CSE6242 Data and Visual Analytics
Project Final Report

Team 85 – Bu Qianqian (qbu7@gatech.edu),
Myles Lefkovitz (mlefkovitz@gatech.edu),
Salim Noorallah Ladak (sladak3@gatech.edu)
Tri Nguyen (tnguyen497@gatech.edu)

Introduction

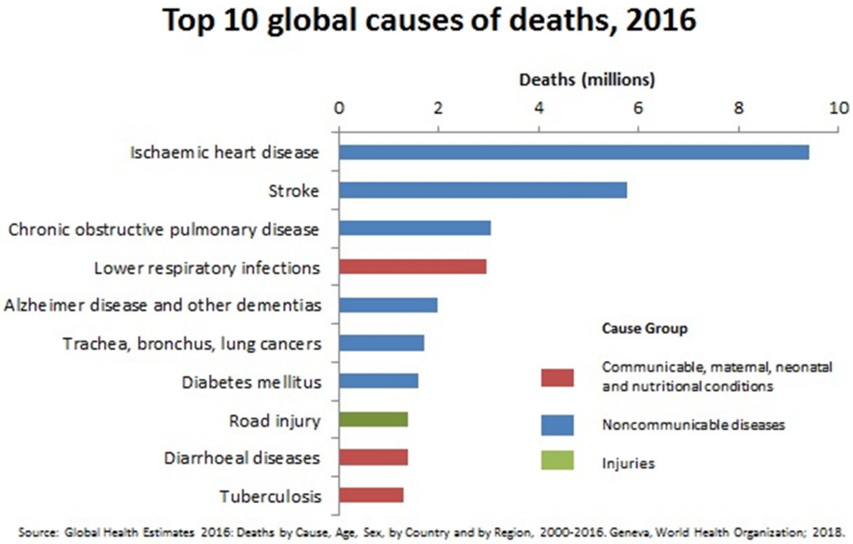
Our objective is to produce an interactive visualization to show the most prevalent diseases in the 500 largest cities in the US, the factors that most contribute to those poor health outcomes and their prevalence within the cities. We base our analysis on data provided by the Centers for Disease Control and Prevention (CDC).

We developed machine learning models to predict the health outcomes and deduce the most important contributing factors. These factors are ranked by the strength of their relationship to each disease. We find that support vector regression (SVR) performs the best in terms of predicting health outcome prevalence; therefore, we use SVR to produce the final ranking for visualization.

Motivation

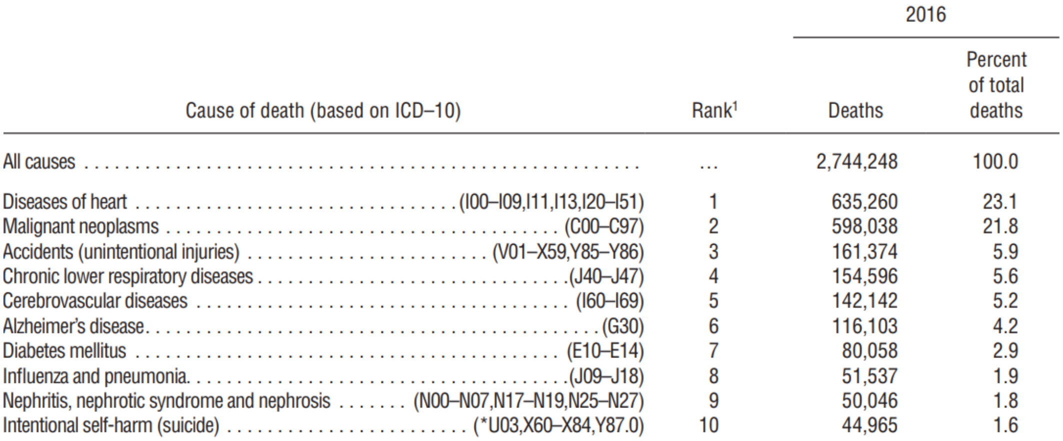
The world is facing increasing demand for health resources, and it requires deeper understanding of causes and comparative burden of diseases and risk factors (Lopez et al., 2006).

Diseases and health conditions represent 9 of the 10 most common causes of death globally (figure 1). Of those 10 causes, chronic diseases represent 6 (WHO, 2018). Chronic disease is also a major cause for loss of life within the US (figure 2) (Heron, 2018).



Source: Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. Geneva, World Health Organization; 2018.

Figure 1. Top 10 global causes of death, 2016 (WHO, 2018)



Cause of death (based on ICD–10)	Rank[1]	2016 Deaths	 Percent of total deaths
All causes	...	2,744,248	100.0
Diseases of heart (I00–I09,I11,I13,I20–I51)	1	635,260	23.1
Malignant neoplasms (C00–C97)	2	598,038	21.8
Accidents (unintentional injuries) (V01–X59,Y85–Y86)	3	161,374	5.9
Chronic lower respiratory diseases (J40–J47)	4	154,596	5.6
Cerebrovascular diseases (I60–I69)	5	142,142	5.2
Alzheimer's disease (G30)	6	116,103	4.2
Diabetes mellitus (E10–E14)	7	80,058	2.9
Influenza and pneumonia (J09–J18)	8	51,537	1.9
Nephritis, nephrotic syndrome and nephrosis (N00–N07,N17–N19,N25–N27)	9	50,046	1.8
Intentional self-harm (suicide) (*U03,X60–X84,Y87.0)	10	44,965	1.6

Figure 2. Top 10 US 2016 causes of death (Heron, 2018)

McKenna and Collins in Remington, 2010 describe the continuum of factors leading to chronic diseases and subsequently disability and fatality (figure 3). Ongoing research on chronic diseases and their causes is therefore of interest to governments (Allen et al., 2014) and the public. The above-mentioned works by Lopez, McKenna, and Allen cover this topic in-depth and provide the motivation for the project, but they lack an assessible medium (e.g. interactive visualization) to help policy makers and the public understand the topic easily.

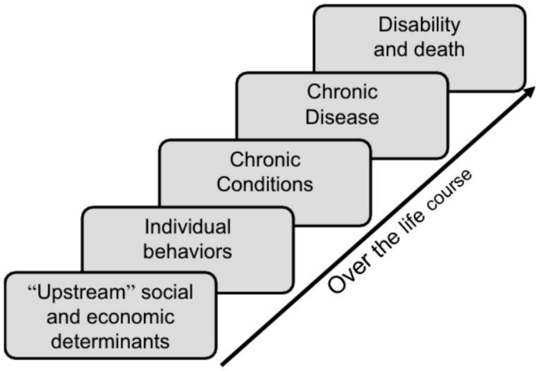
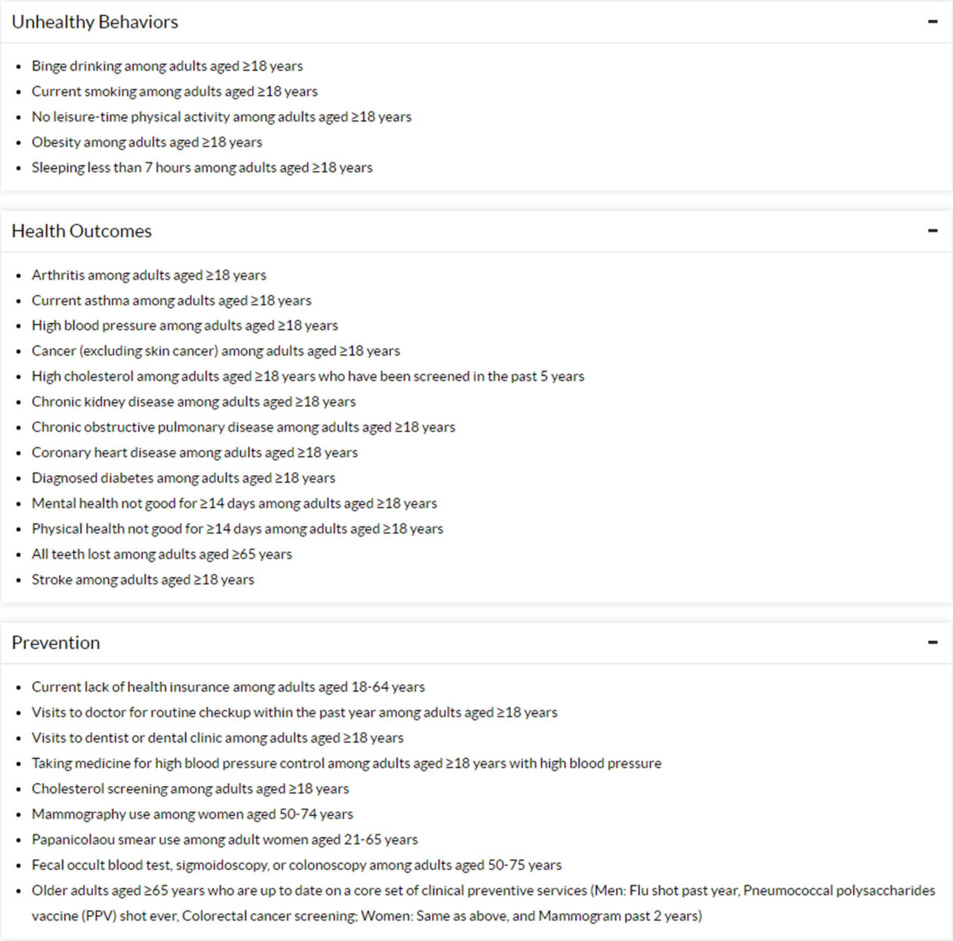


Figure 3. The chronic disease continuum (McKenna and Collins in Remington, 2010)

Problem Definition

Our goal is to identify factors (preventions and unhealthy behaviors) most related to common diseases in 500 largest cities in the US. We create an interactive visualization that enables users to identify the relationships between preventions / unhealthy behaviors and health outcomes.

We will utilize the CDC's 500 Cities project's data, which compiled data for 5 unhealthy behaviors, 9 preventions, and 13 health outcomes for census tracts across the 500 most populous US cities (around 100m people representing 33% of the US population) (500 Cities, CDC).



Unhealthy Behaviors

- Binge drinking among adults aged ≥18 years
- Current smoking among adults aged ≥18 years
- No leisure-time physical activity among adults aged ≥18 years
- Obesity among adults aged ≥18 years
- Sleeping less than 7 hours among adults aged ≥18 years

Health Outcomes

- Arthritis among adults aged ≥18 years
- Current asthma among adults aged ≥18 years
- High blood pressure among adults aged ≥18 years
- Cancer (excluding skin cancer) among adults aged ≥18 years
- High cholesterol among adults aged ≥18 years who have been screened in the past 5 years
- Chronic kidney disease among adults aged ≥18 years
- Chronic obstructive pulmonary disease among adults aged ≥18 years
- Coronary heart disease among adults aged ≥18 years
- Diagnosed diabetes among adults aged ≥18 years
- Mental health not good for ≥14 days among adults aged ≥18 years
- Physical health not good for ≥14 days among adults aged ≥18 years
- All teeth lost among adults aged ≥65 years
- Stroke among adults aged ≥18 years

Prevention

- Current lack of health insurance among adults aged 18-64 years
- Visits to doctor for routine checkup within the past year among adults aged ≥18 years
- Visits to dentist or dental clinic among adults aged ≥18 years
- Taking medicine for high blood pressure control among adults aged ≥18 years with high blood pressure
- Cholesterol screening among adults aged ≥18 years
- Mammography use among women aged 50-74 years
- Papanicolaou smear use among adult women aged 21-65 years
- Fecal occult blood test, sigmoidoscopy, or colonoscopy among adults aged 50-75 years
- Older adults aged ≥65 years who are up to date on a core set of clinical preventive services (Men: Flu shot past year, Pneumococcal polysaccharides vaccine (PPV) shot ever, Colorectal cancer screening; Women: Same as above, and Mammogram past 2 years)

Figure 4. 27 Measures- CDC 500 Cities project (500 Cities, CDC)

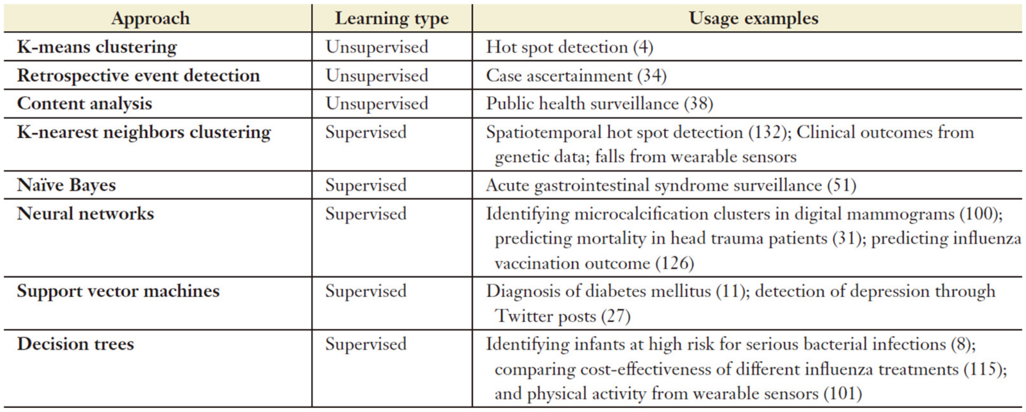
Survey

Study on Relationship between Health Factors

There have been many studies on the relationship between preventive/unhealthy behaviors and chronic diseases. However, behavior and outcome pairs were generally studied in isolation. Obesity, for example, links to coronary heart (Manson et al, 1990) and kidney diseases (Wahba and Mak, 2007). Obesity is further linked to diabetes (Mokdad, 2001) and high blood pressure (Hall, 2014). Omura et al. (2018) studied lack of behavioral counselling for cardiovascular disease. Rippinger et al. (2019) found that abnormal cancer screening findings can discourage follow-up visits. The papers above are useful references because they outline the steps we need to analyze the relationship between specific behaviors and outcomes. These papers, however, only address single pairs, while we will address many.

Health Data Analysis

In healthcare, various machine learning techniques have been used to analyze various problems (Mooney, 2018).



Approach	Learning type	Usage examples
K-means clustering	Unsupervised	Hot spot detection (4)
Retrospective event detection	Unsupervised	Case ascertainment (34)
Content analysis	Unsupervised	Public health surveillance (38)
K-nearest neighbors clustering	Supervised	Spatiotemporal hot spot detection (132); Clinical outcomes from genetic data; falls from wearable sensors
Naïve Bayes	Supervised	Acute gastrointestinal syndrome surveillance (51)
Neural networks	Supervised	Identifying microcalcification clusters in digital mammograms (100); predicting mortality in head trauma patients (31); predicting influenza vaccination outcome (126)
Support vector machines	Supervised	Diagnosis of diabetes mellitus (11); detection of depression through Twitter posts (27)
Decision trees	Supervised	Identifying infants at high risk for serious bacterial infections (8); comparing cost-effectiveness of different influenza treatments (115); and physical activity from wearable sensors (101)

Table 1. Techniques used to solve various healthcare problems (Mooney, 2018)

Multiple techniques can be used effectively for the same problem, and their effectiveness can be measured and ranked. Random forests are the best method for breast cancer screening data, outperforming decision trees and support vector machines in accuracy (Farooqui, 2018). Neural networks were found to be the best method for predicting cardiovascular risk, outperforming random forests, logistic regression, and gradient boosting (Weng, 2017). The three papers above provide us with various examples of techniques that we can apply to our analysis. There are various techniques that were not discussed (a shortcoming) that we will also try to apply.

Current practice frequently compares 3 or 4 algorithms on a single question. Sometimes larger scale analyses are performed with multiple outputs computed simultaneously. Luo, et al (2016) computed 6 health outcomes (5 of which are health outcomes measured by the CDC 500 Cities project) using socio-demographic data.

This paper provides us with a scalable example for computing multiple sets of outputs. The scope is limited in number of outputs and scale of analysis, which we will improve upon by analyzing our data across varying levels of granularity (states and cities).

Health Data Visualization

Data visualization is often used for health data. Tatyana et al. (2017) discussed static mapping techniques for population diseases and provided useful examples, but do not cover animated visualization. Sopan et al. (2012) developed a GIS visualization tool for health data which lacks explicit analysis, but its architecture is a useful reference.

MacQuillan et al. (2017) outline how interpretation from health data varies depending on how data is aggregated geographically (using birth outcome among black women data). Challenge in data-driven healthcare research and visualization include difficulty in combining data sources, bias, incorrect interpretation, etc. (Gotz and Borland 2016). These are valuable insights for our project.

Proposed Method

Intuition

Our project will contribute to the state-of-the-art research in public health in the following areas:

1. While current research focuses on a single factor-disease pair, we will study the relationship between 13 diseases and 14 factors (9 preventions and 5 unhealthy behaviors).
2. Utilizing a large public health data set recently made available by the CDC at census tract level, we can study the relevance of diseases at high geographic granularity.
3. By developing a model that takes several factors contributing to 13 diseases into account we can understand the relative importance of each factor has in predicting each disease.

4. By representing the outcomes of this analysis visually via maps and charts we can glean insights easily.

Approach

Analysis

For each of the 13 health outcomes, we identified the most highly related factors (preventions / unhealthy behaviors) using the following steps:

1. Treat each census tract as one data point, for all census tracts in the country, randomly split them into training and testing subsets.
2. Build various multiple regression models to predict health outcomes (13), using preventions (9) and unhealthy behaviors (5) as features (for a total of 14 features).
3. Refine the model by feature selection / dimensionality reduction and tuning parameters to improve performance on high-dimensional datasets as well as achieve higher accuracy scores for the estimators.
4. From the best model developed through this process, identify the top features that most contribute to the prediction.

For each state, we repeated steps 1-4 using only the census tracts for that state.

Experiment Algorithms

Various machine learning / regression algorithms were initially run for each of the 13 health outcomes. Algorithms encompass various types of multiple regression, including:

- Linear regression,
- Ridge regression,
- Lasso regression, and
- Support vector regression (SVR).

Evaluation of the algorithms is based on the highest *test* score. The test scores for each multiple regression model are based on the R-squared value of test data sets. The top features are selected and ranked per health outcome based on the coefficients of the models or by recursive feature elimination.

For SVR, the GridSearchCV method provided in the Scikit-learn library was used for an exhaustive search to tune the hyper-parameters of the estimator. Hyper-parameters are parameters not directly learned within estimators. Two parameters are specified in the grid search: the kernel type as 'linear' or 'rbf' and the penalty parameter C of the error term with the following values 0.01, 1 and 10.

The kernel functions are used to transform the original dataset into a high-dimensional feature that has a clear dividing margin. Linear kernel is a parametric model while RBF is not. RBF kernel stands for radial basis function kernel which creates a non-linear combination of the variables to find a linear boundary in the higher dimensional space which is an important parameter in support vector machine model.

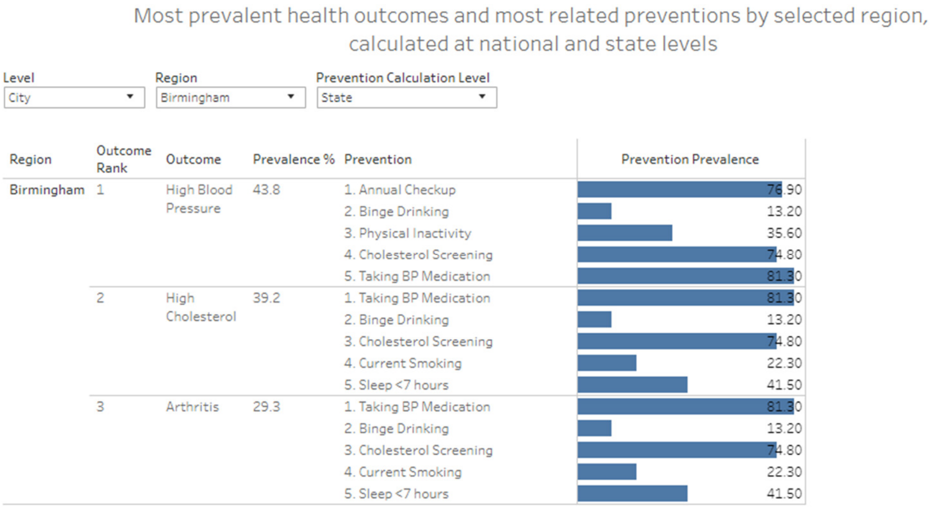
For more details on analysis, refer to the *Experiments/Evaluation* section.

Visualization

The visualization is developed in Tableau.

1. **Health Outcomes and Related Factors:** This is the main interface showing our analysis result. User can select the nation, a state, or a city and see the top diseases in that area.

   For each disease, a bar chart shows the five most strongly related factors (calculated at the state or national level where available) as identified in the analysis and identifies the prevalence of each factor (prevention/unhealthy behavior) and the outcome (disease).



Most prevalent health outcomes and most related preventions by selected region, calculated at national and state levels

Level: City

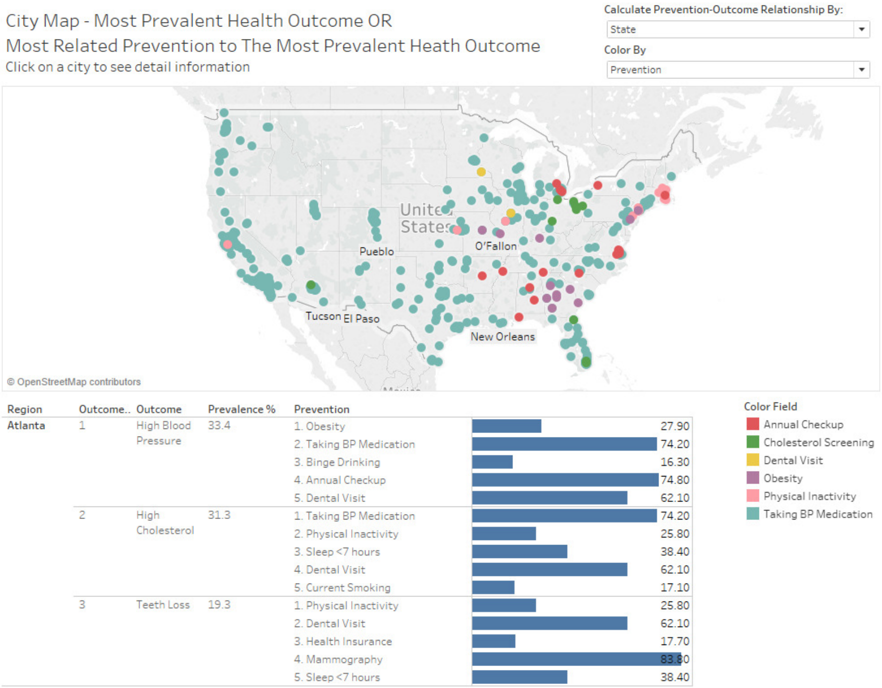
Region: Birmingham

Prevention Calculation Level: State

Region	Outcome Rank	Outcome	Prevalence %	Prevention	Prevention Prevalence
Birmingham	1	High Blood Pressure	43.8	1. Annual Checkup	76.90
				2. Binge Drinking	13.20
				3. Physical Inactivity	35.60
				4. Cholesterol Screening	74.80
				5. Taking BP Medication	81.30
	2	High Cholesterol	39.2	1. Taking BP Medication	81.30
				2. Binge Drinking	13.20
				3. Cholesterol Screening	74.80
				4. Current Smoking	22.30
				5. Sleep <7 hours	41.50
	3	Arthritis	29.3	1. Taking BP Medication	81.30
				2. Binge Drinking	13.20
				3. Cholesterol Screening	74.80
				4. Current Smoking	22.30
				5. Sleep <7 hours	41.50

Figure 5. Health outcomes and related factors bar charts

2. **Interactive Health Information Map:**
    - A map at the city level displaying 500 circles across the US (one for each city in the data set).
    - For each city the circle is colored categorically in line with the most prevalent outcome or the prevention most closely related to the most prevalent outcome.
    - When hovering over a circle relevant prevalence details are displayed.
    - When clicking on a circle the bar charts from the figure above are displayed for that city.



City Map - Most Prevalent Health Outcome OR
Most Related Prevention to The Most Prevalent Heath Outcome
Click on a city to see detail information

Calculate Prevention-Outcome Relationship By: State

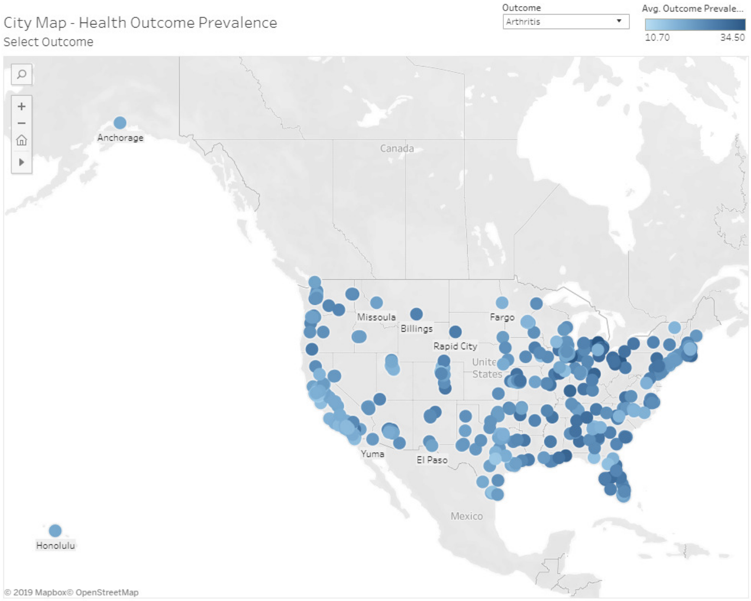
Color By: Prevention

Region	Outcome..	Outcome	Prevalence %	Prevention	
Atlanta	1	High Blood Pressure	33.4	1. Obesity	27.90
				2. Taking BP Medication	74.20
				3. Binge Drinking	16.30
				4. Annual Checkup	74.80
				5. Dental Visit	62.10
	2	High Cholesterol	31.3	1. Taking BP Medication	74.20
				2. Physical Inactivity	25.80
				3. Sleep <7 hours	38.40
				4. Dental Visit	62.10
				5. Current Smoking	17.10
	3	Teeth Loss	19.3	1. Physical Inactivity	25.80
				2. Dental Visit	62.10
				3. Health Insurance	17.70
				4. Mammography	83.80
				5. Sleep <7 hours	38.40

Color Field
■ Annual Checkup
■ Cholesterol Screening
■ Dental Visit
■ Obesity
■ Physical Inactivity
■ Taking BP Medication

Figure 6. Interactive health information

3. **Outcome Prevalence Map:** A map of 500 cities comparing the prevalence of the selected outcome.



City Map - Health Outcome Prevalence
Select Outcome

Outcome: Arthritis

Avg. Outcome Prevale...
10.70 34.50

Figure 7. Outcome Prevalence

4. **Top Prevention by Selected Outcome:** A map that shows the most relevant prevention for the selected outcome for each city.

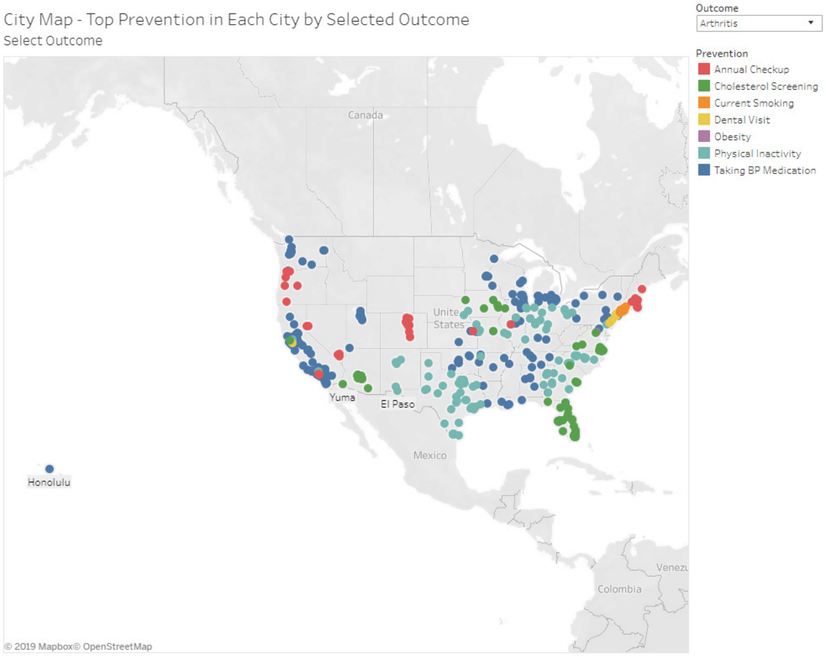


Figure 8. Top Prevention by Selected Outcome

Experiments/Evaluation

Testbed

Our analysis identifies the relationships that exist between (1) preventions and unhealthy behaviors and (2) health outcomes. For each of the 13 health outcomes identified in the CDC 500 Cities dataset, we ran an experiment to identify the most relevant factors (preventions / unhealthy behaviors) at national and state levels.

Data is provided at the census tract levels for 13 health outcomes, 9 preventions and 5 unhealthy behaviors. Data is analyzed both nationally and at state level using census tract data. The data is first split 70% for training and 30% for testing. The selected tool for machine learning / regression is the Python library *Scikit-Learn*.

Experiment Questions:

- For each of the 13 health outcomes, is there an interpretable relationship with the 9 preventions and 5 unhealthy behaviors?
- Is the relationship consistent at national level and among the 50 states?
- What model can be used to best describe the relationships?
- What is the prediction accuracy of the tested models?
- What are the top 5 preventions or unhealthy behaviors for predicting each health outcome at national and state levels?

Experiments

The Python code consists of four .py files: *main.py, util.py, get_data.py* and *constants.py*, for which descriptions follow.

The constants file, *constants.py*, stores column names and other constants predefined and to be used across the project.

The data module, *get_data.py* preforms the initial data cleaning steps. It reads data from the file raw data file, removes additional data rows and columns, then reformats the data. Each row of the resulting data frame represents one data point to be used in the following experiment.

The utilities module, *util.py*, contains the library functions to support the key analysis functionalities including model fitting, hyperparameter optimization, evaluation, and top features ranking.

Finally, *main.py* is the main starting point for this program. It provides an interactive way for users to run it with below options:

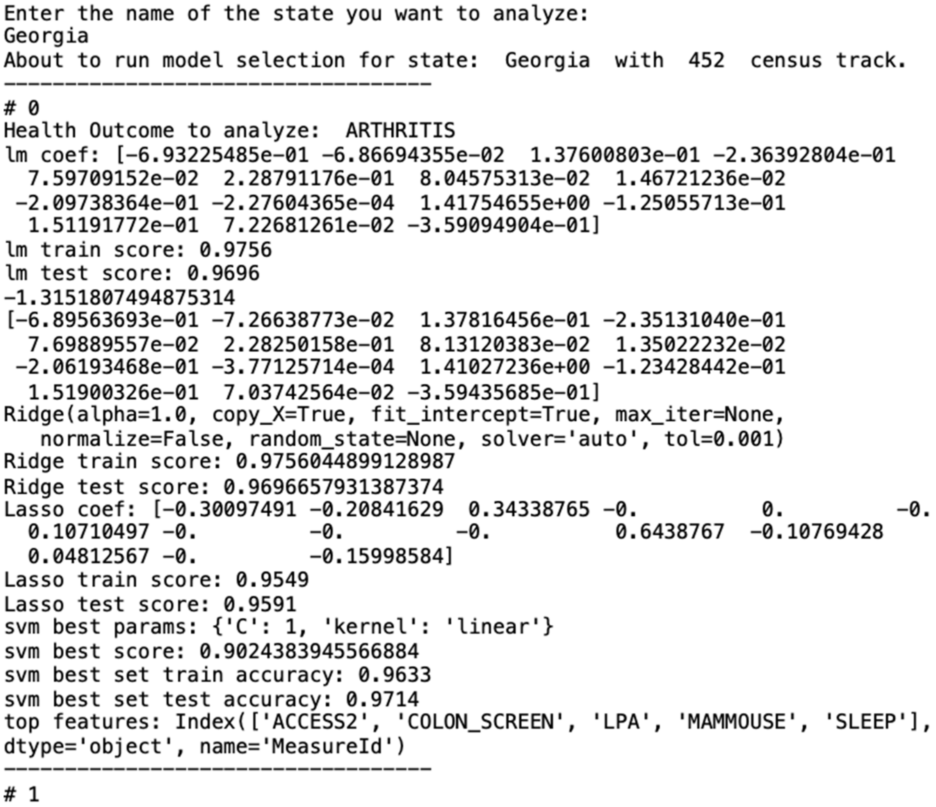
1. Perform top feature selection Analysis.
   Top features are selected and ranked with recursive feature elimination through the feature importance attribute of the best fitted SVR model for each of the 13 health outcomes both at the national and state levels. It also reports the training and test scores of the fitted model.

2. Generate the output file for visualization.
   The output file includes the geographic information, prevalence, model test accuracy for the 5 most significant unhealthy behavior/prevention and health outcome pairs for all the states and cities in US.
3. Run model comparison analysis.
   Runs a comparative analysis using multiple regression models either for a particular state or at national level.

Observations

A comparison of the four different regression algorithms – Linear, Ridge, Lasso Regression and Support Vector Regression (SVR) - was carried out both for the nation and for each of the states across the 13 health outcomes. The train and test scores suggest that generally SVR produce the highest prediction accuracy among the selected models for different health outcomes. Note that for some of the outcomes, the highest test scores achieved by RBF kernel, however, because of the complexity of RBF kernel, it's more expensive to train an RBF kernel SVR and it cannot get the most important original factors after the nonlinear transformation. Thus, taking into consideration the trade-off between model prediction accuracy versus model interpretation, the linear kernel SVR is chosen as the best model to perform further analysis.

Figure 9 shows a sample comparative analysis (option 3 above) for the selected state of Georgia (showing the first health outcome, Arthritis). In this case, SVR test score performs slightly better (0.9714) than lasso, ridge and linear regression.

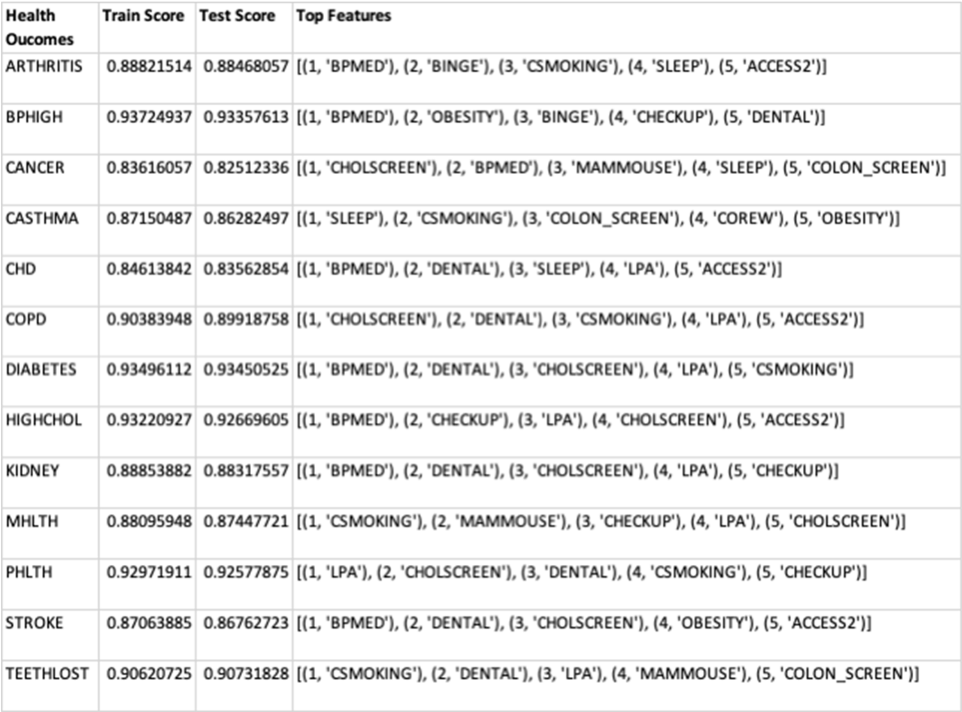


```
Enter the name of the state you want to analyze:
Georgia
About to run model selection for state:  Georgia  with  452  census track.
---------------------------------
# 0
Health Outcome to analyze:  ARTHRITIS
lm coef: [-6.93225485e-01 -6.86694355e-02  1.37600803e-01 -2.36392804e-01
  7.59709152e-02  2.28791176e-01  8.04575313e-02  1.46721236e-02
 -2.09738364e-01 -2.27604365e-04  1.41754655e+00 -1.25055713e-01
  1.51191772e-01  7.22681261e-02 -3.59094904e-01]
lm train score: 0.9756
lm test score: 0.9696
-1.3151807494875314
[-6.89563693e-01 -7.26638773e-02  1.37816456e-01 -2.35131040e-01
  7.69889557e-02  2.28250158e-01  8.13120383e-02  1.35022232e-02
 -2.06193468e-01 -3.77125714e-04  1.41027236e+00 -1.23428442e-01
  1.51900326e-01  7.03742564e-02 -3.59435685e-01]
Ridge(alpha=1.0, copy_X=True, fit_intercept=True, max_iter=None,
   normalize=False, random_state=None, solver='auto', tol=0.001)
Ridge train score: 0.9756044899128987
Ridge test score: 0.9696657931387374
Lasso coef: [-0.30097491 -0.20841629  0.34338765 -0.          0.         -0.
  0.10710497 -0.         -0.         -0.          0.6438767  -0.10769428
  0.04812567 -0.         -0.15998584]
Lasso train score: 0.9549
Lasso test score: 0.9591
svm best params: {'C': 1, 'kernel': 'linear'}
svm best score: 0.9024383945566884
svm best set train accuracy: 0.9633
svm best set test accuracy: 0.9714
top features: Index(['ACCESS2', 'COLON_SCREEN', 'LPA', 'MAMMOUSE', 'SLEEP'],
dtype='object', name='MeasureId')
---------------------------------
# 1
```

Figure 9. Comparative Regression Results for the State of Georgia (Arthritis only)

With the finding that the best-performing algorithm is SVR, we embarked on using it with our remaining experiments for top-feature analysis reporting. For mode 1, Top Feature Selection Analysis, the user is prompted for the number of top features, after which SVR is automatically run at both national level and for all states. After two hours of processing, it creates two output files containing national and state analyses, with training scores, test scores, and the top features.

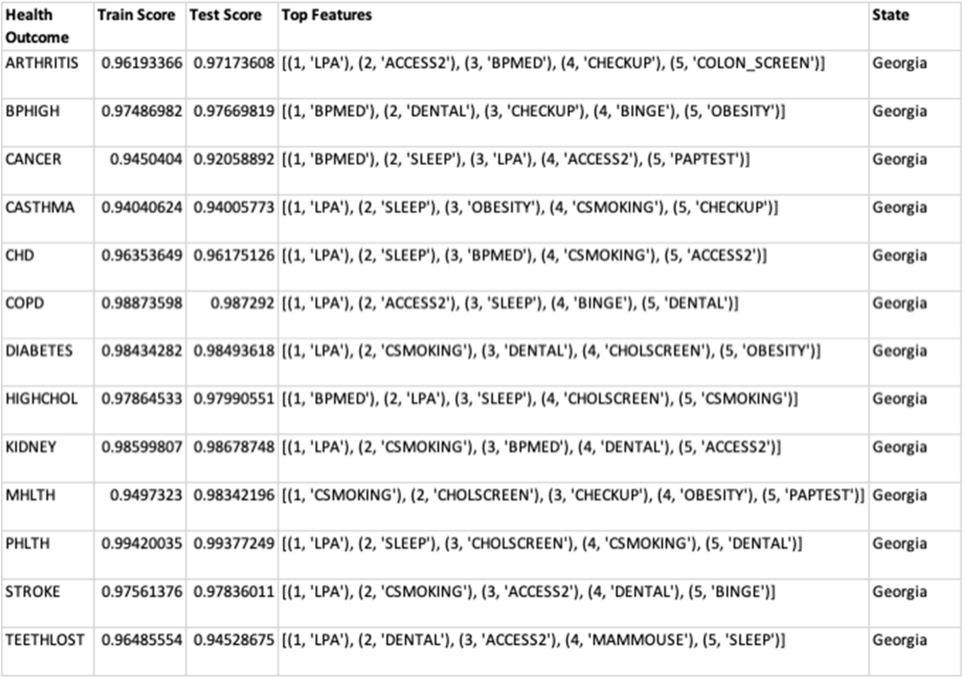
Figure 10 shows the output for country level with the top five features. Depending on the health outcome, test scores range between 0.8251 and 0.9345 with a mean of 0.8893, reaffirming that the regression model provides a good fit and that the data has a linear relationship. The top preventions / unhealthy behaviors are then identified and ranked for each health outcome in order to establish links.



Health Oucomes	Train Score	Test Score	Top Features
ARTHRITIS	0.88821514	0.88468057	[(1, 'BPMED'), (2, 'BINGE'), (3, 'CSMOKING'), (4, 'SLEEP'), (5, 'ACCESS2')]
BPHIGH	0.93724937	0.93357613	[(1, 'BPMED'), (2, 'OBESITY'), (3, 'BINGE'), (4, 'CHECKUP'), (5, 'DENTAL')]
CANCER	0.83616057	0.82512336	[(1, 'CHOLSCREEN'), (2, 'BPMED'), (3, 'MAMMOUSE'), (4, 'SLEEP'), (5, 'COLON_SCREEN')]
CASTHMA	0.87150487	0.86282497	[(1, 'SLEEP'), (2, 'CSMOKING'), (3, 'COLON_SCREEN'), (4, 'COREW'), (5, 'OBESITY')]
CHD	0.84613842	0.83562854	[(1, 'BPMED'), (2, 'DENTAL'), (3, 'SLEEP'), (4, 'LPA'), (5, 'ACCESS2')]
COPD	0.90383948	0.89918758	[(1, 'CHOLSCREEN'), (2, 'DENTAL'), (3, 'CSMOKING'), (4, 'LPA'), (5, 'ACCESS2')]
DIABETES	0.93496112	0.93450525	[(1, 'BPMED'), (2, 'DENTAL'), (3, 'CHOLSCREEN'), (4, 'LPA'), (5, 'CSMOKING')]
HIGHCHOL	0.93220927	0.92669605	[(1, 'BPMED'), (2, 'CHECKUP'), (3, 'LPA'), (4, 'CHOLSCREEN'), (5, 'ACCESS2')]
KIDNEY	0.88853882	0.88317557	[(1, 'BPMED'), (2, 'DENTAL'), (3, 'CHOLSCREEN'), (4, 'LPA'), (5, 'CHECKUP')]
MHLTH	0.88095948	0.87447721	[(1, 'CSMOKING'), (2, 'MAMMOUSE'), (3, 'CHECKUP'), (4, 'LPA'), (5, 'CHOLSCREEN')]
PHLTH	0.92971911	0.92577875	[(1, 'LPA'), (2, 'CHOLSCREEN'), (3, 'DENTAL'), (4, 'CSMOKING'), (5, 'CHECKUP')]
STROKE	0.87063885	0.86762723	[(1, 'BPMED'), (2, 'DENTAL'), (3, 'CHOLSCREEN'), (4, 'OBESITY'), (5, 'ACCESS2')]
TEETHLOST	0.90620725	0.90731828	[(1, 'CSMOKING'), (2, 'DENTAL'), (3, 'LPA'), (4, 'MAMMOUSE'), (5, 'COLON_SCREEN')]

Figure 10. Support Vector Regression Results with Top Features at National Level

With the five top features and the state name added in the last column, Figure 11 shows an extract for the state of Georgia, which has a mean test score of 0.9700.



Health Outcome	Train Score	Test Score	Top Features	State
ARTHRITIS	0.96193366	0.97173608	[(1, 'LPA'), (2, 'ACCESS2'), (3, 'BPMED'), (4, 'CHECKUP'), (5, 'COLON_SCREEN')]	Georgia
BPHIGH	0.97486982	0.97669819	[(1, 'BPMED'), (2, 'DENTAL'), (3, 'CHECKUP'), (4, 'BINGE'), (5, 'OBESITY')]	Georgia
CANCER	0.9450404	0.92058892	[(1, 'BPMED'), (2, 'SLEEP'), (3, 'LPA'), (4, 'ACCESS2'), (5, 'PAPTEST')]	Georgia
CASTHMA	0.94040624	0.94005773	[(1, 'LPA'), (2, 'SLEEP'), (3, 'OBESITY'), (4, 'CSMOKING'), (5, 'CHECKUP')]	Georgia
CHD	0.96353649	0.96175126	[(1, 'LPA'), (2, 'SLEEP'), (3, 'BPMED'), (4, 'CSMOKING'), (5, 'ACCESS2')]	Georgia
COPD	0.98873598	0.987292	[(1, 'LPA'), (2, 'ACCESS2'), (3, 'SLEEP'), (4, 'BINGE'), (5, 'DENTAL')]	Georgia
DIABETES	0.98434282	0.98493618	[(1, 'LPA'), (2, 'CSMOKING'), (3, 'DENTAL'), (4, 'CHOLSCREEN'), (5, 'OBESITY')]	Georgia
HIGHCHOL	0.97864533	0.97990551	[(1, 'BPMED'), (2, 'LPA'), (3, 'SLEEP'), (4, 'CHOLSCREEN'), (5, 'CSMOKING')]	Georgia
KIDNEY	0.98599807	0.98678748	[(1, 'LPA'), (2, 'CSMOKING'), (3, 'BPMED'), (4, 'DENTAL'), (5, 'ACCESS2')]	Georgia
MHLTH	0.9497323	0.98342196	[(1, 'CSMOKING'), (2, 'CHOLSCREEN'), (3, 'CHECKUP'), (4, 'OBESITY'), (5, 'PAPTEST')]	Georgia
PHLTH	0.99420035	0.99377249	[(1, 'LPA'), (2, 'SLEEP'), (3, 'CHOLSCREEN'), (4, 'CSMOKING'), (5, 'DENTAL')]	Georgia
STROKE	0.97561376	0.97836011	[(1, 'LPA'), (2, 'CSMOKING'), (3, 'ACCESS2'), (4, 'DENTAL'), (5, 'BINGE')]	Georgia
TEETHLOST	0.96485554	0.94528675	[(1, 'LPA'), (2, 'DENTAL'), (3, 'ACCESS2'), (4, 'MAMMOUSE'), (5, 'SLEEP')]	Georgia

Figure 11. Support Vector Regression Results with Top Features for the State of Georgia

Conclusions and Discussion

Our analysis shows that for each one of the 13 health outcomes, there is a linear relationship with at least some of the preventions and unhealthy behaviors – we have chosen to rank the top five. This is true both at the national level and across the 50 states. The support vector regression (SVR) algorithm provides the highest test accuracies (averaging 0.8893 nationally) across health outcomes and states.

The visualization allows users to explore the analysis results for their geographic areas of interest. Users can see the most widespread diseases in an area, what unhealthy behaviors and preventions are most related to those diseases. They can also utilize the prevalence map to quickly compare prevalence of those diseases and factors across all cities.

This can be used to assist governmental bodies to assess health outcomes in a city or a state, and more effectively direct their investment in preventative measures. From the map displayed in Figure 6 we can see that “Taking BP Medication” is the prevention most closely related to the most prevalent outcome in a large majority of the cities in the US. The most prevalent outcomes are high cholesterol and high blood pressure. Organizations could use this information to target their efforts on increasing the rate of compliance with blood pressure prescriptions (making sure that physicians are prescribing them and making sure that patients are taking them).

Team Work Distribution

All team members contributed a similar amount of effort.

Project Research and Ideation: Handled by all team members evenly

Project management: Handled by all team members evenly

Analysis: Led by Jane and Salim

Visualization Creation: Led by Tri and Myles

Results Interpretation: Handled by all team members evenly

References

1. WHO (2018, May 24). The top 10 causes of death. *World Health Organization*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death
2. 500 Cities | About our project. *Centers for Disease Control and Prevention*. Retrieved from https://www.cdc.gov/500cities/about.htm
3. Allen, P., Sequeira, S., Best, L., Jones, E., Baker, E. A., & Brownson, R. C. (2014). Peer Reviewed: Perceived Benefits and Challenges of Coordinated Approaches to Chronic Disease Prevention in State Health Departments. *Preventing chronic disease, 11*.
4. Farooqui, N. A. (2018). A Study of Early Prevention and Detection of Breast Cancer Using Three Machine Learning Techniques. *International Journal of Advanced Research in Computer Science, 9* (Special Issue 2), 37.
5. Gotz, D., & Borland, D. (2016). Data-driven healthcare: Challenges and opportunities for interactive visualization. *IEEE computer graphics and applications, 36*(3), 90-96.
6. Hall, M. E., do Carmo, J. M., da Silva, A. A., Juncos, L. A., Wang, Z., & Hall, J. E. (2014). Obesity, hypertension, and chronic kidney disease. *International journal of nephrology and renovascular disease*, 7, 75.
7. Heron, M. P. (2018). Deaths: Leading causes for 2016.
8. Mooney, S. J., & Pejaver, V. (2018). Big data in public health: terminology, machine learning, and privacy. *Annual review of public health*, 39, 95-112.
9. Lopez, A. D., Mathers, C. D., Ezzati, M., Jamison, D. T., & Murray, C. J. (Eds.). (2006). *Global burden of disease and risk factors*. The World Bank.
10. Luo, W., Nguyen, T., Nichols, M., Tran, T., Rana, S., Gupta, S., ... & Allender, S. (2015). Is demography destiny? Application of machine learning techniques to accurately predict population health outcomes from a minimal demographic dataset. *PloS one, 10*(5), e0125602.
11. MacQuillan, E. L., Curtis, A. B., Baker, K. M., Paul, R., & Back, Y. O. (2017). Using GIS mapping to target public health interventions: examining birth outcomes across GIS techniques. *Journal of community health, 42*(4), 633-638.

12. Manson, J. E., Colditz, G. A., Stampfer, M. J., Willett, W. C., Rosner, B., Monson, R. R., ... & Hennekens, C. H. (1990). A prospective study of obesity and risk of coronary heart disease in women. *New England journal of medicine, 322*(13), 882-889.
13. Mokdad, A. H., Ford, E. S., Bowman, B. A., Dietz, W. H., Vinicor, F., Bales, V. S., & Marks, J. S. (2003). Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. *Jama, 289*(1), 76-79.
14. Omura, J. D., Bellissimo, M. P., Watson, K. B., Loustalot, F., Fulton, J. E., & Carlson, S. A. (2018). Primary care providers' physical activity counseling and referral practices and barriers for cardiovascular disease prevention. *Preventive medicine, 108*, 115-122.
15. Remington, P. L., Brownson, R. C., & Wegner, M. V. (2010). *Chronic disease epidemiology and control* (No. Ed. 3). American public health association.
16. Rippinger, N., Heinzler, J., Bruckner, T., Brucker, J., Dinkic, C., Hoffmann, J., ... & Schott, T. C. (2019). The impact of a cervical dysplasia diagnosis on individual cancer prevention habits over time: a bicentric case–control study. *Archives of gynecology and obstetrics*, 1-9.
17. Sopan, A., Noh, A. S. I., Karol, S., Rosenfeld, P., Lee, G., & Shneiderman, B. (2012). Community Health Map: A geospatial and multivariate data visualization tool for public health datasets. *Government Information Quarterly, 29*(2), 223-234.
18. Tatyana V. Kotova, Svetlana M. Malkhazova, Vladimir S. Tikunov, Temenoujka Bandrova (2017). Visualization of Public health dinamics. *Geography, Environment, Sustainability,* Vol.10, No 4, p. 27-42.
19. Wahba, I. M., & Mak, R. H. (2007). Obesity and obesity-initiated metabolic syndrome: mechanistic links to chronic kidney disease. *Clinical Journal of the American Society of Nephrology, 2*(3), 550-562.
20. Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data?. *PloS one, 12*(4), e0174944.