Quantifying the Relationship Between Health Outcomes and Unhealthy Habits

Swapna S. Gokhale
Dept. of Computer Science & Engg.
Univ. of Connecticut, Storrs, CT 06269
{swapna.gokhale}@uconn.edu

Abstract

Chronic health outcomes impact the quality of life of affected individuals and their families and also lead to huge health care costs. Most of the chronic health outcomes can be attributed to few unhealthy behaviors, however, the extent to which these behaviors can explain the variation in the common outcomes is not known. This paper explores the relationship between: (i) unhealthy behaviors using principal components analysis; and (ii) unhealthy behaviors and chronic health outcomes using multiple linear regression. The 500 Cities data, released by the Center for Disease Control, forms the basis of this investigation. PCA suggests that the unhealthy behaviors can be projected along two dimensions, each punctuated by the common age of occurrence. The results of linear regression are consistent with expectations for some outcomes, but reveal unexpected trends for the others.

1 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 impact the quality of lives of the affected individual as well as their families. Moreover, they are the leading causes of death and disability, and drivers of the nation's $3.3 trillion in annual health care costs. The CDC estimates that six in ten adults in the U.S. have one chronic disease, and four in ten adults have two or more [6].

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 [14]. The association between these risk factors and chronic diseases is known qualitatively. What is not known, however, is the relationship of these unhealthy behaviors with each other, and the extent to which these behaviors contribute to specific chronic health outcomes. It is crucial to quantify the level of variance in the different health outcomes that can be explained by risky behaviors; because then the search for what leads to unexplained or residual variance can begin in earnest. These additional causes, beyond unhealthy or risky behaviors, may be found in other factors such as environmental stressors and genetic predisposition.

In this paper, we explore the relationship among the unhealthy behaviors themselves, and between unhealthy behaviors and chronic health outcomes. The 500 cities data [7], which provides city and census-tract level small area estimates for chronic disease risk factors, health outcomes, and clinical preventive service use for the largest 500 cities in the United States forms the basis of our investigation. Principal components analysis is used to study how unhealthy behaviors cluster together, and multiple linear regression is used to relate these behaviors to the health outcomes. Our results suggest that the five unhealthy behaviors can be mapped to two dimensions. The first dimension accounts for approximately 68% of the variation and comprises of habits that may mostly develop around the middle age, whereas the second dimension includes only binge drinking which is more prevalent among the younger population. The results of multiple linear regression confirm that a large percentage of variation in coronary heart disease, stroke, high cholesterol, COPD, and diabetes can be attributed to unhealthy behaviors. However, a relatively lower percentage of variation in high blood pressure, which is viewed as a risk factor for heart disease and stroke, and asthma which is considered a risk factor for COPD can be explained by unhealthy habits. Moreover, it appears surprising that over 80% of the variation in arthritis and teeth loss, two conditions that co-exist with aging-related deterioration, is attributable to unhealthy habits. Finally, only about 50% of the variance in cancer is explainable by unhealthy behaviors, suggesting the presence of strong genetic and/or environmental influences.

The rest of the paper is organized as follows: Section 2 summarizes the 500 cities data. Section 3 and Section 4

DOI reference number: 10.18293/SEKE2020-116

discuss principal components and linear regression analysis respectively. Section 5 compares related research. Section 6 offers concluding remarks and future research directions.

2 The 500 Cities Data

The 500 Cities Project is a collaboration between the Center for Disease Control (CDC), the Robert Wood Johnson Foundation, and the CDC Foundation. The purpose of the 500 Cities Project is to provide city and census-tract level small area estimates for 5 unhealthy behaviors, 13 health outcomes, and 11 clinical preventive service use for the largest 500 cities in the United States [7]. These 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 [9]. These measures are estimated using the raw data from the CDCs Behavioral Risk Factor Surveillance System [8], using a multi-level statistical modeling framework [10].

In this paper, we considered the 13 health outcomes and 5 unhealthy behaviors from the 500 Cities Project. Tables 1 and 2 offer a brief summary, significance, and mean prevalence of these measures. In Table 1, all the health outcomes, except for mental and physical health, are formally diagnosed by medical professionals whereas estimates of (lack of) mental and phsyical health are self-reported [9].

3 Principal Components Analysis

Principal Components Analysis (PCA) uses an orthogonal transformation to convert a set of observations with correlated variables into a set of values of linearly uncorrelated variables called principal components [12]. The first principal component has the largest possible variance, that is, it accounts for as much variability in the data as possible. Each succeeding principal component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. PCA creates as many new independent variables as there exist in the original data. Usually, however, the first few independent variables can explain a large percentage of the variation in the data and are retained for analysis, while the others that contribute very little to the variability are eliminated in favor of model parsimony. PCA is therefore also referred to as a feature extraction or dimensionality reduction procedure.

We apply PCA to uncover the relationships among the 5 unhealthy behaviors. The cumulative variability explained by the principal components is represented using a scree plot in Figure 1, which shows that the first two dimensions account for about 84% of the variation. Focusing on these two dimensions, our next step was to investigate the contribution of unhealthy behaviors to each as shown in Figures 3 and 4. The contribution of each variable is represented as a percentage, where the red dashed lines are reference lines that correspond to the expected contribution if each variable pitched uniformly. With 5 original variables, the reference lines are shown at 20%. Variables with contributions above the reference line are considered important for that dimension. According to this heuristic, in the figure, three variables, namely, lack of physical activity, obesity, and smoking are important contributors to the first dimension. Of these, lack of physical activity and obesity contribute predominantly, while smoking is just barely above the reference line. Because lack of activity and obesity usually develop around middle age, we label this dimension as "Midlife Crisis". For the second dimension, only binge drinking contributes more than 20%, which tends to occur in younger adults, and hence, we label this dimension as "Youthful Adventures". The graph of PCA variables shows the orthogonal projection of the five behaviors along the two dimensions as shown in Figure 2. Figures 5 and 6 show that the top 20 cities contribute more than the uniform 0.2% towards each dimension. Contributors to Midlife Crisis concentrate in the Midwest and Mountain States, whereas Youthful Adventures cluster along the East and West coasts as shown in Figure 7.

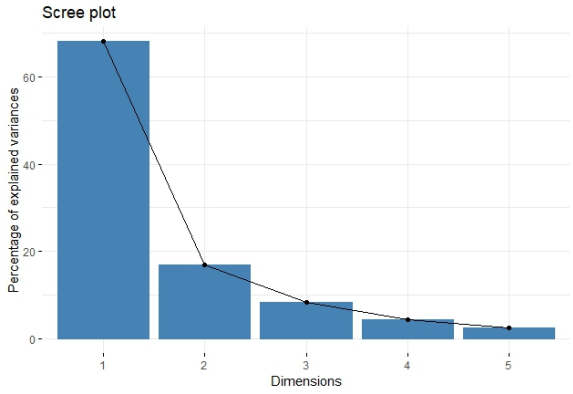

Figure 1. Scree Plot

4 Multiple Linear Regression

We postulate a linear relationship between health outcome i, and unhealthy behaviors $UB_1, \ldots, UB_5$ given by:

$$HO_i = \beta_{i,0} + \sum_{j=1}^{5} \beta_{i,j} * UB_j + e_i \qquad (1)$$

The key assumption underlying least squares linear regression models is homoskedasticity, which implies that the variations for all the observations in a data set are equal.

Table 1. Chronic Health Outcomes: Significance & Prevalence

	Health Outcome	Mean
HO_1	**Arthritis:** Reduces physical function, quality of life.	22.39
HO_2	**Asthma:** ED visits, hospitalizations, missed work, comorbid depression.	9.18
HO_3	**Cancer:** Still a leading cause of death, second to heart disease.	5.98
HO_4	**Chronic Kidney Disease:** Ninth leading cause of death, but most affected don't know.	2.75
HO_5	**Chronic Obstructive Pulmonary Disease (COPD):** Impaired pulmonary function, which often goes undiagnosed.	6.05
HO_6	**Coronary Heart Disease (CHD):** Common form of heart disease, leading cause of death	5.73
HO_7	**High BP:** Responsible for $20-30\%$ CHD, $20-50\%$ Stroke, cardiovascular complications.	30.39
HO_8	**High Cholesterol:** Responsible for $30-40\%$ CHD, $10-20\%$ strokes.	31.35
HO_9	**Diabetes:** Impaired glucose function, complications if not managed.	10.25
HO_{10}	**Mental Health:** Not good for more than 14 days. Related to diabetes, cancer, cardiovascular disease, asthma, obesity. Many risk factors; physical inactivity, smoking, binge drinking, insufficient sleep also contribute to mental illness.	12.44
HO_{11}	**Physical Health:** Not good for more than 14 days. Related to health-related quality of life.	12.57
HO_{12}	**Stroke:** 1 out of 20 deaths, serious long-term disability.	3.05
HO_{13}	**Teeth Loss:** Reduces quality of life, self-image, and daily functioning (>65 years old).	14.51

Table 2. Unhealthy Behaviors: Significance & Prevalence

	Unhealthy Behavior	Mean
UB_1	**Current Smoking:** Greater than 100 cigarettes and smoke every day or most days. Increases the risk for heart disease, stroke, multiple types of cancer, and chronic lung disease.	17.58
UB_2	**Binge Drinking:** Five or more drinks (men), four or more drinks (women) at one time. Accounts for over 40,000 deaths and 1 million years of potential life lost annually. Health and social problems such as motor-vehicle crashes, violence, suicide, hypertension, acute myocardial infarction, STDs, unintended pregnancies, fetal alcohol spectrum disorders, sudden infant death syndrome.	16.53
UB_3	**No Leisure Time Physical Activity (LoPA):** Other than their regular job, did not participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking. Improve the health and quality of life of all ages, regardless of chronic disease or disability. Lower the risk for early death, coronary heart disease, stroke, high blood pressure, type 2 diabetes, breast and colon cancer, falls, and depression.	25.86
UB_4	**Obesity:** Body mass index (BMI) greater than $30.0\ kg/m^2$. Increases the risk for multiple chronic diseases, including heart disease, stroke, hypertension, type 2 diabetes, osteoarthritis, and certain cancers.	29.31
UB_5	**Sleeping less than 7 hours (LoS):** Insufficient sleep (<7 hours), on an average, during a 24-hour period. Associated with chronic conditions such as diabetes, cardiovascular disease, hypertension, obesity, and depression. May cause motor vehicle crashes and industrial errors, causing substantial injury and disability. Reduces productivity and quality of life.	35.69

Most real-world data sets will probably be heteroskeastic [15], but it is possible to use the least squares model for large enough sample sizes, which is the case here. In Equation (1), HO_i's are the predicted or response variables, and $UB_1, \ldots, UB_5$ are the independent or predictor variables, often known as regressors. The coefficients $\beta_{i,j}$, $j = 0, \ldots, 5$ are estimated by minimizing the sum of squared unexplained parts. The coefficient of determination R^2 is given by Equation (2), where $\hat{HO}_{i,k}$ is the estimate of the health outcome i for the k^{th} city produced by the model, and $\bar{HO}_i$ is the mean value of the outcome i across all the 500 cities. R^2 measures the proportion of variation in HO_i that can be explained by the regressors $UB_1, \ldots, UB_5$.

$$R_i{}^2 = \frac{ModelSS}{TotalSS} = \frac{\sum_{k=1}^{500}(\hat{HO}_i - \bar{HO}_{i,k})^2}{\sum_{k=1}^{500}(HO_{i,k} - \bar{HO_i}{}^2)} \quad (2)$$

For each health outcome, a p-value is also estimated by

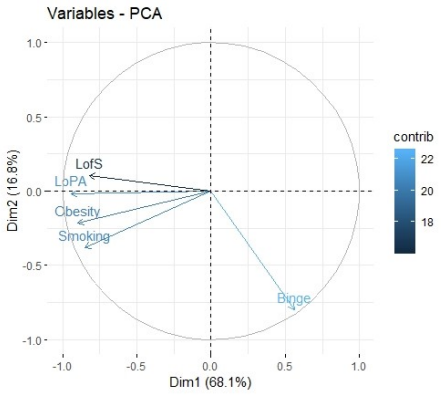

Figure 2. Graph of Variables – PCA

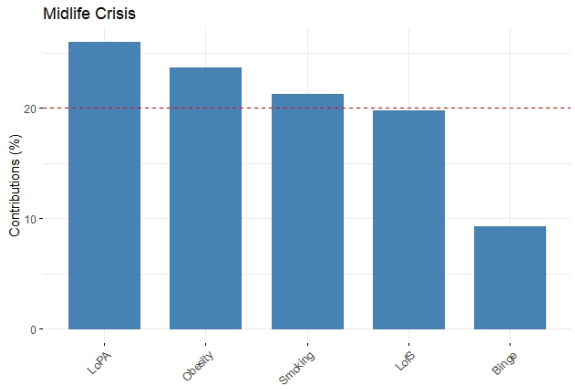

Figure 3. Unhealthy Behaviors ==> Midlife Crisis

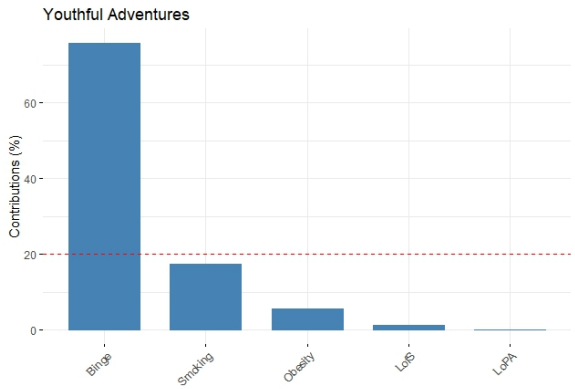

Figure 4. Unhealthy Behaviors ==> Youthful Adventures

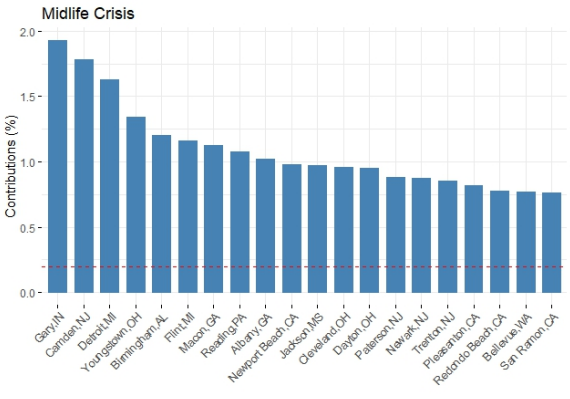

Figure 5. Cities ==> Midlife Crisis

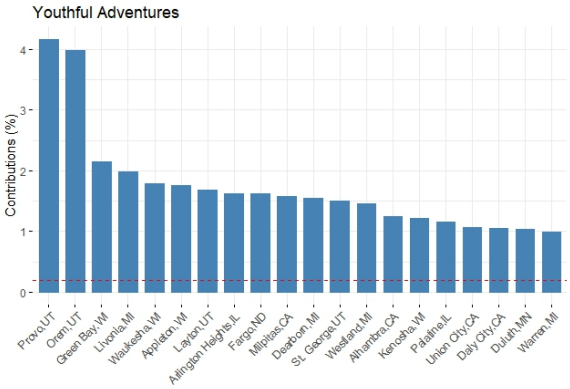

Figure 6. Cities ==> Youthful Adventures

Figure 7. Cities ==> Dimensions, Geographical Spread

the model for all unhealthy behaviors, which is compared against the typical level of significance $\alpha = 0.05$. If the p-value is less than α, then the effect of that specific behavior on the particular health outcome is significant. Table 3 shows the results of the regression model. For each health outcome, the table lists the t-statistic and p-values for each unhealthy behavior, and R^2 which explains the total varia-

tion that can be attributed collectively to these behaviors. We divide the outcomes into two groups I and II, these groups respectively comprise of the outcomes for which over 80% and less than 80% is explained by the unhealthy habits. The results confirm certain expectations, but also reveal anomalies. Only for about 50% of the outcomes, all the five unhealthy behaviors are statistically significant. These include CHD, stroke, teeth loss, diabetes, cancer, asthma, and physical health. High blood pressure and high cholesterol, are the two common precursors to CHD and stroke, however, these belong to first and second groups respectively. Thus, although high blood pressure may be mostly attributed to unhealthy behaviors making up the midlife crisis, high cholesterol may have additional origins. Transitioning from high cholesterol which is relatively benign, to life threatening conditions such as stroke and CHD, however, may be precipitated by lifestyle choices. A similar relationship can be seen between COPD and asthma, COPD belongs to the first group, but asthma which is considered a risk factor belongs to the second. Although all the unhealthy behaviors are statistically significant for cancer and asthma, collectively they explain only about 50% and 73% variation respectively. This suggests that in cancer and asthma genetics and the environment [5] may interplay with lifestyle choices. A few additional significant and interesting observations include: Smoking is significant for all the health outcomes except high cholesterol. Binge drinking is not significant for high blood pressure and COPD. Lack of mental health is not influenced by either binge drinking or lack of sleep. Obesity is not significant for kidney disease. Finally, although teeth loss and arthritis are mainly dominant in aging populations, they are also members of the first group. This indicates that the influence of lifestyle choices is not limited to metabolic conditions of high blood pressure, high cholesterol and diabetes.

5 Related Research

The association between chronic diseases and lifestyle choices is generally known, however, very few studies have sought to quantify this association. Adaji *et. al.* [1] use logistic regression to identify the risk factors associated with some common chronic conditions (arthritis, angina, stroke, diabetes, and chronic lungs disorder) among people over 50 years in India. The model includes socioeconomic and demographic factors and the interplay between the conditions. A similar study by Ismail *et. al.* [11] is conducted for the younger Indian population but only for coronary heart disease. Zhao *et. al.* [16] estimate the prevalence and correlates of chronic diseases in an elderly population in Haikou. Four major chronic conditions, namely, hypertension, diabetes, COPD and stroke and sociodemographic characteristics and lifestyle factors are considered in the study. Regression analysis has been used in the context of chronic conditions to estimate the various types of burdens, including health care costs, absenteeism and employer costs associated with these conditions [3, 13]. In contrast, our research analyzes how the variance in a variety of chronic conditions can be attributed to five core unhealthy behaviors, regardless of the other socioeconomic and demographic factors.

6 Conclusions and Future Research

This paper explores the relationship among common unhealthy behaviors, and their influence on prevalent chronic health outcomes quantitatively. The analysis uses the 500 Cities data, which provides small area estimates of 27 health-related measures for 500 largest cities in the United States. PCA is used to map the unhealthy behaviors to orthogonal dimensions to understand their co-occurrence, and multiple linear regression is used to explore how these unhealthy behaviors relate to chronic health outcomes. PCA dimensions can be readily interpreted within the context of age. However, the results of multiple linear regression expose some interesting, and unexpected tendencies.

Our future research involves relating the health outcomes at the level of census tracts to demographic and socioeconomic data available from the U.S. Census Bureau [4].

References

[1] E. E. Adaji, A. S. Ahankari, and P. R. Myles. "An Investigation to Identify Potential Risk Factors Associated with Chronic Diseases Among the Older Population in India". *Indian J. Community Med*, 42(1):46–52, 2017.

[2] E. C. Alexopoulos. "Introduction to Multivariate Regression Analysis". *Hippokratia*, 14(Suppl 1):23–28, December 2010.

[3] G. R. B. Asay, K. Roy, J. E. Lang, R. L. Payne, and D. H. Howard. "Absenteeism and Employer Costs Associated with Chronic Diseases and Health Risk Factors in the US Workforce". *Prev Chronic Dis*, 2016.

[4] United States Census Bureau. "Census Data API User Guide". `https://www.census.gov/data/developers/guidance/api-user-guide.html`, June 2017. Accessed: 2019-01-21.

[5] Illinois Disability and Health Program. "What is Chronic Disease? Important Things to Know About Chronic Diseases for Persons with Disabilities". `http://www.idph.state.il.us/idhp/idhp_ChronicDisease.htm`. Accessed: 2020-01-21.

Table 3. Results of Multiple Linear Regression Model

HO	T-statistic, p-values					F-statistic	R^2
	Smoking	Obesity	LoS	LoPA	Binge		
Group I: Over 80% Variation							
Arthritis	$1.04e-08$	$<2e-16$	$<2e-16$	0.948	$<2e-16$	$<2.2e-16$	0.8389
High BP	$6.29e-14$	$6.86e-08$	$<2e-16$	$<2e-16$	0.267	$<2.2e-16$	0.8621
CHD	$<2e-16$	$<2e-16$	$<2e-16$	$<2e-16$	0.00016	$<2.2e-16$	0.8706
COPD	$3.43e-15$	$<2e-16$	0.00101	0.02320	0.4887	$<2.2e-16$	0.903
Diabetes	$3.29e-12$	$1.43e-09$	$<2e-16$	$<2e-16$	$<2e-16$	$<2.2e-19$	0.8223
Stroke	$<2e-16$	$<2e-16$	$<2e-16$	$<2e-16$	0.0263	$<2.2e-19$	0.8442
Teeth Loss	$<2e-16$	$<2e-16$	0.00281	0.00180	$1.06e-09$	$<2.2e-19$	0.8669
Group II: Less than 80% Variation							
Cancer	0.04591	$<2e-16$	0.00312	$2.34e-15$	$<2e-16$	$<2.2e-16$	0.517
Asthma	0.000209	$<2e-16$	$3.07e-05$	$5.78e-12$	$2.62e-15$	$<2.2e-16$	0.641
High Chol.	0.1504	0.0673	$<2e-16$	0.4080	$1.81e-07$	$<2.2e-16$	0.6354
Kidney	$1.71e-11$	0.517	$1.65e-13$	$2.32e-09$	$8.59e-05$	$<2.2e-16$	0.7317
Mntl. Hlth	$5.89e-08$	$<2e-16$	0.694	$<2e-16$	0.140	$<2.2e-16$	0.7394
Phys. Hlth	$3.65e-11$	$2.13e-10$	0.0205	$1.34e-06$	$2.99e-08$	$<2.2e-16$	0.7372

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

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

[8] Center for Disease Control and Prevention. "Behavioral Risk Factor Surveillance System". https://www.cdc.gov/brfss/index.html, April 2017. Accessed: 2019-01-21.

[9] Center for Disease Control and Prevention. "The 500 Cities Project: Local Data for Better Health, Measures Definition". https://www.cdc.gov/500cities/measure-definitions.htm, November 2017. Accessed: 2019-01-21.

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

[11] B. Ismail and M. Anil. "Regression Methods for Analyzing the Risk Factors for a Life Style Disease Among the Young Population of India". *Indian Heart J.*, 66(6):587–592, November-December 2014.

[12] K. Khan. "Principal Component Analysis - An Introduction with R Implementation". https://rpubs.com/koushikstat/pca. Accessed: 2020-01-21.

[13] H. H. Konig, H. Leicht, H. Bicket, A. Fuchs, and et. al. J. Genischen. "Effects of Multiple Chronic Conditions on Health Care Costs: An Analysis based on an Advanced Tree-Based Regression Model". *BMC Health Services Research*, 2013.

[14] W. C. Willett, J. P. Koplan, R. Nugent, C. Dusenbury, P. Puska, and T. A. Gaziano. "Prevention of Chronic Disease by Means of Diet and Lifestyle Changes". In D. T. Jamison, J. G. Breman, and A. R. Measham, editors, *Disease Control Priorities in Developing Countries, 2nd Edition*, chapter 44. he International Bank for Reconstruction and Development / The World Bank, 2006.

[15] C. Yobero. "Methods for Detecting and Resolving Heteroskedasticity". https://rpubs.com/cyobero/187387, June 2016. Accessed: 2020-01-21.

[16] C. Zhao, L. Wong, Q. Zhu, and H. Yang. "Prevalence and Correlates of Chronic Diseases in an Elderly Population in an Elderly Population: A Community Survey in Haikou". *PLoS One*, June 2018.