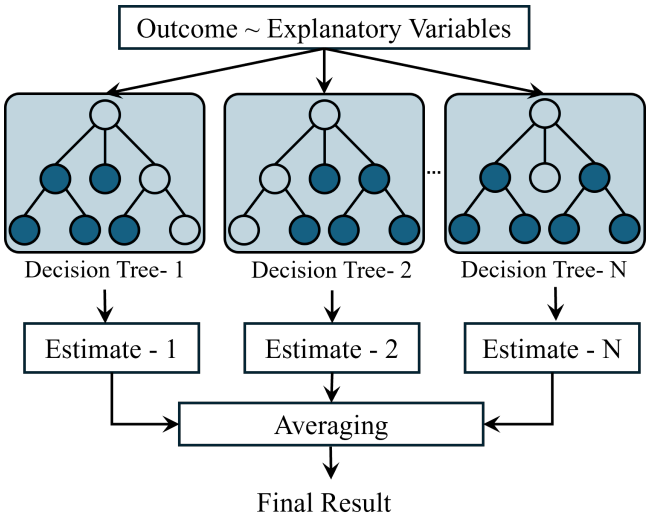

Figure 1: Schematic Representation of Random Forests

We chose random forests because of their interpretability in that they expose the importance of each SDoH measure in explaining a given health outcome. This indicates how much each SDoH measure contributes to the accuracy of the model. The importance of each SDoH measure is computed by measuring how much each measure decreases the impurity in the trees, averaged over all trees in the forest. Thus, for each health outcome, once the strongest SDoH index is identified, the relative importance of the measures comprising that index can be further revealed using our random forest-based approach.

3 DATA DESCRIPTION

We illustrate our approach using data from three public-domain sources: the PLACES (Population Level Analysis and Community Estimates) project [19], Social Deprivation Index (SDI) [14] and Social Vulnerability Index (SVI) [16]. County-level data for 3041 counties from the contiguous United States is pulled from all three sources. We chose county-level data, since for the initial illustration

we felt that census-tract level data may be too detailed. This section provides an overview and also presents an exploratory analysis of these data sets.

3.1 Chronic Health Outcomes

PLACES is a collaboration between the Centers for Disease Control and Prevention (CDC), the Robert Wood Johnson Foundation, and the CDC Foundation [19]. It is the source of age-adjusted prevalence rates of the following 12 chronic health outcomes estimated in adult population over 18 years of age [19], [20].

- **Arthritis:** Affects millions, leads to joint pain and disability that severely impacts daily activities and quality of life [21].
- **High Blood Pressure (High BP):** Major risk factor for heart disease and stroke, significantly affects global mortality and healthcare costs [22].
- **Asthma:** Causes chronic inflammation of the airways, leading to severe breathing difficulties that affect work and physical activity, and result in high healthcare utilization [23].
- **Cancer:** Leading cause of death worldwide, cancer impacts various body parts and necessitates extensive healthcare support [24].
- **Coronary Heart Disease (CHD):** Most common type of heart disease, leading to heart attacks and significant morbidity [25].
- **Chronic Obstructive Pulmonary Disease (COPD):** Includes a group of lung diseases that block airflow and make breathing difficult, significantly reducing quality of life and increasing healthcare costs [26].
- **Depression:** Leading cause of disability, this mental health disorder decreases productivity and increases disability [27].
- **Diabetes:** Affecting blood sugar processing, diabetes is a major cause of complications such as blindness, kidney failure, heart attacks, and stroke [28].
- **High Cholesterol (High Chol):** Leads to atherosclerosis, heightens the risk of CHD and stroke, and is a significant silent risk factor for cardiovascular diseases [29].
- **Kidney Disease:** Causes gradual loss of kidney function, linked to severe health complications and high medical costs.
- **Obesity:** Increases the risk of chronic diseases like diabetes, heart disease, and certain cancers, and is a leading preventable cause of death [30].
- **Stroke:** Leading cause of death and long-term disability, occurring when blood flow to an area of the brain is interrupted, necessitating extensive rehabilitation and imposing significant healthcare costs [31].

Table 1 summarizes the descriptive statistics of 12 chronic outcomes. Significant disparities in the prevalence and impact of the health outcomes can be observed. High blood pressure and obesity have the highest prevalence rates, with high standard deviations indicating considerable variability across regions. High cholesterol, arthritis, and depression show moderate prevalence, with arthritis and high cholesterol showing similar variability, while depression has more regional variation. Cancer and kidney disease have low prevalence and variability, suggesting a more uniform impact. Asthma and stroke also show moderate prevalence with low variability, indicating stable distribution. This illustrates that while some conditions like high blood pressure and obesity are widespread with substantial variation, others like cancer and kidney disease are more localized and consistent.

Fig. 2 represents the geographical locations of the extreme (minimum and maximum) prevalence rates of the 12 health outcomes. Each location represents the centroid (latitude and longitude) of a county. The blue circles present locations with minimum prevalence rates, while the red circles show locations with the maximum prevalence rates. Maximums (and minimums) of multiple health outcomes overlap at some locations, and the sizes of the bubbles indicate the number of overlapping outcomes at each location, the larger the bubble the higher is the number of health outcomes with extreme values at that location. In Loving County, Texas (TX), there are three minimum rates for asthma, cancer and kidney disease. Conversely, in Todd County, South Dakota (SD) four maximum rates are concentrated, corresponding to CHD, diabetes, kidney disease, and stroke.

Table 1: Descriptive Statistics - Chronic Health outcomes

Health outcomes	Mean	Std.Dev (SD)
Arthritis	24.99	2.7
High BP	60.31	3.68
Asthma	10.38	0.99
Cancer	6.22	0.29
CHD	5.89	0.85
COPD	7.21	1.63
Depression	23.09	3.25
Diabetes	10.57	2.25
High Chol	31.33	2.31
Kidney	2.94	0.36
Obesity	37.46	4.52
Stroke	3.07	0.54

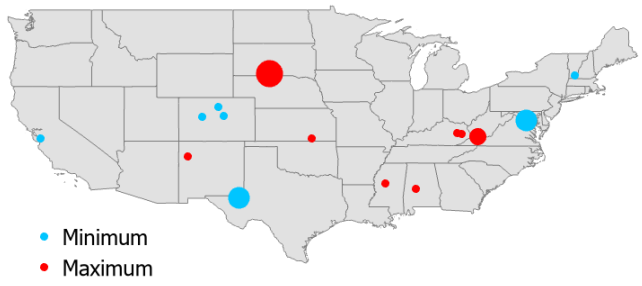

Figure 2: Extremes Prevalence of Chronic Health Outcomes

3.2 SDoH Indices

In this section, we review Social Deprivation Index (SDI) and Social Vulnerability Index (SVI) for the initial illustration of our approach. The SDI serves as a representative index that narrowly focuses on socio-economic disadvantage, and the SVI as a representative that includes broad characteristics beyond socio-economic conditions that make communities more vulnerable. We considered

Area Deprivation Index (ADI) and Neighborhood Deprivation Index (NDI) as alternatives to SDI. However, the ADI may be both under and over inclusive compared to the SDI. The ADI is over inclusive because it contains multiple measures to model a single type of disadvantage (for example, three measures for poverty and two for education). At the same time, it is under inclusive because it still focuses only on deprivation and disadvantage. A similar issue arises with NDI, which focuses mostly on multiple metrics of disadvantage. Moreover, both the ADI and the NDI are estimated at the level of census tracts, and as discussed in Section 3 for our initial illustration we prefer county-level data. The SDI is parsimonious in including disadvantage measures and estimated at the county level as a more compact representation of social deprivation [32–34].

SDI evaluates area-level deprivation using seven metrics from the American Community Survey (ACS) [14], while the SVI assesses the capacity of communities to respond to external stresses on public health such as natural disasters, economic downturns, and disease outbreaks [13, 35]. In addition to social deprivation, the SVI also introduces measures such as housing cost burden, lack of health insurance, age-related vulnerabilities (for both elderly and young), disability, language proficiency, housing structure type, racial minority status, mobile homes, and group quarters [11]. In total, the SVI includes 16 measures, also collected from the American Community Survey [36].

Table 4 provides a side by side comparison of the SDI and SVI measures highlighting overlaps and key distinctions. Abstractly, the SDI and SVI share 6 measures: poverty, education, unemployment, crowded units, single-parent households, and no vehicle. The two indices, however, differ in how some of these are defined and estimated. SDI considers individuals below 100% of the Federal Poverty Level (FPL), whereas SVI uses a higher threshold of 150% FPL, potentially identifying a larger group of economically vulnerable people. Similarly, SDI and SVI both include unemployment, but SVI defines it as non-employed persons aged 16 and older, while SDI limits the age range to 16-64 years and includes people who are not in the labor force. Each SDI and SVI measure is calculated as a percentage of individuals or households or families exhibiting that characteristic.

Table 2 summarizes descriptive statistics of SDI measures. Non employment has the highest mean, and also the highest spread, suggesting very high unemployment rates in some regions coupled with very low unemployment rates in some others. On an average, about one-fifth live in renter units, and the spread is more modest. About one sixth live in poverty, and lack high school diploma. Both measures have moderate standard deviations, which indicates regional disparities. On an average, a very small percentage lack access to a vehicle and live in crowded units. Both these measures also enjoy low standard deviations, indicating low but uniform impact. Single parent households fall somewhere in between the highest and the lowest measures both for the mean and spread. Fig. 3 represents the locations of the extremes of SDI measures. The maximums and minimums do not overlap, and are distributed across different counties, indicating that no county is particularly replete with or significantly devoid of resources and opportunities. Most of the lowest SDI measures occur in the Midwest, and neither the maximums nor the minimums can be found in the West.

Table 2: Descriptive Statistics - SDI

SDI outcomes	Mean	Std. Dev (SD)
Poverty	15.11	6.33
Education	13.06	6.26
Non-Employed	46.52	21.06
Renter-Units	28.31	8.10
Crowded-Units	2.31	1.90
Single-Parent Households	12.40	4.10
No-Vehicle	6.13	3.64

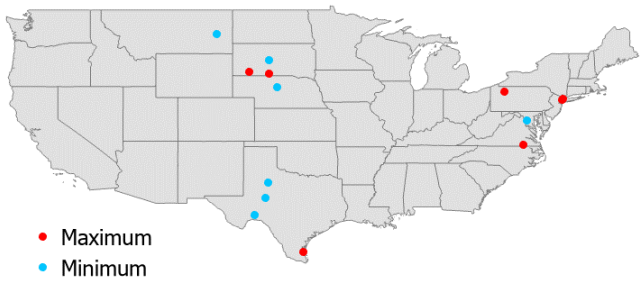

Figure 3: Extremes of SDI Measures

Table 3 summarizes the descriptive statistics of SVI measures. Percentages of population living in poverty and holding racial minority status have very high means and high standard deviations suggesting that some counties may be homogeneous and affluent, whereas, some counties may house a sizeable racial minority population living in poverty. Means of housing cost burden and proportion of population under the age of 17 is also high, but the standard deviation is comparatively modest suggesting that these vulnerabilities are consistent and pervasive across the country. Finally, the percentage of population lacking education has a modest mean and moderate variability. On the lower end, crowding, living in group quarters, non employment, and single parent families have low means with relatively low standard deviations, suggesting that these vulnerabilities are less prevalent but more consistent across regions. Fig. 4 represents the locations of the extremes of SVI measures [16]. Similar to SDI, most extreme SVI measures are spread but unlike SDI there is some overlap. New York, NY registers highest values of two measures of vulnerability; lack of access to a vehicle and multi-unit housing. Only one county in the southwest (Loving County, TX) shows the lowest values with respect to poverty and single-parent households.

Differences in how each index defines the measures manifest as inconsistencies in their descriptive statistics in Tables 2 and 3. For example, non employment is the SDI measure with the highest mean, yet it is the SVI measure with the lowest mean, because SDI computes this measure over a limited age range of 16-64 years, whereas, SVI includes the entire adult population while calculating the percentage unemployed. Similarly, SVI's more aggressive definition of poverty raises the mean percentage compared to SVI. Finally, the approach used by SDI to calculate the composite index

Table 3: Descriptive Statistics - SVI

SVI outcomes	Mean	Std.Dev (SD)
Poverty	24.51	8.48
Non-Employed	5.15	2.53
Housing Cost Burden	22.24	5.21
Education	12.40	6.04
No Insurance	9.40	5.05
Age 65 and older	19.23	4.65
Age 17 and younger	22.13	3.48
Disability	16.00	4.50
Single Parent Households	5.88	2.38
English Proficiency	1.58	2.66
Racial Minority	23.68	19.92
Multi-Unit Structures	4.68	5.70
Mobile Homes	12.47	9.27
Crowded-Units	2.27	1.91
No-Vehicle	6.03	3.66
Group Quarters	3.39	4.34

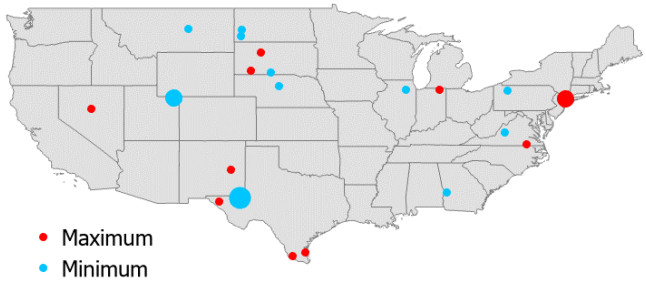

Figure 4: Extreme Values of SVI Measures

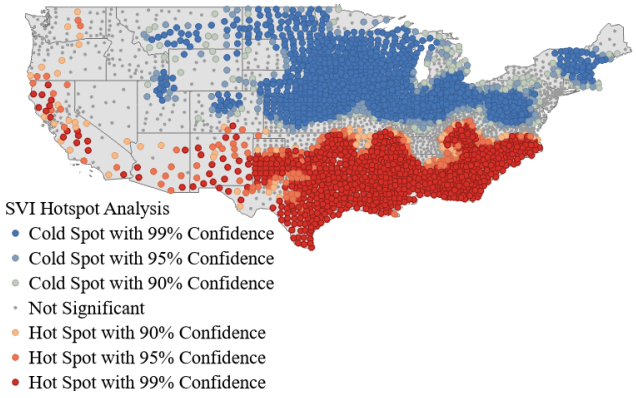

Figure 5: SVI – Hot Spots and Cold Spots

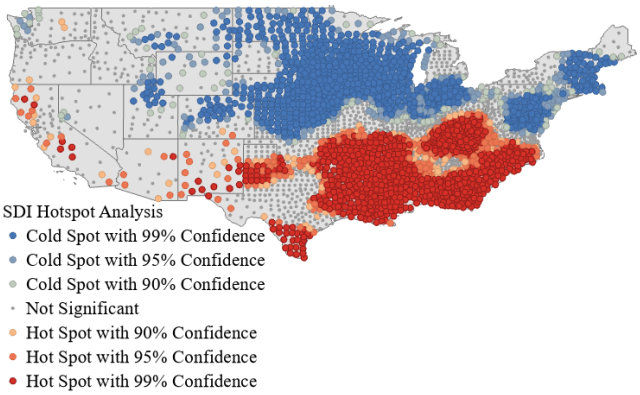

Figure 6: SDI – Hot Spots and Cold Spots

differs from the approach used by the SVI to compose the aggregate index [35].

Spatial spreads of SVI and SDI, analyzed using hot spots and cold spots [37], represent areas with statistically significant clustering. Hot spots indicate regions with high concentrations of SVI and SDI indicate higher social vulnerabilities, while cold spots represent areas with low concentrations of SVI and SDI indicating lower social vulnerabilities as shown in Fig. 5 and Fig. 6 respectively. These figures reveal similar spatial distributions, with both indices identifying significant hot spots in the southeastern states, such as Alabama, Mississippi, and Louisiana, and parts of the Southwest, including New Mexico and Arizona. Conversely, cold spots according to both indices are predominantly found in the northern regions, particularly in Midwestern states of Minnesota, North Dakota, and Wisconsin. There is a notable 9.5% increase in the area of cold spots for SVI compared to SDI which suggests that SVI may encompass a broader range of challenges which the SDI does not fully capture. Sizeable spatial overlap in hot spots between SVI and SDI can be attributed to their shared variables, such as poverty, education, and employment. Social deprivation may be thus a critical driver of social vulnerability compared to other factors such as disability, minority status, and language barriers.

Next, we formalize the spatial relationship observed visually using local bivariate analysis [37], which seeks to determine whether the relationship between any two variables is statistically significant and how their values depend on each other or vary over geographical space. It is conducted by calculating an entropy statistic in each local neighborhood to measure how much information the two variables share [37]. For the sake of convenience, we designate SVI as the dependent variable, and SDI as the explanatory variable [38]. Each region is classified into the five types of relationships between the SDI and SVI: (i) Positive Linear, (ii) Negative Linear, (iii) Positive Concave, (iv) Positive Convex, and (v) Not significant.

As shown Fig. 7, the analysis classifies the 3041 counties into four classes. Specifically, 2,479 counties, an overwhelming 80% exhibit a positive linear relationship which is spread across the United States, which indicates that in these areas, higher SVI is associated with higher SDI. 135 counties demonstrated a concave relationship which means that while SVI increases as SDI increases, the rate of increase slows. Initially, higher SDI leads to significant increases in SVI, but as deprivation continues to rise, its impact on SVI becomes less substantial. Positive concave relationship was mainly observed in the East South Central and East North Central regions. 396 counties exhibited a convex relationship. A convex relationship is the opposite of concave: the increase in SVI starts

Table 4: SDI vs SVI

Social Deprivation Index (SDI)	Social Vulnerability Index (SVI)
1. **Poverty** - % of individuals less than 100% FPL. 2. **Education** - % of individuals 25 years or more with less than 12 years of education. 3. **Unemployment** - % of non-employed individuals aged 16-64 years. 4. **Crowded Units** - % of occupied housing units with more people than rooms. 5. **Single Parent** - % of single-parent families with dependents under 18. 6. **No Vehicle** - % of households with no vehicle available. 7. **Renter Units** - % of households living in renter-occupied housing units.	1. **Poverty** - % of individuals below 150% FPL. 2. **Education** - % of individuals with no high school diploma (age 25+). 3. **Unemployment** - % of non-employed individuals aged 16+. 4. **Crowded Units** - % of occupied housing units with more people than rooms. 5. **Single Parent** - % of single-parent households with children under 18. 6. **No Vehicle** - Percentage of households with no vehicle available. 7. **Housing Cost Burden** - Percentage of housing cost-burdened occupied housing units with annual income less than $75,000. 8. **No Health Insurance** - % of uninsured in the total civilian non-institutionalized population. 9. **Persons 65 years of age or older** - % of individuals aged 65 and older. 10. **Persons 17 years of age or younger** - % of individuals aged 17 and younger. 11. **Disability** - % of civilian non-institutionalized population with a disability. 12. **English Language Proficiency** - % of individuals (age 5+) who speak English "less than well". 13. **Multi-Unit Structures** - % of housing in structures with 10 or more units. 14. **Racial Minority** - % of minority (Hispanic or Latino (of any race); Black and African American, Not Hispanic or Latino; American Indian and Alaska Native, Not Hispanic or Latino; Asian, Not Hispanic or Latino; Native Hawaiian and Other Pacific Islander, Not Hispanic or Latino; Two or More Races, Not Hispanic or Latino; Other Races, Not Hispanic or Latino). 15. **Mobile Homes** - % of mobile homes. 16. **Group Quarters** - % of individuals in group quarters.

slowly but then accelerates as SDI continues to rise. In these areas, at low SDI, the impact on SVI is small, but as SDI increases, the effect on SVI becomes more substantial. This pattern was observed across the West North & South Central and Middle & South Atlantic areas. Only 31 counties do not show any significant relationship between SDI and SVI. Finally, not a single county shows negative linear relationship. The Pearson correlation coefficient between SDI and SVI is 0.89, further supporting the strong positive relationship in most regions.

SVI, being more comprehensive than SDI, is likely to be more strongly linked to health outcomes. However, the use of SVI in public health studies has raised legal concerns due to its inclusion of race [39]. Should the use of SVI be prohibited, the relative strength of associations between SDI and SVI must be understood to determine whether the SDI can serve as an alternative to SVI. Moreover, whether the SDI and SVI can be used interchangeably across all the health outcomes and also across the entire United States, or within some narrow regions and a subset of the health outcomes needs

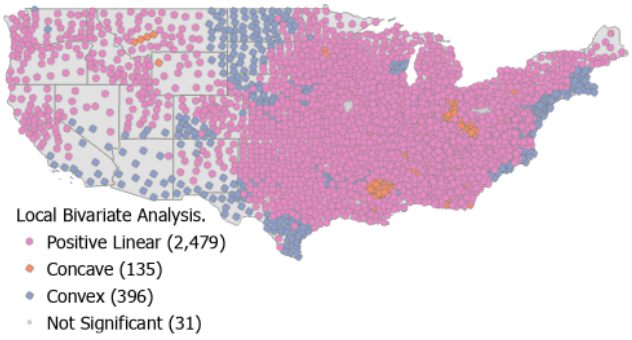

Figure 7: Bivariate Analysis – SVI and SDI

to be determined. Our methodology can be used to explore these questions and define the contours of such investigations.

4 RESULTS & DISCUSSION

Our comprehensive approach considers the associations between SDI and SVI individually with each of the 12 chronic health outcomes, with and without geography using a collection of 48 random forest models. The coefficient of determination (R^2) of these models, is summarized in Table 5. For all health outcomes, SVI shows higher R^2, indicating better explanation of the associations between SDoH and health outcomes. However, this improvement is moderated by geography; incorporating geo-coordinates reduces the percentage improvement in R^2 of SVI significantly, ranging from twice to more than five times. We observe that the impact of SVI is not uniform across all health outcomes; the greatest improvement of about 20% is observed for cancer, a modest improvement of about 9% for kidney disease, obesity, arthritis, and COPD, and 6% or lower improvement for the remaining health outcomes.

Because the SVI outperforms SDI, we choose the SVI to determine the relative significance of the individual SDoH variables through importance analysis. We organize the importance scores of the 16 SVI measures according to their sub themes as shown in Figure 8. These sub themes include Socio-Economic Characteristics (SES), Household Characteristics (HC), Racial Minority Status (RM), and Access to Transportation & Housing (HT) as shown in Fig 8. A normalized importance score for each sub theme is calculated by pooling the scores for all the measures in each sub theme and dividing the pooled score by the total number of measures in that sub theme. We also introduce a fifth sub theme named "Place," (PL) which pools and normalizes the importance scores of latitude and longitude. Fig. ?? shows the relative importance of the five sub themes, and according to the most dominant sub theme, the 12 health outcomes can be classified into three groups.

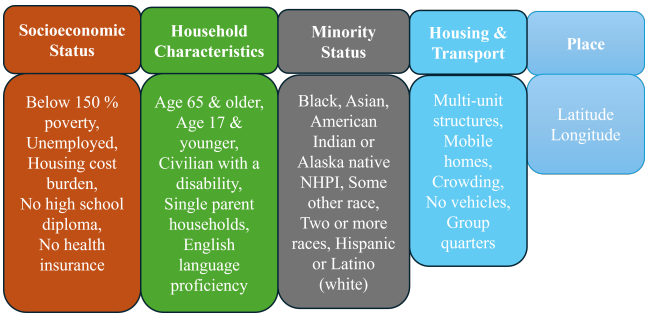

Figure 8: SVI Themes

- **Group 1: Geography/Place:** For arthritis, high blood pressure, asthma, obesity, and high cholesterol, geography/place emerges as the most dominant sub theme. The three outcomes (high blood pressure, obesity and high cholesterol) that are precursors to coronary heart disease are dominantly associated with geography rather than socio-economic status unlike many other studies have shown [40]. The standing of the four SVI sub themes relative to geography is similar for all four chronic outcomes except high cholesterol. For high cholesterol, the importance of place is heightened compared to the other SVI sub themes, which also rationalizes the lowest improvement shown by SVI over SDI. High blood pressure also shows only a marginal improvement for SVI.
- **Group 2: Racial Minority Status:** Being a racial minority has the highest impact for cancer, depression, diabetes and kidney disease. Significant health disparities among different racial minority groups may arise due to unequal access to healthcare, chronic stress, and discrimination. The impact of racial minority status is not uniformly pronounced across the four health outcomes. Place is nearly as important as racial minority status for depression and socio-economic status is nearly as important as racial minority status for diabetes and kidney disease. For cancer, racial minority status overwhelmingly dominates the four other sub themes. This also sheds light on why SVI shows the highest improvement in explaining the associations between SDoH and cancer.
- **Group 3: Socio-economic Status:** For COPD, stroke and CHD, socio-economic status is the most significant. Socio-economic status presents a much higher association with coronary heart disease compared to the other four sub themes, however, for COPD the other four themes are not as insignificant, in fact racial minority status falls right below socio-economic status.

Other than racial minority status, place and socio-economic status emerge as the most important SDoH drivers for many health outcomes, which indicates that the physical and social environment and economic conditions, such as income and education, play a crucial role in a person's health. Although the two other sub themes of SVI, namely, housing costs and access to affordable transportation are linked to chronic health outcomes, they never appear as the topmost predictors. Overall, our results underscore the necessity of considering the moderating effect of place on producing an accurate picture of how social determinants influence health outcomes. Moreover, SDI may be a viable alternative to SVI at least for those health outcomes where SVI offers only marginal improvement, and in those regions where there is a significant overlap between the two indices.

5 RELATED RESEARCH

In this section, we summarize related works that link social deprivation and social vulnerability indices to health outcomes.

Many studies have shown that SDI is closely associated with a range of adverse health outcomes such as cardiovascular disease and related mortality [41], diabetes and respiratory diseases [42, 43], cancer [44, 45], and arthritis [46]. Social deprivation is also linked to mental health disorders, including depression and anxiety, as it often leads to psychological stress, likely due to the compounding effects of economic hardship and social isolation [47].

Although the SVI was conceived to assess the resilience of a community to man-made and natural disasters [16], recent studies have linked the SVI to health outcomes [48–51]. SVI has been shown to relate to heart disease [52], heart disease related mortality among individuals with diabetes particularly younger adults [53, 54], elevated mortality rates from cancer [55, 56] and asthma-related hospital visits and higher asthma prevalence [57, 58]. SVI is further

Table 5: Performance Comparison of Random Forest Models

Health Outcomes	Mean	RF Baseline			RF + Geo-Coordinates		
		SVI R-Squared	SDI R-Squared	% Gain	SVI R-Squared	SDI R-Squared	% Gain
Arthritis	24.98	0.72	0.59	22.03	0.81	0.74	9.46
High BP	32.78	0.82	0.73	12.33	0.91	0.88	3.41
Cancer	6.22	0.82	0.50	64.00	0.89	0.75	18.67
Asthma	10.38	0.60	0.46	30.43	0.78	0.74	5.41
CHD	5.89	0.86	0.77	11.69	0.88	0.83	6.02
COPD	7.21	0.86	0.72	19.44	0.90	0.83	8.43
Depression	23.09	0.61	0.36	69.44	0.75	0.71	5.63
Diabetes	10.57	0.91	0.80	13.75	0.93	0.87	6.90
High Chol	31.33	0.50	0.45	11.11	0.75	0.73	2.74
Kidney	2.94	0.94	0.84	11.90	0.95	0.87	9.20
Obesity	37.46	0.61	0.47	29.79	0.72	0.66	9.09
Stroke	3.07	0.93	0.85	9.41	0.94	0.88	6.82

linked to mortality and cognitive impairment in older adults, even after adjusting for age, sex, and frailty [50, 59–61]. Few studies also consider sub themes of the SVI, reporting that socio-economic status and household composition and disability were most strongly associated with poor postoperative outcomes among patients [62]. The relationship between social vulnerability and COVID-19 outcomes have also been studied, finding that increased social vulnerability correlates with higher rates of COVID-19 cases [63], and incidence [64, 65].

Despite the extensive use of SDI and the emerging use of SVI to associate SDoH with health outcomes, very few works directly compare the SDI and SVI [12]. Our research fills this gap by comparing the relative strengths of the SDI and SVI in explaining the associations between social determinants and age-adjusted prevalence rates of 12 chronic health outcomes while adjusting for geography. It identifies a subset of chronic health outcomes and regions for which the SDI may be a viable alternative to SVI in understanding how social determinants shape chronic health.

6 CONCLUSION AND FUTURE RESEARCH

This paper presents a ML based approach to compare the relative strengths of SDoH indices in explaining the associations between social determinants and a range of chronic health outcomes while controlling for geography. The approach exploits the flexibility offered by random forests, an ensemble learning paradigm to consider different types of covariates in a single framework. The methodology is illustrated by comparing the relative merits of the Social Deprivation Index (SDI) and Social Vulnerability Index (SVI) in explaining the associations between social determinants and age-adjusted prevalence rates of 12 chronic health outcomes from the CDC PLACES project. Our results indicate that the SVI, although conceived to assess the resilience of communities to natural and man-made disasters, is more effective at explaining these associations. However, the relative strength of the SVI over SDI varies across the chronic health outcomes. These findings, combined with the spatial relationships between the SDI and SVI suggest that the SDI may be used in place of SVI for those health outcomes where the improvement is low to modest, and in regions where there is a strong positive relationship between these two indices. This finding is significant especially because the SVI may be banned from public health studies due to its inclusion of race as a component measure. The importance scores of the SVI measures are normalized into its four sub themes, with geography being introduced as the fifth sub theme. The health outcomes are then classified into three groups, depending on which sub theme emerges as the most dominant. The first group ties chronic health outcomes to place/geography as the dominant sub theme, the second group ties them to racial minority status, and the third group ties them to socio-economic conditions. The emergence of geography/place as one of the dominant sub themes alongside SDoH measures underscores the value of our approach which allows an integrated consideration of geography with SDoH measures to honestly assess their relevance.

Our future research involves applying our approach on other indices such as the Area Deprivation Index (ADI) and Neighborhood Deprivation (NDI), for a more detailed census-tract level analysis. We also plan to investigate how SDoH measures can be linked to health-risk behaviors and the use of preventive health services.

ACKNOWLEDGEMENTS

This research is supported by a seed grant from the Institute for Collaboration on Health, Intervention, and Policy, University of Connecticut, 2006 Hillside Road, Storrs, CT, 06269-1248, USA.

REFERENCES

[1] U. Bauer, P. Briss, R. Goodman, and B. Bowman, "Prevention of chronic disease in the 21st century: Elimination of the leading preventable causes of premature death and disability in the usa," *Lancet*, vol. 384, 07 2014.
[2] WHO, "Non communicable diseses." https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases, 2023. Accessed: 2024-07-15.
[3] CDC, "Deaths and mortality." https://www.cdc.gov/nchs/fastats/deaths.htm, 2024. Accessed: 2024-07-25.
[4] CDC, "Health and economic costs of chronic conditions." https://www.cdc.gov/chronic-disease/data-research/facts-stats/index.html, 2024. Accessed: 2024-07-25.
[5] CDC, "Social determinant of health outcomes." https://www.cdc.gov/about/priorities/why-is-addressing-sdoh-important.html, 2024. Accessed: 2024-07-25.

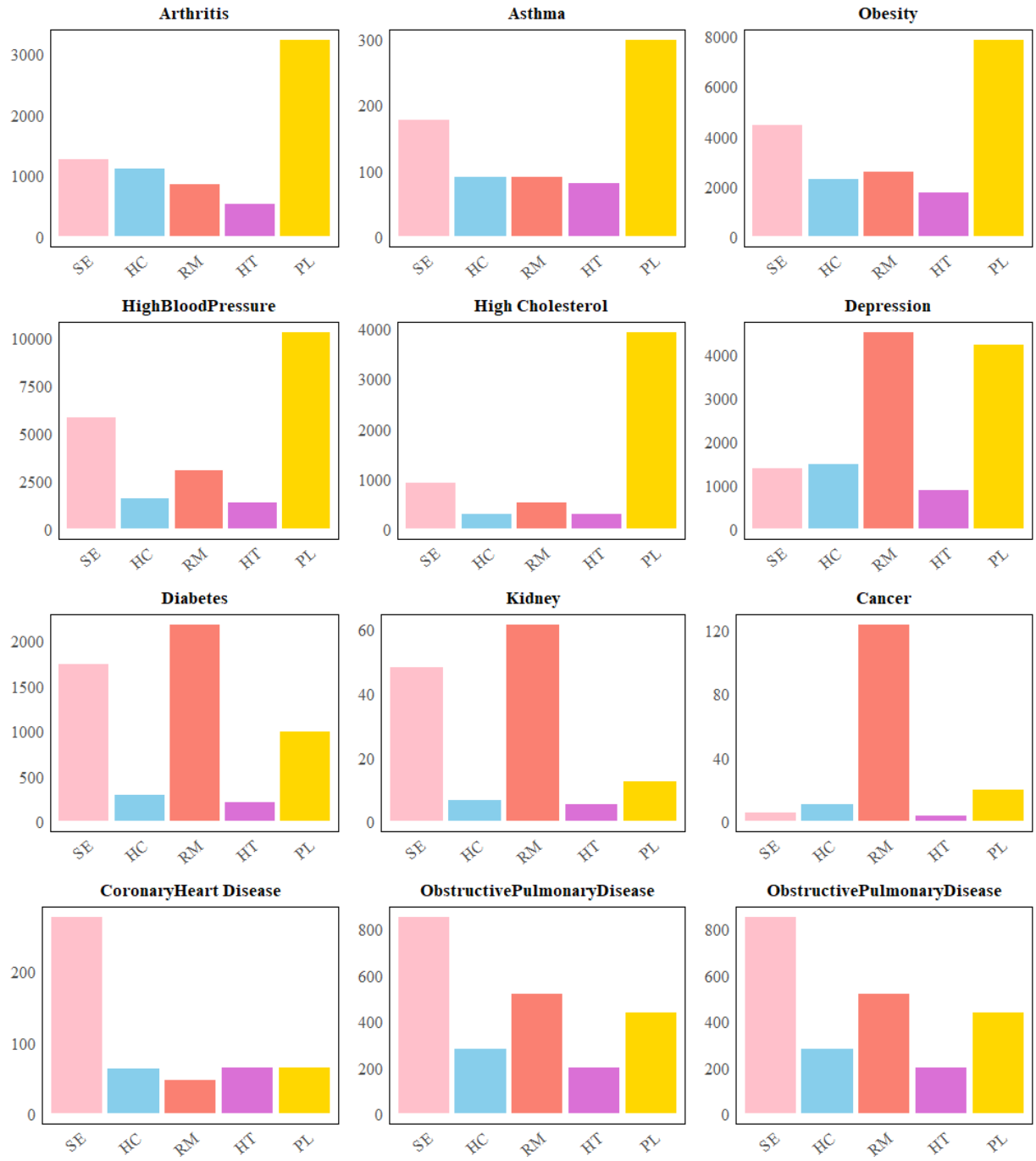

Figure 9: Importance of SVI Themes

[6] NAC, "National academic of sciences." https://nap.nationalacademies.org/catalog/25467/integrating-social-care-into-the-delivery-of-health-care-moving, 2024. Accessed: 2024-07-25.

[7] N. Arpey, A. Gaglioti, and M. Rosenbaum, "How socioeconomic status affects patient perceptions of health care: A qualitative study," *Journal of Primary Care Community Health*, vol. 8, p. 215013191769743, 07 2017.

[8] C. Barakat and T. Konstantinidis, "A review of the relationship between socioeconomic status change and health," *International journal of environmental research and public health*, vol. 20, 06 2023.

[9] F. Boscoe, B. Liu, and F. Lee, "A comparison of two neighborhood-level socioeconomic indexes in the united states," *Spatial and Spatio-temporal Epidemiology*, vol. 37, p. 100412, 02 2021.

[10] T. Kim, "Relationship of neighborhood and individual socioeconomic status on mortality among older adults: Evidence from cross-level interaction analyses," *PloS one*, vol. 17, p. e0267542, 05 2022.

[11] M. Mujahid, S. Maddali, X. Gao, K. Oo, L. Benjamin, and T. Lewis, "The impact of neighborhoods on diabetes risk and outcomes: Centering health equity," *Diabetes care*, vol. 46, 06 2023.

[12] C. Park, T. Schappe, S. Peskoe, D. Mohottige, N. Chan, N. Bhavsar, L. Boulware, J. Pendergast, A. Kirk, and L. Mcelroy, "A comparison of deprivation indices and application to transplant populations," *American Journal of Transplantation*, vol. 23, 01 2023.

[13] P. ONE, "Association between sdi and svi." https://journals.plos.org/plosone/article/figures?id=10.1371/journal.pone.0292281, 2023. Accessed: 2024-07-15.

[14] ArcGIS, "Social depriviation index." https://www.graham-center.org/maps-data-tools/social-deprivation-index.html, 2024. Accessed: 2024-07-20.

[15] N. Waters, "Tobler's first law of geography," 12 2017.

[16] CDC, "Atsdr svi data and documentation." https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html, 2022. Accessed: 2024-07-20.

[17] M. Schonlau and R. Zou, "The random forest algorithm for statistical learning," *The Stata Journal: Promoting communications on statistics and Stata*, vol. 20, pp. 3–29, 03 2020.

[18] T. Hengl, M. Nussbaum, M. Wright, G. Heuvelink, and B. Graeler, "Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables," *PeerJ*, vol. 6, p. e5518, 08 2018.

[19] CDC, "Places health data." https://www.cdc.gov/places/, 2023. Accessed: 2024-07-20.

[20] K. Greenlund, H. Lu, Y. Wang, K. Matthews, J. LeClercq, B. Lee, and S. Carlson, "Places: Local data for better health," *Preventing Chronic Disease*, vol. 19, 06 2022.

[21] Arthritis. https://www.cdc.gov/arthritis/index.htm, 2024.

[22] WHO, "Highbloodpressure." https://www.who.int/health-topics/hypertension#tab=tab_1, 2023. Accessed: 2024-07-24.

[23] CDC, "Asthma." https://www.cdc.gov/asthma/most_recent_national_asthma_data.htm, 2023. Accessed: 2024-07-15.

[24] WHO, "Cancer." https://www.who.int/health-topics/cancer#tab=tab_1, 2023. Accessed: 2024-07-15.

[25] NHS, "Coronary heart diasease." https://www.nhs.uk/conditions/coronary-heart-disease/, 2024. Accessed: 2024-07-15.

[26] CDC, "Obstructive pulmonary diasease." https://www.cdc.gov/copd/index.html#:~:text=What%20is%20COPD%3F,Americans%20who%20have%20this%20disease., 2024. Accessed: 2024-07-15.

[27] WHO, "Depression." https://www.who.int/health-topics/depression#tab=tab_1, 2023. Accessed: 2024-07-15.

[28] WHO, "Diabetes." https://www.who.int/health-topics/diabetes#tab=tab_1, 2023. Accessed: 2024-07-15.

[29] CDC, "High cholesterol." https://www.cdc.gov/cholesterol/index.htm, 2024. Accessed: 2024-07-15.

[30] CDC, "Obesity." https://www.cdc.gov/obesity/index.html, 2023. Accessed: 2024-07-15.

[31] CDC, "Chronic stroke." https://www.cdc.gov/stroke/about.htm#:~:text=A%20stroke%2C%20sometimes%20called%20a,term%20disability%2C%20or%20even%20death, 2023. Accessed: 2024-07-15.

[32] NDI, "Codebook – neighborhood deprivation index data." https://www.gis.cancer.gov/research/NeighDeprvIndex_Codebook.pdf, 2024. Accessed: 2024-07-25.

[33] K. Rollings, G. Noppert, J. Griggs, R. Melendez, and P. Clarke, "Comparison of two area-level socioeconomic deprivation indices: Implications for public health research, practice, and policy," *PLoS ONE*, vol. 18, 10 2023.

[34] CDC, "Preventing chronic disease." https://www.cdc.gov/pcd/issues/2016/16_0221.htm, 2024. Accessed: 2024-07-25.

[35] P. ONE, "Comparison of adi and svi items.." https://journals.plos.org/plosone/article/figures?id=10.1371/journal.pone.0292281, 2023. Accessed: 2024-07-15.

[36] ACS, "American county survey." https://www.census.gov/data/developers/data-sets/acs-5year.html, 2024. Accessed: 2024-07-25.

[37] ArcGIS, "How local bivaraiet analysis worls- arcgis pro." https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/learnmore-localbivariaterelationships.htm#:~:text=The%20Local%20Bivariate%20Relationships%20tool,relationships%20vary%20over%20geographic%20space., 2024. Accessed: 2024-07-20.

[38] ArcGIS, "Local bivariate analysis." https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/localbivariaterelationships.htm, 2024. Accessed: 2024-07-20.

[39] SVI, "Svi better for geographic areas." https://ihpi.umich.edu/sites/default/files/2021-06/ADIvSVI-COVID-MI_brief_Tipirneni_050521.pdf, 2024. Accessed: 2024-07-25.

[40] ACS, "Geographic variation in obesity." https://www.cdc.gov/pcd/issues/2021/21_0094.htm, 2024. Accessed: 2024-07-25.

[41] D. Butler, S. Petterson, R. Phillips, and A. Bazemore, "Measures of social deprivation that predict health care access and need within a rational area of primary care service delivery," *Health services research*, vol. 48, 07 2012.

[42] M. Mujahid, S. Maddali, X. Gao, K. Oo, L. Benjamin, and T. Lewis, "The impact of neighborhoods on diabetes risk and outcomes: Centering health equity," *Diabetes care*, vol. 46, 06 2023.

[43] K. Wang, C.-k. Law, J. Zhao, A. Y.-K. Hui, B. Yip, E. Yeoh, and R. Chung, "Measuring health-related social deprivation in small areas: development of an index and examination of its association with cancer mortality," *International Journal for Equity in Health*, vol. 20, 09 2021.

[44] Z. Balogun, L. Gardiner, J. Li, E. Moroni, M. Rosenzweig, and M. Nilsen, "Neighborhood deprivation and symptoms, psychological distress, and quality of life among head and neck cancer survivors," *JAMA otolaryngology– head neck surgery*, vol. 150, 02 2024.

[45] M. Arcaya, R. Tucker-Seeley, R. Kim, A. Schnake-Mahl, M. So, and S. Subramanian, "Research on neighborhood effects on health in the united states: A systematic review of study characteristics," *Social Science Medicine*, vol. 168, pp. 16–29, 08 2016.

[46] M. J. Harrison, K. J. Tricker, L. Davies, A. Hassell, P. Dawes, D. L. Scott, S. Knight, M. Davis, D. Mulherin, and D. P. M. Symmons, "The relationship between social deprivation, disease outcome measures, and response to treatment in patients with stable, long-standing rheumatoid arthritis.," *The Journal of Rheumatology*, vol. 32, no. 12, pp. 2330–2336, 2005.

[47] A. Orben, L. Tomova, and S.-J. Blakemore, "The effects of social deprivation on adolescent development and mental health," *The Lancet Child & Adolescent Health*, vol. 4, no. 8, pp. 634–640, 2020.

[48] S. Gokhale, "Comparing the impact of unhealthy behaviors and preventive services on chronic health outcomes," pp. 798–802, 12 2020.

[49] A. Ganguly, K. Alvarez, S. Mathew, V. Soni, S. Vadlamani, B. Balasubramanian, and K. Bhavan, "Intersecting social determinants of health among patients with childcare needs: a cross-sectional analysis of social vulnerability," *BMC Public Health*, vol. 24, 02 2024.

[50] J. Mah, J. Penwarden, H. Pott, O. Theou, and M. Andrew, "Social vulnerability indices: a scoping review," *BMC Public Health*, vol. 23, 06 2023.

[51] L. Wallace, O. Theou, F. Pena, K. Rockwood, and M. Andrew, "Social vulnerability as a predictor of mortality and disability: cross-country differences in the survey of health, aging, and retirement in europe (share)," *Aging Clinical and Experimental Research*, vol. 27, no. 3, pp. 365–372, 2015.

[52] G. Bevan, A. Pandey, S. Griggs, J. Dalton, D. Zidar, S. Patel, S. Khan, K. Nasir, S. Rajagopalan, and S. Al-Kindi, "Neighborhood-level social vulnerability and prevalence of cardiovascular risk factors and coronary heart disease," *Current Problems in Cardiology*, vol. 48, p. 101182, 03 2022.

[53] H. Bashar, O. Kobo, K. Khunti, A. Banerjee, R. Bullock-Palmer, N. Curzen, and M. Mamas, "Impact of social vulnerability on diabetes-related cardiovascular mortality in the united states," *Journal of the American Heart Association*, vol. 12, 10 2023.

[54] R. Pham, E. Gorodeski, and S. Al-Kindi, "Social vulnerability and location of death in heart failure in the united states," *Current Problems in Cardiology*, vol. 48, p. 101689, 03 2023.

[55] R. Ibrahim, E. Sainbayar, H. Nhat, M. Shahid, A. Saleh, Z. Javed, S. Khan, S. Al-Kindi, and K. Nasir, "Social vulnerability index and cardiovascular disease care continuum," *JACC: Advances*, p. 100858, 03 2024.

[56] S. Ganatra, S. Dani, A. Kumar, S. Khan, R. Wadhera, T. Neilan, P. Thavendiranathan, A. Barac, J. Hermann, M. Leja, A. Deswal, M. Fradley, J. Liu, D. Sadler, A. Asnani, L. Baldassarre, D. Gupta, E. Yang, A. Guha, and A. Nohria, "Impact of social vulnerability on comorbid cancer and cardiovascular disease mortality in the united states," *JACC: CardioOncology*, vol. 4, pp. 326–337, 09 2022.

[57] I. Aijazuddin, A. Alloghbi, and A. Sukari, "Associations between micro-geographic social vulnerability and disparities in cancer incidence.," *Journal of Clinical Oncology*, vol. 41, pp. e18523–e18523, 06 2023.

[58] A. Bakshi, A. Doren, C. Maser, K. Aubin, C. Stewart, S. Soileau, K. Friedman, and A. Williams, "Identifying louisiana communities at the crossroads of environmental and social vulnerability, covid-19, and asthma," *PLOS ONE*, vol. 17, p. e0264336, 02 2022.

[59] M. Andrew and K. Rockwood, "Social vulnerability predicts cognitive decline in a prospective cohort of older canadians," *Alzheimer's dementia : the journal of the Alzheimer's Association*, vol. 6, pp. 319–325.e1, 07 2010.

[60] M. Andrew, A. Mitnitski, and K. Rockwood, *Social vulnerability, frailty and mortality in elderly people*, pp. 89–105. 04 2016.

[61] M. Andrew, A. Mitnitski, S. Kirkland, and K. Rockwood, "The impact of social vulnerability on the survival of the fittest older adults," *Age and ageing*, vol. 41, pp. 161–5, 03 2012.

[62] H. Labiner, M. Hyer, J. Cloyd, D. Tsilimigras, D. Dalmacy, A. Paro, and T. Pawlik, "Social vulnerability subtheme analysis improves perioperative risk stratification in hepatopancreatic surgery," *Journal of Gastrointestinal Surgery*, vol. 26, 01 2022.

[63] E. Biggs, P. Maloney, A. Rung, E. Peters, and W. Robinson, "The relationship between social vulnerability and covid-19 incidence among louisiana census tracts," *Frontiers in Public Health*, vol. 8, 01 2021.

[64] B. Neelon, F. Mutiso, N. Mueller, J. Pearce, and S. Benjamin Neelon, "Spatial and temporal trends in social vulnerability and covid-19 incidence and death rates in the united states," *medRxiv : the preprint server for health sciences*, 09 2020.

[65] S. Kim and W. Bostwick, "Social vulnerability and racial inequality in covid-19 deaths in chicago," *Health Education Behavior*, vol. 47, p. 109019812092967, 05 2020.