Comparing the Impact of Unhealthy Behaviors and Preventive Services on Chronic Health Outcomes

[1]Swapna S. Gokhale
Dept. of Computer Science & Engineering
Univ. of Connecticut, Storrs, CT 06269
swapna.gokhale@uconn.edu

***Abstract*—Chronic health outcomes are a leading cause of death and disability, and also prominent drivers of health care costs. Most chronic health outcomes can be attributed to a few risky behaviors. It is believed that chronic health outcomes and their burdens can be alleviated by the use of clinical preventive services. This paper seeks to assess the relative influence of unhealthy behaviors and preventive services on chronic health outcomes using the 500 Cities data. The approach comprises of three-way clustering of the 500 cities, one each based on health outcomes, preventive services, and unhealthy behaviors, and then measuring the pairwise similarity between the clustering solutions based on health outcomes and preventive services, and health outcomes and unhealthy behaviors. A variant of the Rand Index is defined to assess this clustering similarity. Higher similarity between clusterings based on health outcomes and unhealthy behaviors compared to the clusterings based on health outcomes and preventive services is observed. These findings suggest a greater influence of unhealthy behaviors over preventive services on chronic health outcomes. The paper concludes with a question of whether investing in facilitating healthier lifestyle choices will yield a higher return towards promoting health as opposed to investing in improving access to preventive services.**

***Index Terms*—500 Cities, Health Outcomes, Unhealthy Behaviors, Clustering, Rand Index.**

I. Introduction & Motivation

Chronic diseases are broadly defined as conditions that last longer than a year or more and require ongoing medical attention or limit activities of daily living or both. Chronic conditions such as heart disease, cancer, and diabetes affect the quality of lives of the affected individuals and their families. Moreover, they are the leading causes of death and disability in the United States [7]. They are also the primary drivers of our nation's $3.3 trillion in annual health care costs [7].

Many chronic diseases may be attributed to a short list of risky behaviors: (i) tobacco use and exposure to secondhand smoke; (ii) poor nutrition, including diets low in fruits and vegetables and high in sodium and saturated fats; (iii) lack of physical activity; and (iv) excessive alcohol use [7]. Many forms of clinical preventive strategies are believed to reduce both the incidence and burden of chronic diseases [14]. Routine medical checkups may include a discussion of health-related risk behaviors, and vaccines to prevent infections and their associated complications. Cancer screening may allow for early detection, while cholesterol and blood pressure screening may keep it under control through medications and reduce the risk for cardiovascular disease and stroke. Overall, the goal of clinical preventive services is to prevent and detect conditions in their earlier, more treatable stages [11].

Given the emphasis on clinical preventive services, one might be (mis)led into believing that somehow these services can compensate for risky and unhealthy behaviors. For example, in the recent years, the Affordable Care Act (ACA) has led to a higher usage of preventive care services, but it has also led to an increase in some risk behaviors including smoking, gaining weight and lack of exercise [3]. The question then arises is what matters more: risky behaviors or the use of preventive services, and whether the detrimental influence of unhealthy behaviors can be mitigated by the use of preventive services. This paper compares the relative impact of unhealthy behaviors and the use of clinical preventive services on chronic health outcomes using data from the 500 cities project. This project provides high-quality, small-area (city and census-tract level) epidemiological estimates for 5 chronic disease risk factors, 13 health outcomes, and 11 preventive services for the largest 500 cities in the United States [5].

The analysis approach consists of two steps. The first step clusters the 500 cities based on each of the three types of measures. The second step computes the similarity score between the clustering solutions based on: (i) chronic health outcomes and unhealthy behaviors; and (ii) chronic health outcomes and the use of clinical preventive services. The underlying hypothesis is that a higher similarity score is indicative of greater influence. Our results show that the similarity score between health outcomes and unhealthy behaviors is considerably higher than the similarity score between health outcomes and the use of preventive services. This suggests that unhealthy behaviors have an adverse impact on chronic health outcomes, which probably cannot be eliminated by the use of clinical preventive services.

The rest of the paper is organized as follows: Section II provides an overview of the 500 Cities data. Section III presents the analysis approach and discusses the results. Section IV compares and contrasts related work. Section V offers concluding remarks and directions for future research.

II. The 500 Cities Data

The 500 Cities Project is a collaboration between the Center for Disease Control (CDC), the Robert Wood Johnson Foun-

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

dation, and the CDC Foundation. It consists of 27 measures that include 5 unhealthy behaviors, 13 health outcomes, and 9 prevention practices [6]. The measures include major risk behaviors that lead to illness, suffering and early death related to chronic diseases and conditions, as well as the conditions and diseases that are the most common, costly, and preventable of all health problems.

The mean prevalence of the health outcomes is summarized in Table I. The estimates of health outcomes represent the percentage of people that have been told that they have the condition by a health care professional. All health outcomes, except for teeth loss have been measured for adults over 18 years of age. Teeth loss is estimated in adults over 65 years. Mental (physical) health refers to the percentage of respondents who felt that their mental (physical) health was not good for more than 14 days. COPD refers to Chronic Obstructive Pulmonary Disease. Table II summarizes the prevalence of five unhealthy habits. Each of these five risk behaviors contributes to multiple health outcomes. In this table, smoking includes those who smoke greater than 100 cigarettes per day or smoke every day. Binge drinking includes men who consume more than 5 drinks per day, and women who consume more than four drinks per day. Lack of physical activity and obesity are self explanatory, whereas lack of sleep estimates the percentage of adults who get insufficient sleep ($<$ 7 hours) every night. Finally, the mean prevalence of the use of clinical preventive services is listed in Table III. The last two rows estimate the percentage of men and women who are up to date on their routine checkups and immunizations. Pap smear, colonoscopy, mammograph and cholesterol screening are for early detection of chronic health outcomes, and BP meds are for controlling blood pressure which poses a risk for heart disease and stroke.

Health Outcome	**Mean**
Arthritis	22.39
Asthma	9.18
High BP	30.39
Cancer	5.98
High Cholesterol	31.35
Kidney Disease	2.75
COPD	6.05
Heart Disease	5.73
Diabetes	10.25
Mental Health	12.44
Physical Health	12.57
Stroke	3.05
Teeth Loss	14.51

Table I: Health Outcomes: Prevalence

III. Analysis Approach

Our analysis approach consists of clustering the 500 cities in three ways; one each based on each of the three types of measures, namely, health outcomes, preventive services and unhealthy behaviors. Intuitively, we seek to determine whether cities that have better health outcomes, lower levels of unhealthy behaviors, and higher degrees of the use of preventive health services belong to the same cluster. Therefore, in three-way clustering, we measure the pairwise similarity of the clustering solutions based on health outcomes and preventive measures, and health outcomes and unhealthy behaviors. Our hypothesis is that if the similarity index between the clustering based on health outcomes and preventive measures is higher than the similarity index between the clustering based on health outcomes and unhealthy behaviors, then preventive measures have a stronger influence on health outcomes or vice versa. Our analysis approach thus consists of two-steps, namely, clustering analysis and similarity assessment.

Unhealthy Behavior	**Mean**
Smoking	17.58
Binge Drinking	16.63
Lack of Physical Activity	25.86
Obesity	29.31
Lack of sleep	35.69

Table II: Unhealthy Behaviors: Prevalence

Preventive Services	**Mean**
Lack of health insurance	16.47
Routine checkup	68.07
Dental visit	62.03
BP Meds	58.11
Cholesterol	74.27
Mammography	77.80.
Pap Smear	80.64
Colonoscopy cancer	62.63
Clinical (Men)	31.85
Clinical (Women)	30.86

Table III: Preventive Services: Prevalence

A. *Clustering Analysis*

The purpose of clustering is to find subgroups of similar observations within a data set. We used both k-means and k-medoids clustering [10], implemented in the following steps:

Scaling the Features/Variables:
We normalize the features because their values in each group are radically different from one another. For example, among health outcomes the prevalence of high cholesterol is in the range of 30-40, and that of stroke lies between 2.0 and 3.0. Among unhealthy behaviors, lack of sleep is almost twice as likely as smoking. Amongst the use of preventive services, the mean prevalence of the lack of health insurance is one-fifth of that of the use of mammography services.

Assessing Clustering Tendency:
We statistically examine the tendency of the 500 cities

to cluster, for each type of measures by computing the Hopkins statistic [21]. Hopkins statistic is to be interpreted heuristically, if it is close to zero (far below 0.5), the data is considered significantly clusterable [21]. The Hopkins statistic is 0.162, 0.183 and 0.187 for clusterings based on health outcomes, unhealthy behaviors, and preventive services respectively, indicating a similar clustering tendency based on all three types of measures.

Computing Distance Measures:
We use the Euclidean distance to measure dissimilarity for the k-means and Manhattan distance for the k-medoids approach to assess pairwise distance between the cities [10].

Running the Clustering Algorithms:
The basic idea behind clustering is to define the clusters so that the total intra-cluster variation (known as total within-cluster variation) is minimized. In k-means clustering, we need to specify k, the optimal number of clusters. Each cluster is represented by its center (i.e. centroid) which corresponds to the mean of the points assigned to the cluster, and the total within-cluster variation is defined as the sum of squared Euclidean distances between items and the corresponding centroid. We used the Elbow method to determine optimal k where the within cluster sum of squares (WSS) is computed for a different number of clusters, k [10]. Then, WSS is plotted against k. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters [10]. Based on the elbow plots, the optimal number of clusters is 2 for each of the three types of measures.

Because the k-means algorithm represents centroids using means, the clustering is highly sensitive to outliers. For better clustering stability, we also used k-medoids (partitioning around medoids) clustering, where centroids are represented using one representative observation from each cluster referred to as medoids. The goal is to find medoids that minimize the sum of the dissimilarities of the observations assigned to each cluster. We constructed two clusters (the same number which was determined to be optimal in the k-means method), by assigning each observation to one of the two nearest medoids.

Interpreting the Clusters:
We labeled the two clusters based on each group of measures as "Good" and "Bad". Tables IV, V, and VI show the k-means centroids for health outcomes, unhealthy behaviors, and preventive services respectively. Good clusters have lower levels of chronic outcomes and unhealthy behaviors, and higher levels of preventive services, bad clusters are the opposite. Table IV shows that the prevalence of all chronic health outcomes except for cancer is lower in the good cluster over the bad cluster. The prevalence of cancer is higher in the good cluster, but only slightly, suggesting that cancer may arise from a combination of unhealthy behaviors, lack of preventive services, and genetic tendencies [9]. Mean unhealthy behaviors are all lower in the good cluster compared to the bad cluster as shown in Table IV. Clustering based on preventive services produces mixed results; on all measures except for routine checkup, good cluster fares better, whereas, the prevalence of routine check up is very slightly lower in the good cluster compared to the bad cluster.

Measure	Good	Bad
Arthritis	20.38	24.86
High BP	27.15	34.37
Cancer	6.05	5.87
CHD	5.01	6.59
COPD	5.08	7.25
Diabetes	8.63	12.25
Cholesterol	30.14	32.84
Kidney	2.48	3.09
Mental Health	11.14	14.03
Stroke	2.55	3.67
Teeth Loss	11.09	18.72
Phys Health	10.86	14.69

Table IV: Centroids – Health Outcomes

Measure	Good	Bad
Smoking	14.41	20.75
Obesity	24.77	33.83
Lack Sleep	32.92	38.45
Lack Phys Act	20.76	30.95

Table V: Centroids – Unhealthy Behaviors

Measure	Good	Bad
Lack Insurance	12.64	21.29
Routine Checkup	67.75	68.48
BPMed	56.86	59.68
Cholesterol Screen	75.81	72.32
Colon Screen	66.20	58.12
Core Men	34.06	29.04
Core Women	33.34	27.05
Dental	67.22	55.47
Mammogram	79.43	75.74
Pap	82.60	78.18

Table VI: Centroids – Preventive Services

B. Similarity Assessment

Each city is labeled as "G" or "B" in each of the three clustering solutions, yielding three labels after the clustering process. We define the Rand Index to assess the similarity between clustering solutions, based on the example of unhealthy behaviors and chronic outcomes. We build a 2 x 2 contingency table from the two labels based on two measures assigned to each city as shown in Table VII. $n_{G,G}$ represents the number of cities labeled as "G" in both clustering solutions

Unhealthy Behaviors	Health Outcomes		
		G	B
	G	$n_{G,G}$	$n_{G,B}$
	B	$n_{B,G}$	$n_{B,B}$

Table VII: Contingency Table

Comparison	k-means	k-medoids
HO vs. UB	0.90	0.92
HO vs. PS	0.636	0.654

Table VIII: Clustering Similarity – Rand Index

based on chronic health outcomes and unhealthy behaviors and $n_{B,B}$ represents the number of cities labeled as "B" in both solutions. Cells (G, G) and (B, B) consist of cities that conform to intuition, where good health outcomes coincide with lower levels of unhealthy behaviors, and bad health outcomes coincide with higher levels of unhealthy behaviors. Cells (G, B) and (B, G) represent cities that are a mismatch. (G, B) indicate cities with bad health outcomes despite having lower degrees of unhealthy behaviors, whereas (B, G) indicate cities with better health outcomes despite higher degrees of unhealthy behaviors. This contingency table is akin to confusion matrix in supervised machine learning.

This two-way contingency table allows us to define the Rand Index [20] to measure the similarity between the two clustering solutions in Equation (1). Usually, the Rand index is used to measure the level of agreement between different clusterings produced on the same data set using different methods [20]. Here we use the index in a novel way to assess the agreement between clusterings of the same set of observations on two different feature sets. The Rand index takes a value between 0 and 1, with 0 indicating that the two data clusterings do not agree on any pair of points, and 1 indicating that they agree exactly. We define the Rand index as:

$$Rand\ Index = \frac{n_{G,G} + n_{B,B}}{n_{G,G} + n_{G,B} + n_{B,G} + n_{B,B}} \quad (1)$$

Rand indices were also computed for clusterings based on chronic health outcomes and preventive services for both the k-means and k-medoids methods. These values are reported in Table VIII. The table shows that regardless of the clustering method (k-means vs. k-medoids), the agreement between the clusterings based on health outcomes and unhealthy behaviors is over 90% and it is 15% higher compared to the agreement between clusterings based on health outcomes and preventive services. These results suggest that unhealthy behaviors have a stronger influence on chronic health outcomes, and this adverse influence may not be mitigated by the use of clinical preventive services. Recent guidelines for pursuing clinical preventive strategies have turned more restrictive, including lower tolerance thresholds for blood pressure [19] and cholesterol [17] to trigger prescription medications. This poses the question whether these lowered thresholds produce any benefit, and if managing risk factors aggressively can turn healthy people into chronic patients without much advantage [13].

An overwhelming majority – over 90% of the cities showed agreement between the degree of chronic health outcomes and unhealthy behaviors. Just about 40 cities showed a mismatch; that is, they belonged to the good cluster according to their health outcomes, and bad cluster according to the unhealthy behaviors, and vice versa. Next, we investigate whether there exist any geographical trends among these mismatch cities. Figure 1 shows the geographical spread of the cities with both types of mismatch. In the figure, *B-HO/G-UB* represents cities where the health outcomes are bad despite their residents maintaining a low degree of unhealthy behaviors. On the other hand, *G-HO/B-UB* represent cities with good health outcomes despite their residents having a higher degree of unhealthy behaviors. As the figure shows, *B-HO/G-UB* is mostly a coastal phenomenon, whereas, *G-HO/B-UB* is more or less an inland occurrence. Of these, many *B-HO/G-UB* cities appear to be from California and the Pacific Northwest. Charlotte and Boston are the only large/medium cities on the east coast, and North Carolina is the only state on the east coast with three cities in this group. New York, New Jersey and Rhode Island are the two coastal states in the *G-HO/B-UB* group.

IV. Related Research

Since its release, the 500 cities data has been the subject of several research efforts. Klompas *et. al.* [12] obtain the prevalence of diabetes, obesity, hypertension, and asthma using health records, and compare these to the small-area estimates from the 500 cities data. Liu *et. al.* [15] apply clustering analysis to the 500 cities data to identify patterns of kidney disease. Camille Seaberry apply clustering analysis to group census tracts from the eight cities in Connecticut into two groups, and ultimately link them to social determinants of health [18]. Our previous work has used clustering to compare the health outcomes in the San Francisco and Boston metro areas [8]. The 500 Cities data has also been included in the city health dashboard tool [1], [2] that allows users to visualize health data from multiple sources.

These works limit their scope either by the data or the geography or both. Liu *et. al.* consider all the 50 U.S. states, Camille Seaberry restrict themselves to the cities in Connecticut, and our prior work considers metro areas on two opposite coasts. Moreover, Liu *et. al.* analyze only one health outcome, while Camile Seaberry and our prior work consider all the health outcomes but not the other measures. The present work includes all three types of measures as well as all the 500 cities, and thus enjoys enjoys a broader scope both in terms of the data and geography.

The work closest to the research reported in this paper is by Liu *et. al.* [16], which finds associations between unhealthy behaviors, clinical preventive service use, and chronic disease outcomes. Their work finds two-way associations between health outcomes and the other two measures, but it does not compare the relative influence of these two types of measures.

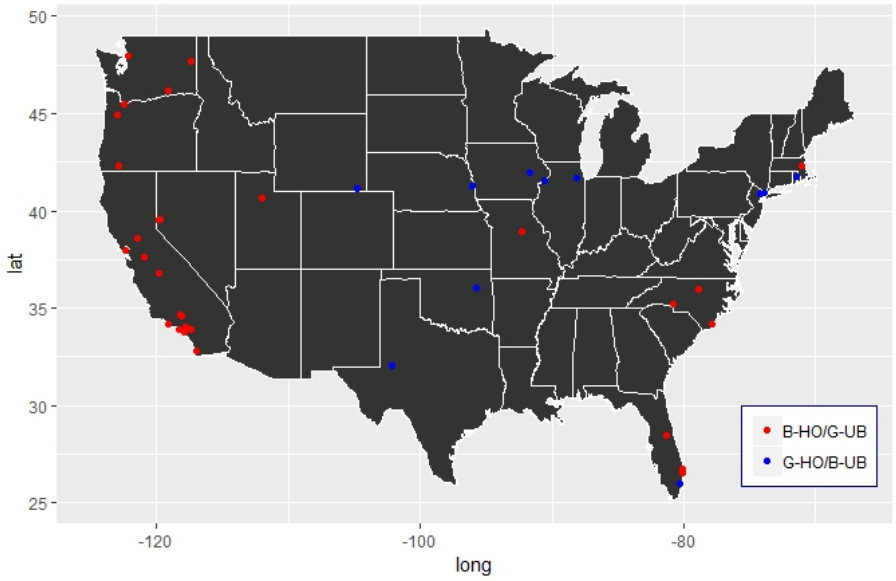

Figure 1: Geographical Spread of Mismatch Cities

By contrast, the goal of our approach is to determine the relative impact of the two types of measures.

V. Conclusions & Future Research

This paper assesses the relative impact of unhealthy behaviors and clinical preventive services on chronic health outcomes using a two-step approach. The results indicate that unhealthy behaviors have a significantly higher adverse influence on health outcomes, which perhaps cannot be alleviated by the use of clinical preventive services. This finding raises the question whether more investments are necessary in promoting healthy choices rather than aggressively pushing people to avail clinical preventive services.

Data from the 500 Cities project can also be used to understand how socioeconomic, demographic and geographical factors correlate with health measures [4]. Our future research is concerned with understanding these interactions.

References

[1] "City Health Dashboard". https://med.nyu.edu/pophealth/city-health-dashboard. Accessed: 2019-01-21.

[2] "City Health Dashboard Expanding to 500 More Cities". https://wagner.nyu.edu/news/story/city-health-dashboard-expanding-500-more-cities#. Accessed: 2019-01-21.

[3] M. Bryant. "ACA Increased Primary Care Usage, but Also Some Risky Behaviors, Report Finds". https://www.healthcaredive.com/, April 2018.

[4] K. M. Fitzpatrick, X. Shi, D. Willis, and J. Niemeier. "Obesity and Place: Chronic Disease in the 500 Largest U.S. Cities". *Obes Res Clin Pract*, 12(5):421–425, Sep-Oct 2018.

[5] Center for Disease Control and Prevention. "500 Cities: Local Data for Better Health". https://www.cdc.gov/500cities/index.htm, November 2017. Accessed: 2019-01-21.

[6] Center for Disease Control and Prevention. "The 500 Cities Project: Local Data for Better Health". https://www.cdc.gov/500cities/pdf/500Cities-FactSheet-082217.pdf, November 2017. Accessed: 2019-01-21.

[7] Center for Disease Control and Prevention. "About Chronic Diseases". https://www.cdc.gov/chronicdisease/about/index.htm, October 2019. Accessed: 2019-01-21.

[8] S. Gokhale. "Comparing Health Outcomes in San Francisco and Boston Metro Areas". In *Proc. of IEEE Annual Computer Software and Applications Conference*, pages 740–745, Milwaukee, WI, July 2019.

[9] S. Gokhale. "Quantifying the Relationship Between Health Outcomes and Unhealthy Habits". In *Proc. of Intl. Conf. on Software Engineering and Knowledge Engineering*, Pittsburgh, PA, July 2020.

[10] UC Business Analytics R Programming Guide. "K-Means Cluster Analysis". https://uc-r.github.io/kmeans_clustering. Accessed: 2020-01-21.

[11] HealthyPeople.gov. "Clinical Preventive Services". https://www.healthypeople.gov/2020/leading-health-indicators/2020-lhi-topics/Clinical-Preventive-Services, 2020. Accessed: 2020-07-21.

[12] M. Klompas, N. M. Cocoros, J. J. Menchaca, E.Hafer, B. Herrick, and M. Josephson. "State and Local Disease Surveillance using Electronic Health Record Systems". *American Journal of Public Health*, August 2017.

[13] M. J. Kreiner and L. M. Hunt. "The Pursuit of Preventive Care for Chronic Illness: Turning Healthy People into Chronic Patients". *Sociol Health Illn.*, 36(6):870–874, July 2013.

[14] S. Levine, E. Malone, A. Lekiachvili, and P. Briss. "Health Care Industry Insights: Why the Use of Preventive Services Is Still Low". *Preventing Chronic Disease*, 16, March 2019.

[15] S. H. Liu, Y. Li, and B. Liu. "Exploratory Cluster Analysis to Identify Patterns of Chronic Kidney Diseases in the 500 Cities Project". *Preventing Chronic Diseases*, May 2018.

[16] S. H. Liu, B. Liu, and Y. Li. "Risk Factors Associated with Multiple Correlated Health Outcomes in the 500 Cities Project". *Preventive Medicine*, 112:126–129, July 2018.

[17] Johns Hopkins Medicine. "2018 Cholesterol Guidelines for Heart Health Announced". https://www.hopkinsmedicine.org/, November 2018. Accessed: 2020-07-21.

[18] M. Camille Seaberry. "Merging 500 Cities and Connecticut Data on Health Equity". https://preview.tinyurl.com/yypnzm5, February 2017. Accessed: 2019-01-21.

[19] ACC News Story. "New ACC/AHA High Blood Pressure Guidelines Lower Definition of Hypertension". https://www.acc.org/latest-in-cardiology/articles/2017/11/08/11/47/mon-5pm-bp-guideline-aha-2017, November 2017. Accessed: 2020-07-21.

[20] D. Tang. "The Rand Index". https://davetang.org/muse/2017/09/21/the-rand-index/, September 2017. Accessed: 2019-01-21.

[21] L. YiLan and Z. RuTong. "Check the Clustering Tendency". https://cran.r-project.org/web/packages/clustertend/clustertend.pdf, May 2015. Accessed: 2020-07-21.