CRI: Health Sub-Index

1. Introduction

The objective of this project is to identify areas of need within all communities in the Dallas-Fort Worth metroplex and help end child poverty. Health is an important factor in determining the well-being of children both in current and later stages of life. Allocating the appropriate health resources would provide promise in the future of children's' lives to give them the best opportunity to succeed. The purpose of this report is to show the process of how we obtained our results for the Health Sub-Index. We initially obtained as much information as we could about the Health Sub-Index from the COI 2.0 paper and other online resources to give us a good foundation on existing methodologies and have a starting point for the type of data to look at. From the data we collected, we composed categories of indicators that would make up our Health Sub-Index. These categories include healthcare, physical health, mental health, environmental health, and traffic accidents. From these categories, we selected what we think are the best indicators to represent the health of a community. Along with our selection process, we developed a methodology for arriving at a final Health Score. Below we explain how we gathered, merged, imputed, standardized, weighted, scaled, and ultimately calculated the cumulative Health Score.

2. Data

a. Final Indicators

Table 1. Final Health Indicators, Descriptions, and Sources

	Indicator	Description	Source
HEALTHCARE	Health Insurance	Percentage individuals ages 0-64 with health insurance coverage	http://data.diversitydatakids.org/dataset/coi20-child-opportunity-index-2-0-database/resource/f16fff12-b1e5-4f60-85d3-3a0ededa30a0#dictionary_anchor
	Routine Check-Up	Model-based estimate for crude prevalence of visits to doctor for routine checkup within the past year among adults aged >=18 years	https://chronicdata.cdc.gov/500-Cities/500-Cities-Census-Tract-level-Data-GIS-Friendly-Fo/k86t-wghb
PHYSICAL HEALTH	Healthy Food	Percentage households without a car located further than a half-mile from the nearest supermarket	http://data.diversitydatakids.org/dataset/coi20-child-opportunity-index-2-0-database/resource/f16fff12-b1e5-4f60-85d3-3a0ededa30a0#dictionary_anchor
	Walkability	EPA Walkability Index	http://data.diversitydatakids.org/dataset/coi20-child-opportunity-index-2-0-database/resource/f16fff12-b1e5-4f60-85d3-3a0ededa30a0#dictionary_anchor
	High Blood Pressure	Model-based estimate for crude prevalence of high blood pressure among adults aged >=18 years	https://chronicdata.cdc.gov/500-Cities/500-Cities-Census-Tract-level-Data-GIS-Friendly-Fo/k86t-wghb
	Stroke	Model-based estimate for crude prevalence of stroke among adults aged >=18 years	https://chronicdata.cdc.gov/500-Cities/500-Cities-Census-Tract-level-Data-GIS-Friendly-Fo/k86t-wghb
MENTAL HEALTH	Sleep	Model-based estimate for crude prevalence of sleeping less than 7 hours among adults aged >=18 years	https://chronicdata.cdc.gov/500-Cities/500-Cities-Census-Tract-level-Data-GIS-Friendly-Fo/k86t-wghb
	Alcohol Use	Model-based estimate for crude prevalence of binge drinking among adults aged >=18 years	https://chronicdata.cdc.gov/500-Cities/500-Cities-Census-Tract-level-Data-GIS-Friendly-Fo/k86t-wghb
ENVIRONMENT	Vacancy	Percentage housing units that are vacant	http://data.diversitydatakids.org/dataset/coi20-child-opportunity-index-2-0-database/resource/f16fff12-b1e5-4f60-85d3-3a0ededa30a0#dictionary_anchor
	Superfund Site	Average number of Superfund sites within a 2-mile radius, converted to natural log units	http://data.diversitydatakids.org/dataset/coi20-child-opportunity-index-2-0-database/resource/f16fff12-b1e5-4f60-85d3-3a0ededa30a0#dictionary_anchor
TRAFFIC	Pedalcyclist	Texas Peace Officer's Crash Reports (CR-3) that involved a pedalcyclist that were received and processed by TxDOT	https://cris.dot.state.tx.us/public/Query/app/public/query/results
	Drug Accidents	Traffic accidents involving drugs	https://cris.dot.state.tx.us/public/Query/app/public/query/results

Healthcare Indicators

We grouped medical check-up and insurance data as indicators of healthcare access and usage. Since there is not enough data about this for children, we used certain adult level data that would most closely be representative of the children's health care as well. A good health insurance availability and routine check-up visits would mean that children have access to healthcare facilities as well.

Physical Health Indicators

This group represents physical health of individuals and the measure of opportunity for good health in an area. Children born into areas without sufficient open public spaces or access to healthy food would be at a higher risk of adverse outcomes. Furthermore, a family history of high blood pressure and stroke would most likely be indicative of poor health of the child.

Mental Health Indicators

Happier adults raise happier children. Since mental well-being of the adults in a family would most definitely shape that of the children, we have included data on the average hours of sleep and alcohol use to assess the mental health of the individuals in an area. Additionally, children tend to adopt the lifestyle habits of their parents as they become adults.

Environmental Indicators

The environmental indicators represent the important information about the neighborhoods that affect the quality of life of a child that lives in it. We have included housing vacancy, as a high percentage would show that people do not want to live in the area and are related to higher crime and reduced feelings of safety. Superfund site is the presence of hazardous waste dumps in the nearby area.

Traffic Indicators

Poorer neighborhoods are known to have less safe road designs. There are studies that show that there were significantly more accident related injuries in poorer areas. Hence, traffic indicators would be a good measure of safety. Data about accidents involving drugs was also included in this indicator.

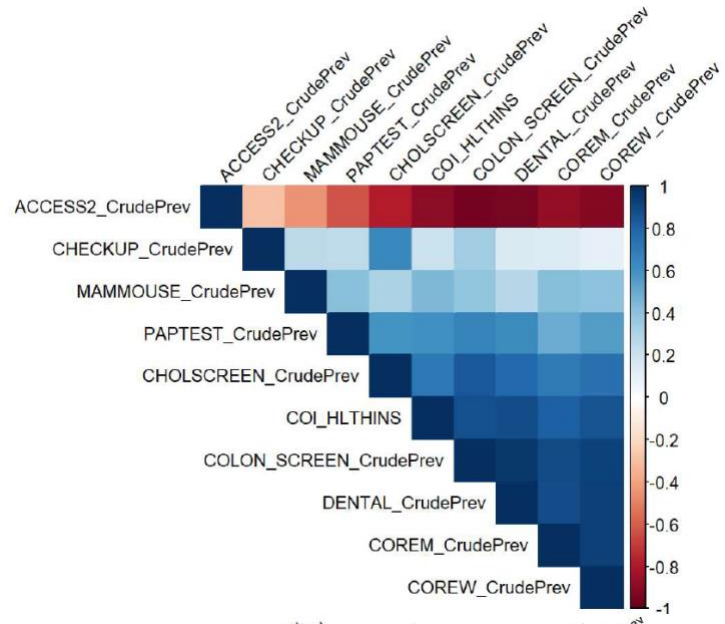
b. Indicator Selection

In our first iteration of indicator selection, our method was to run correlation plots for each of our four categories at the time - healthcare, physical health, mental health, and environmental. With such a large number of potential indicators, it would have been difficult to interpret a single correlation plot. Therefore we ran a separate plot for each category as there is a high chance of similar indicators having high correlation.

From these plots, we selected indicators based on correlation. We wanted to limit the possibility of collinearity issues, so we focused on indicators with low correlation with other indicators. We also considered an indicator's theoretical impact on childrens' health. Therefore in cases where many indicators were highly correlated (particularly in physical health), we discussed and chose the indicator we thought was most likely to affect a child's health and well-being, and which was a good representative for the group of highly correlated indicators. **Table 2** shows our initially selected health indicators. The referenced correlation plots are in **Figures 1-4**.

Table 2. Initial Health Indicators

	Indicator	Description
Healthcare	**Health insurance coverage**	Percentage individuals ages 0-64 with health insurance coverage.
	Routine check ups	Model-based estimate for crude prevalence of visits to doctor for routine checkup within the past year among adults aged >=18 years
Physical Health	**Access to healthy food**	Percentage households without a car located further than a half-mile from the nearest supermarket.
	Access to green space	Percentage impenetrable surface areas such as rooftops, roads or parking lots.
	Walkability	EPA Walkability Index.
	Obesity	Model-based estimate for crude prevalence of obesity among adults aged >=18 years
Mental Health	**Housing vacancy rate**	Percentage housing units that are vacant.
	Alcohol use	Model-based estimate for crude prevalence of binge drinking among adults aged >=18 years
	Smoking	Model-based estimate for crude prevalence of current smoking among adults aged >=18 years
Environmental	**Hazardous waste dump sites**	Average number of Superfund sites within a 2-mile radius, converted to natural log units.
	Industrial pollutants in air, water or soil	Index of toxic chemicals released by industrial facilities, converted to natural log units.
	Airborne microparticles	Mean estimated microparticle (PM2.5) concentration.
	Ozone concentration	Mean estimated 8-hour average ozone concentration.
	Extreme heat exposure	Summer days with maximum temperature above 90F.



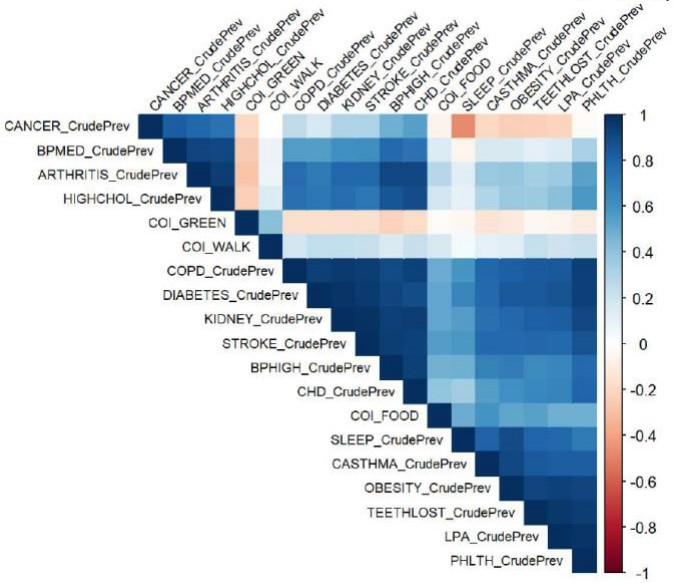
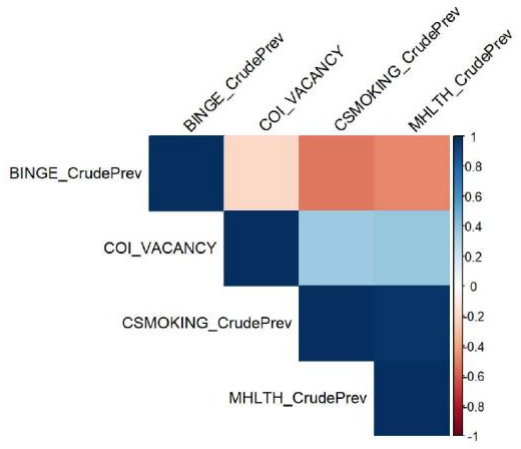


Figure 1. Correlation of Healthcare Indicators **Figure 2. Correlation of Physical Health Indicators**



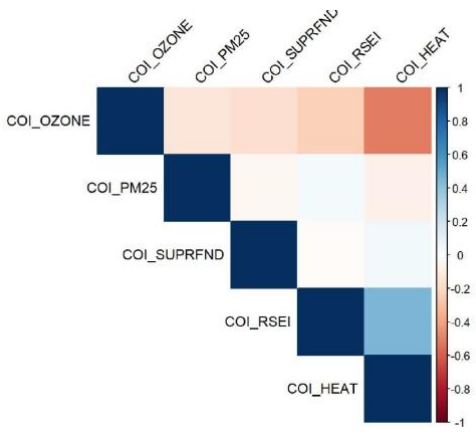


Figure 3. Correlation of Mental Health Indicators **Figure 4. Correlation of Environmental Indicators**

We improved our indicator selection during a second iteration. During this iteration, we used a multiple linear regression model with life expectancy as the dependent variable. Life expectancy is also one of the outcome variables we used to calculate indicator weights, as discussed later in **Section 3e**. We chose to use life expectancy, rather than physical health issues or mental health issues (our other two outcome variables in **Section 3e**), for indicator selection because we believed life expectancy to be a more holistic measurement of health.

We start with a simple base model of life expectancy. We then perform a sort of stepwise loop where we add one variable at a time. The new variable must meet two criteria. First, it should increase R_2 by the highest impact. Second, the coefficient should be greater than 0 (have a positive effect on life expectancy) but less than 1 (not dominate our model). We are working with z-score data at this point (see **Section 3d**), so a coefficient of 1 is very big. The loop runs until introducing a new variable no longer improves R_2.

After we arrived at a model, we removed some insignificant variables, investigated collinearity issues, and re-ran the model. Lastly, we regressed physical health issues on the same selected indicators to check the reasonableness of the results. Our final indicators are as listed in **Table 1**.

3. Method

a. Gather & merge source data

Our final Health Score includes data from three different data sources - COI, 500 Cities Project, TX DoT, and CDC. In order to work with the data, we first needed to merge all data based on GeoID. Each 11-digit GeoID represents a census tract.

b. Resolve duplicate census tracts

After merging the data from different sources, we found that we had duplicate census tracts. In one of our datasets, they would list a GeoID twice if a census tract spanned across more

than one city. In order to arrive at one row per GeoID, we took a weighted average of the data in those rows based on 2015 population.

c. Impute missing data

Upon merging the data, we also found that we had cases of missing data. We imputed missing data to be equal to the nearest census tract. We used a script which calculated the distance between latitude and longitude points to determine the nearest census tract. If the nearest census tract also happened to be missing the same data point, the script would find the next nearest census tract. Details on the technicalities of this can be found in our Python code.

d. Standardize data

In order to both calculate indicator weights and arrive at a final cumulative Health Score, we needed to get all the indicators on the same scale and in the same direction. For this we normalized all the indicators using z-score standardization. Then we multiplied many of the indicators by -1 to get all the indicators in the same direction; i.e. as the indicator value increases from low to high, it indicates a more desirable score. Lastly, to limit extreme values, we winsorized the upper bounds of the data at 95% for each indicator.

List of variables which required a change in directionality:
Healthy food
High blood pressure
Stroke
Sleep
Alcohol use
Vacancy
Superfund sites
Pedalcyclists
Drug accidents

e. Calculate weights

To calculate the weights of each indicator on the overall Health Score we thought of regressing these indicators on factors which closely represent good and poor health, which is what our Health Score aims to measure. We chose life expectancy, physical health issues, and mental health issues as the factors that, taken in combination, closely mirror our ideal Health Score.

We followed the below mentioned steps to come up with final weights of indicators:

1. We scaled all the indicator variables using z-score standardization as mentioned in **Section 3d**. From here on we refer to these indicators as X variables.
2. We scaled the three factors which closely represent the health score using z-score standardization. From here on we refer to these factors as Y variables.
3. We regressed each Y variable on all X variables and calculated the coefficients of each X variable. The value of each coefficient can be found in **Table 3** below.
4. To combine the three different coefficients we derived for each indicator, we used a weighted average method. We used weights of 0.5 for life expectancy coefficient, 0.25 for physical health coefficient, and 0.25 for mental health coefficient. We applied a

higher weight to life expectancy because we think that it better represents the overall health of a community. The weighted average of coefficients can be found in **Table 3** under the Weighted Average column.

5. From the weighted average of coefficients, we calculated the final weights by scaling these weighted averages in such a way that they all add up to be 1.

Table 3. Indicator-Outcome Coefficients, Weighted Average Weights, and Final Weights

Indicator		Life Expectancy	Physical Health	Mental Health	Weighted Average	Final Weight
HEALTHCARE	Health Insurance	0.310	0.280	0.266	0.292	0.129
	Routine Check-Up	0.337	0.471	0.627	0.443	0.195
PHYSICAL HEALTH	Healthy Food	0.089	-0.057	-0.015	0.027	0.012
	Walkability	0.093	-0.073	-0.044	0.017	0.008
	High Blood Pressure	0.239	0.531	0.297	0.326	0.144
	Stroke	0.275	0.506	0.411	0.367	0.162
MENTAL HEALTH	Sleep	0.308	0.385	0.754	0.439	0.193
	Alcohol Use	0.140	0.329	0.464	0.269	0.118
ENVIRONMENT	Vacancy	0.113	-0.038	-0.009	0.045	0.020
	Superfund Site	0.105	-0.061	-0.054	0.024	0.010
TRAFFIC	Pedalcyclist	0.065	-0.042	-0.042	0.011	0.005
	Drug Accidents	0.029	-0.014	-0.009	0.009	0.004

f. Re-scale data

Since all indicator values are in z-scores which are difficult to interpret, we re-scaled the data to be between 0 to 100 for all the indicators. For that we used a min-max scaler. This allows users to easily interpret a census tract's performance for an indicator.

The three tables below show the descriptive statistics for our data in each of its three forms - raw, standardized, and scaled.

Table 4. Descriptive Statistics for Raw Data

Variable Name	Mean	SD	Min	Max
Access to healthy food	3.36	4.7	0.0	25.9
Alcohol use	18.19	3.0	10.3	30.0
Hazardous waste dump sites	-13.57	1.8	-13.8	0.0
Health insurance coverage	80.36	11.5	60.4	99.2
High blood pressure	31.41	6.9	12.6	55.3
Housing vacancy rate	6.97	4.8	0.0	27.6
Routine check ups	66.31	3.7	56.0	81.3
Sleep	34.52	4.4	23.0	48.6
Stroke	2.87	1.3	0.6	8.3
Traffic accidents drugs	0.41	2.1	0.0	55.6
Traffic accidents pedalcyclist	0.57	1.8	0.0	28.0
Walkability	10.39	3.3	2.8	16.9

Table 5. Descriptive Statistics for Z-Score Standardized Data

Indicators	Mean	SD	Min	Max
Access to healthy food	0.00	1.00	-0.69	7.76
Alcohol use	0.00	1.00	-2.03	5.99
Hazardous waste dump sites	0.00	1.00	-4.99	5.04
Health insurance coverage	0.00	1.00	-1.27	6.18
High blood pressure	0.00	1.00	-2.27	6.81
Housing vacancy rate	0.00	1.00	-1.42	4.36
Routine check ups	0.00	1.00	-1.08	7.48
Sleep	0.00	1.00	-1.85	6.81
Stroke	0.00	1.00	-1.64	5.75
Traffic accidents drugs	0.00	1.00	-0.20	26.63
Traffic accidents pedalcyclist	0.00	1.00	-0.33	15.62
Walkability	0.00	1.00	-2.18	5.74

Table 6. Descriptive Statistics for Scaled Data

Variable Name	Mean	SD	Min	Max
Access to healthy food	91.83	11.8	0.0	100.0
Alcohol use	74.69	12.5	0.0	100.0
Hazardous waste dump sites	50.25	10.0	0.0	100.0
Health insurance coverage	17.07	13.4	0.0	100.0
High blood pressure	74.97	11.0	0.0	100.0
Housing vacancy rate	75.39	17.3	0.0	100.0
Routine check ups	12.63	11.7	0.0	100.0
Sleep	78.63	11.6	0.0	100.0
Stroke	77.82	13.5	0.0	100.0
Traffic accidents drugs	99.26	3.7	0.0	100.0
Traffic accidents pedalcyclist	97.96	6.3	0.0	100.0
Walkability	27.48	12.6	0.0	100.0

g. Calculate cumulative Health Score

Finally, to calculate the cumulative Health Score, we multiplied each of the scaled indicator values with the final weights calculated in **Section 3e** and added them. Then we re-scaled this score to be between 0 to 100 using a min-max scaler. This gave us the final cumulative Health Score for each census tract..

4. Conclusion

Our final results are visualized in the attached Tableau dashboard. Dark blue indicates a low score, high need census tract. Light teal indicates a high score, healthier census tract. The map is consistent with our expectations - Southeast Dallas has a high concentration of higher need census tracts. The same appears true for Southeast Fort Worth as well.

The dashboard is meant to be interactive. When you click a census tract on the map, the bar chart is filtered for that census tract only and you can view its Health Score ranking among all 1172 census tracts. You can also view the census tract's performance in the 12 indicators compared to the average performance across all census tracts.

The final Health Sub-Index has a diverse and informative set of indicators. We have at least two indicators in each of the five categories - healthcare, physical health, mental health, environment, and traffic - which can be used to guide organizations on where to dedicate resources. For example, Census Tract 1132.02 in Arlington, Tarrant County is ranked last among all the counties. Although this census tract has decent health care coverage and performs well in routine checkups, it performs very poorly in stroke, high blood pressure, sleep, and access to healthy food. The idea is not to funnel resources directly into care for people who have experienced strokes or have high blood pressure, but instead to look at the overall message communicated by these results. It's important to note that the indicators included in this model are not an

exhaustive list of factors that contribute to the health of a community. They are *indicators* and are meant to be representative of factors not included in the model. For example, stroke and high blood pressure are highly correlated with an array of complications such as lung disease, heart disease, and diabetes. Viewed in combination with low sleep scores and low access to healthy food, perhaps one answer is to provide health education for students and parents on the importance of diet and exercise and provide accompanying resources to enable healthy lifestyle choices.

This is just one example of how to utilize the Health Sub-Index and dashboard. There are many more insights to be gained for each census tract. We look forward to any improvements future UTD students may contribute to our indicator selection and methodology and hope our work may make an impact in CPAL's mission to reduce child poverty in DFW communities.