Behavioral Drivers of Preventative and Chronic Health: A Multivariate Approach

Harini Reddy Pisati

Executive Summary

This study sought to define the relevant relationship drivers between chronic health concerns, preventative health measures, and human behavior. As this is a broad topic with myriad approaches possible, the team dove into narrower research questions and applied varying analytic techniques. The team's goal was to create an initial framework to understand which human behaviors bore the most bearing on common chronic illnesses. To approach this topic, the team relied on the Behavioral Risk Factor Surveillance System, a large-scale annual survey conducted by the Centers of Disease Control. More than 400,000 American adults were surveyed during the 2015 study period, creating a varied and representative data pool. The team chose to tackle four questions during this study:

1. How do chronic health conditions affect both quality of life and feelings of being emotionally supported?
2. How do diabetes risk factors relate? How can these factors be grouped into meaningful latent constructs?
3. How do behavioral, demographic, and health factors impact an individual's overall health?
4. What clusters of behavior do women exhibit when it comes to preventative health?

Finally, because the study produces an enormously rich data set annually, exploring more than 300 questions with each of the 400,000+ participants, the team elected to tackle a fifth question:

5. What variables can be grouped together into principal components to reduce the number of variables available and explain the variation in the study's individuals?

A distinct analytic technique was used to explore each of these five questions. For question 1, the team deployed a canonical correlation analysis to explore how quality of life and feelings of being emotionally supported varied with health condition variables. This first analysis covers the broadest spectrum of chronic illness concerns, including variables pertaining to heart disease, stroke, cancer, asthma and other lung ailments, arthritis, diabetes, and depression. The second question, a deep dive into diabetes risk factors, with variables selected in consultation with previous work in the field, relied upon an initial exploratory factor analysis as well as a supplementary confirmatory factor analysis testing the three discovered factors[AD1] . Our third question was answered through a multiple regression exploring predictors of general health status[AD2] . For the fourth question, pertaining to women's behaviors and preventative health, a hierarchical agglomerative clustering algorithm helped define group characteristics. This method used a variety of demographic and behavioral questions, in addition to questions about preventative health measures The last question employed a principal component analysis on continuous measures to attempt to create a model for reducing the variables available with this annual survey.

Our results provide an exciting glimpse into the underlying relationship between human behavior and incidence of chronic disease. In answer to the first research question, the team found a negative relationship between feelings of emotional support and life satisfaction and having a chronic disease. This finding supports existing research into how chronic illness has a detrimental

effect on one's overall emotional well-being. The analysis also found that depression, out of all chronic illnesses included, had the most impact on this relationship. Our exploration of diabetes risk factors revealed three factors: Lifestyle (primarily composed of physical activity, fruit/vegetable consumption, and alcohol consumption), Medical History (with variables such as diabetes, high blood pressure, history of heart disease/attack, and general health), and Socioeconomic Status (comprised of variables such as mental health, income, and inability to seek medical attention due to cost). Furthermore, the team found a negative relationship between medical history and socioeconomic status; in other words, an increased incidence of health concerns in one's life was associated with lower socioeconomic status. Because this analysis found mental health to be associated with socioeconomic status, this finding dovetails nicely with the first research question's finding of depression's influence on life satisfaction and emotional support.

Our third question, concerning predictors of overall health, found that high blood pressure, activity limitation due to health problems, history of heart disease or attack, diabetes, and other illness variables lowered overall health status. Other behavioral and demographic factors, such as education and employment type, also played a role in predicting overall health. The regression model used explains approximately 57% of the variation in overall health, which is a good result for an initial model. The fourth health question, regarding clusters of women's behaviors around preventative health, found distinctive group behaviors regarding preventative health measures such as mammograms, Pap smears, mental health, smoking, and HIV testing. These groups were largely organized around demographic factors, such as age, income, and employment.

Finally, our attempt to simplify the quantitative variables present in the survey discovered component groupings in Physical/Mental Health (including height and weight variables), Vegetable Consumption, and Number of Household Members (children and adults). A fourth component suggests mental and physical health could stand independently of height/weight variables.

While our study has promising initial findings, there are limitations to broadly applying the models. Most notably, because the study is conducted by telephone, the responses are limited to those who were available for a lengthy, and at times deeply probing, phone call from a stranger. As with any survey, answers may be overall more positive than reality. The study's respondents also skewed towards older adults. Future work in this area should consider examining individual chronic conditions one at a time - though our study attempted a broad overview of chronic disease at large, it is likely that the results could look very different if narrowed to one or two (related) illnesses. Future study could also entail intentionally surveying younger, working adults to have a broader sample from a younger demographic, and analyzing the results again using lenses of race, gender identity, and sexual orientation, which may display their own predictive or clustering behaviors.

In conclusion, many behaviors have influences on a person's overall health and well-being. Our analyses complemented a general understanding of human behavior, demographic conditions, and incidence of disease. Our team found that behaviors do influence incidence of chronic disease, and in turn, suffering a chronic physical illness influences behavior and mental health. We also found interesting patterns among behaviors towards preventative health screenings.

Behavioral Drivers of Preventive and Chronic Health: A Multivariate Approach

Rasa Willette, Jessica Bicek, Alexandra DeGrandchamp, Mahender Kunchala, Harini Reddy Pisati

Abstract

This analysis was conducted to better understand risk behaviors that can influence aspects of physical well-being. The study looks at the following: how chronic illness impacts quality of life, diabetes risk factor relationships, how physical activity and other similar factors impact individual health, what behaviors women exhibit around preventative health, and whether principal component analysis (PCA) can be used to reduce dimensionality by creating smaller subsets that explain most of the variability in the original dataset. The results were processed using a variety of methods in addition to the PCA mentioned before: canonical component analysis (CCA), confirmatory factor analysis (CFA), multiple regression, and hierarchical cluster analysis (HAC). This study was conducted using data collected via cellular or landline telephone from the Centers for Disease Control and Prevention (CDC). The results revealed that chronic illness can negatively impact quality of life, there are three main risk factors contributing to the incidence of diabetes, chronic illnesses and lifestyle factors impact an individual's general health, and women fall into four main cluster groupings of behaviors around preventative health measures. In conclusion, it was determined that many different risk factors influence various aspects of a person's health. It is important to conduct these kinds of studies to better educate the general population on actions they can take to prevent avoidable illness and disease.

Introduction

In an era plagued with sky-high hospital bills, myriad diet fads, a complex insurance landscape, and a society still reeling from the ravages of a global pandemic, the stakes to the questions "What is health? And how do my behaviors keep me healthy?" have never seemed higher. A staggering 45% of the United States population (equivalent to 133 million people) is burdened by at least one chronic illness (Raghupathi & Raghupathi, 2018). Can behavioral patterns lessen the burden of chronic disease? Can overall health be buoyed by lifestyle and demographic patterns? What role does preventative health play in this landscape?

These questions urged the authors to discern risk factors that might lurk beneath common chronic conditions, particularly when some causative elements evade current understanding. By harnessing the extensive data within the 2015 Behavioral Risk Factor Surveillance System (BRFSS), an expansive phone-based[1] survey overseen by the Centers for Disease Control and Prevention (CDC), an avenue emerges for uncovering correlations and patterns that cast light on the intricate interplay of behaviors and health outcomes. This, in turn, can inform the development of effective public health strategies and interventions. The BRFSS dataset, encompassing 375,059 complete interviews and an extensive array of 310 variables, is meticulously organized into distinct modules addressing various health concerns, such as women's and men's health, diabetes, arthritis, memory impairment, asthma, and demographic information. This comprehensive survey design holds immense potential to draw connections between behavioral patterns and incidence of disease, yielding insights that can enhance the well-being of millions affected by chronic health conditions.

Literature Review

This study builds on previous works examining the intersection of behavioral factors and chronic health. Conner & Norman (2002) lay out a comprehensive framework for understanding key health and cognitive behaviors, defined as smoking, diet, exercise, health screenings, sexual behaviors, and alcohol use. The authors use this framework to understand contributors to general morbidity and mortality.

Megari (2013) approaches these broad questions from the reverse lens, examining the impact of health-related quality of life on physical, psychological, and social functioning. Megari adeptly summarizes the pertinent literature on this reverse relationship, establishing the conceptual model from which to statistically analyze chronic disease's impact on lifestyle.

While the work from Conner & Norman and Megari examine the broad strokes of behavior and disease, other pieces seek to explain specific conditions' behavioral influences. For instance, Kontogianni, Farmaki, et al (2010) examine eating behaviors and obesity in children and adolescents, centering their population of interest in Greece. Here, the authors seek to frame a principal component analysis to classify and simplify the multitude of variables available to explain this relationship.

The literature abounds with studies reliant on the vast resource of data stemming from the annual BRFSS. Raghupathi & Raghupathi (2018) explore relationships between CDC-collected data (much of the variables coincide with data collected through the BRFSS) regarding behavioral, demographic and socioeconomic factors and incidences of chronic disease. Raghupathi & Raghupathi seek to expound on the Conner & Norman framework as well as other works by contributing a means to conduct exploratory data analysis and visualization, in order to bring rigor to the study of these relationships. In a similar vein, Xie, Nikolayeva, et al (2019) build a model of predictors of Type 2 diabetes, relying on the 2014 BRFSS for their training set. Xie, Nikolayeva, et al were instrumental in determining how to conduct our own exploration of diabetes predictors.

With strong conceptual frameworks guiding the overarching space of behavioral patterns, preventative health, and chronic disease, as well as proposed statistical models explaining relationships between specific diseases and behaviors, this author group seeks to examine research questions covering the gamut of relationships between preventative health, chronic health issues, and lifestyle and personal behaviors. Specifically, this author group seeks to understand how chronic health conditions affect quality of life and feeling emotionally supported, how diabetes risk factors relate and can be grouped into meaningful latent constructs, how factors such as physical activity, education, health conditions, and employment status impact an individual's overall health, and what clusters of behavior women exhibit when it comes to preventative health. This author group also seeks to make sense of the myriad variables contributing to this broad space and undertook a principal component analysis to attempt to reduce dimensionality in independent variables that explain the underlying variability among the population.

Methods

This study employed five statistical techniques on various subsets of the study to attempt to answer the five research questions. All statistical techniques were executed in RStudio and relied on common R statistical packages.

The first research question's study, using the canonical correlation analysis (CCA) technique, was performed on two groups of variables: 10 variables selected that represented chronic illnesses and two variables that represented quality of life. A total of 441,456 responses were

used, although some were rendered to zero to represent a non-answer. The CCA technique attempted to identify the relationship between chronic illness and life satisfaction and served to provide a model for dimensionality reduction within the broader study.

This group's second research question was answered through a confirmatory factor analysis (CFA), with a contributing initial deployment of Exploratory Factor Analysis (EFA). These techniques were used to understand how the risk factors for diabetes can be grouped into meaningful latent constructs[AD3] . The techniques helped confirm the structure and relationship of observed variables and assisted in understanding the underlying factors that may increase risk of developing diabetes. To perform CFA, we inspected the structure and summary of the dataset and identified missing values, substituting missing values with their respective mean values. Following this, we checked the dataset for infinite values, variables with zero standard deviation, and other potential inconsistencies, ensuring they were addressed before proceeding. With the cleaned data, we first conducted an exploratory factor analysis, represented by the `scree` plot in Figure 1. Building on this, a Confirmatory Factor Analysis (CFA) was designed with three latent variables, namely: 'Lifestyle', 'Medical History', and 'Socioeconomic Status', each capturing different sets of observed variables.

Our third research question was modeled through multiple regression. Here, our goal was understanding the impact of different factors on an individual's general health. Regression illustrated how different variables together influence the overall health of an individual, proposed a model of the relative importance of each factor, and predicted how changes in these factors might affect a person's general well-being. We started with preprocessing steps on all available variables in the study, including handling missing values and selecting relevant columns for further analysis. Then, we visualized the data to understand the distribution and correlation among variables. Selected correlation is depicted in Figure 2.

Initial feature engineering alluded to multicollinearity in the model, which was addressed by selecting variables that are not highly correlated with each other. The variance inflation factor (VIF) was used to confirm that the multicollinearity issue was resolved.

Below figures show before and after removing multicollinearity[AD4] .

Hierarchical agglomerative clustering (HAC) was deployed in answer to the fourth research question, seeking to find subsets of the general population of female adults in order to understand behavioral characteristics of the groups. The HAC technique was used over k-means to distinguish how many clusters were appropriate for the subset. The dataset was first filtered on sex = 'Female' (SEX in dataset), HPVTEST = 1 or HPVTEST = 2 ('yes' or 'no' result), and HADPAP2 = 1 or HADPAP2 = 2 ('yes' or 'no' result). 34 variables pertaining to both women's

health and lifestyle/demographics were retained and analyzed on the resulting 21,632 observations. Variables were normalized using min-max, z-score, or decimal scaling, depending on the relevant variable; most variables were on a scale of 0-13 before beginning the exercise, so limited scaling was needed. The variables were represented by categorical and quantitative metrics; a complete list of variables and their descriptions used for clustering can be found in Appendix X. Variables were mixed, with binary and continuous variables equally represented. Interval and ordinal variables were also present in the study variables. Nine combinations of distance and linkage functions were tested; Euclidean distance and Ward-D linkage produced the least skewed dendrogram, presenting four, six, or eight clusters as optimal. A high degree of noise was observed when traversing through the dendrogram to leaf clusters. (See Figure 4).
Principal Component Analysis (PCA) was used on the dataset's quantitative values, some 48 potential variables. Variables were assessed for completeness (any variable with 11,000 or more incomplete answers, usually represented by some coding containing 7s or 9s) were removed from the set. The remaining 14 variables handled missing values by coding to the median. A scree plot aided development of principal components of interest; any principal component with an eigenvalue >= 1 was retained for analysis (see Figure 5).

Discussion and Results

The CCA used to define the relationship between chronic illness and a combination variate of emotional support and life satisfaction resulted in two dimensions, with the first achieving statistical significance at the 0.05 level. The Bartlett's Chi-Squared Test is shown in Figure 6.

The shared variance from the first canonical correlation (representing chronic illnesses) is only 0.064%. Based on the structural loadings shown in Figure 7, the major influencer in the chronic illness variables (listed as x variables in the figure) was depression. This suggests that for every unit increase in the depression variable, the first variate increases by 0.765, which suggests a positive relationship. For the emotional support and life satisfaction variables (y variables), both variables are major influencers. This suggests that for every unit increase in both the emotional support and life satisfaction variables, the first variate decreases by -0.94 and -0.96, respectively. This indicates a strongly negative relationship between the variables and the variate. The low shared variance given by the first variate suggests that this grouping of variables is too broad to interpret valuable information from the relationship between these groups of variables.

Variable Name	Description	Loadings Value
	X Variables (Chronic Illness)	
CVDFRIN4	Heart Attack	-0.09105347
CVDCRHD4	Angina/coronary heart disease	-0.06596601
CVDSTRK3	Stroke	-0.3619894
ASTHMA3	Asthma	-0.10455287
CHCSCNCR	Skin cancer	-0.20076757
CHCOCNCR	Other types of cancer	0.03710732
CHCCOPD1	Chronic obstructive pulmonary disease (COPD)	-0.26021806
HAVARTH3	Any form of arthritis	-0.1696095
ADDEPEV2	Depression	0.76515678

DIABETE3	Diabetes	-0.23846343
Y Variables (Emotional Support, Life Satisfaction)		
EMTSUPRT	Social/emotional support frequency	-0.9399837
LSATISFY	Life satisfaction	-0.9626035

Figure 7: Structural Loadings of CCA's X and Y Variables

A CFA analysis was conducted in answer to the second research question, intended to group the diabetes risk factors in meaning full latent constructs. Scree plot demonstrates an ‘elbow’ at three factors (see Figure 1, above in ‘Methods’ section). Factor analysis extracted three factors via a varimax rotation. Based on the grouped variables’ characteristics, the three factors were named:

- Factor1 ("**Lifestyle**"): This has strong positive loadings for PhysActivity (0.880), Fruits (0.832), Veggies (0.860), and HvyAlcoholConsump (0.607). This factor relates to physical activity and dietary habits.
- Factor2 ("**Medical History**"): It has notable loadings with Diabetes (0.417), HighBP (0.288), HeartDiseaseorAttack (0.386), and GenHlth (-0.627). This factor captures aspects related to one's medical history and overall health status.
- Factor3 ("**Socioeconomic Status**"): This has loadings associated with NoDocbcCost (0.235), MentHlth (0.466), and Income (0.119). It seems to capture aspects related to healthcare access and mental health likely influenced by socioeconomic conditions.

Figure 8: Latent Variables from CFA

After fitting our CFA model to the data, we extracted fit statistics and summary metrics. To visually assess how well our model performed, we plotted the observed versus the fitted covariances, providing an illustrative representation of the model's goodness-of-fit.

- Goodness of Fit: The Comparative Fit Index (CFI) was 0.960 and Tucker-Lewis Index (TLI) value was 0.948, both close to 1, indicating a good fit. The RMSEA value was 0.046, further suggesting that the model has a decent fit with the data. See Figure 9 for the Goodness of Fit Plot.
- Factor Loadings: The variables loaded strongly on their respective latent factors. For instance, PhysActivity had a loading of 0.880 on the 'Lifestyle' factor, Diabetes had a loading of 0.417 on 'Medical History', and NoDocbcCost had a loading of 0.235 on 'Socioeconomic Status', confirming their associations.
- Covariances: There exists a demonstrable relationship between 'Lifestyle' (value: 0.049) and 'Medical History' as well as 'Lifestyle' (value: 0.103) and 'Socioeconomic Status'. The negative covariance (value: -0.471) between 'Medical History' and 'Socioeconomic Status' implies an inverse relationship.

Research question three was approached through a multiple regression. Through the stepwise selection process, we identified significant features that have strong relationships with 'GENHLTH.'

The resulting model exhibited a good fit, explaining approximately 57% of the variance in 'GENHLTH.' Features such as 'QLACTLM2,' '_MICHD,' '_RFHLTH,' 'DIABETE3,' 'PHYSHLTH,' 'EMPLOY1,' 'BPHIGH4,' and 'EDUCA' were identified as influential in determining general health. This shows that general health is influenced by various aspects like physical health, employment status, education, diabetes, and high blood pressure, among others. See Figure X for a full

summary of the regression model. The Mean Squared Error (MSE) of the model is 0.4311 on the test data, providing further insight into the predictive ability of the model.

Initial clustering attempts hinted at some four, six, or eight clusters as a potential answer to research question four. Supplementary principal component analysis on these cluster models was used to test and visualize optimal cluster numbers. Ultimately, four clusters proved optimal, with the most definition between clusters. The most overlap occurred between clusters 1 & 3 and clusters 2 & 4. Clusters were analyzed and named based on the most notable driving characteristics, **_AGEG5YR and EMPLOY1.** See Figure 11 for cluster visualization and Figure 12 for name and description of each cluster.

Cluster Name	Cluster Number	# & % of Respondents in Cluster
Older Retirees	3	3,436 (15.88%)
Younger Retirees	1	6,779 (31.34%)
Mid-Life Workers	2	8,542 (52.78%)
Young Professionals	4	2,875 (13.29%)

Figure 12: Cluster Names and Descriptive Statistics

Though age group and working status drove most cluster behavior, other notable findings include:

- Older retirees have gone the longest since their last mammogram or Pap smear, weigh a bit less, and have the lowest estimated functional capacity;
- Younger retirees struggle the most with weight (this grouping collected the heaviest women), and were most likely to have had poor health stop them from doing activities in the past 30 days;
- Mid-Life Workers enjoyed the strongest mental health in the past 30 days and were likely to have exercised in the past 30 days; and
- Young Professionals were least likely to have had a mammogram, HPV test, or Pap smear, but were the most likely to have been tested for HIV. They were most likely to have children residing in the household, least likely to have smoked 100 cigarettes in their lifetime, and enjoyed the highest functional capacity.

Appendix X provides density plots for each of the four clusters over 16 variables of interest.

The first four principal components, as evidenced by the scree plot in Figure 5 (see 'Methods' section, above) are defined by their contributing variables:

- PC1, comprised of Physical Health, Mental Health, Height, and Weight;
- PC2, comprised of Green-colored vegetables, orange-colored vegetables, and other vegetables;
- PC3, comprised of number of children and number of adults; and
- PC4, mental health and physical health (reprised from PC1[AD5]).

All analyses performed followed one clear path: many factors impact health aspects. The insights derived from this analysis could be valuable in identifying key health determinants and informing

healthcare policies and interventions[AD6] . Educating people on how they can build 'healthy habits', and understanding more deeply what those habits *are*, may keep people healthy as they age. It should be noted that, since the data was collected via telephone survey, the subject pool was comprised of people who were available at home at the time of the call and were willing to answer personal health questions. Therefore, the broader applicability of results may be limited in the general American adult population. A further limitation is the over-sampling of older adults. According to the 2020 US census, approximately 1 in 6 (16.67%) of American adults were over the age of 65; this survey's complete interviews featured 35.87% of respondents in that category.

Future work can build on this author group's broad study to build a deeper understanding on how behavioral risk factors can affect health conditions. The exploration of research question 1 through CCA uncovered an opportunity to delve deeper into understanding of individual illnesses or related groups of illnesses to better understand the effect of chronic illness on overall emotional support and life satisfaction. This would give a better understanding on how resources in communities should be distributed to patients who report an overall lower amount of emotional support and life satisfaction. It would also help to better understand the 18-45 population, particularly regarding female behavior, and understand the clusters of behavior in females under the lenses of race, gender identity, and sexual orientation.

Conclusion

The research reveals that chronic illness can have a detrimental effect on life, while opening an array of future research options to further hone these initial findings. CFA showed that diabetes risk factors can be grouped into three factors: lifestyle, medical history, and socioeconomic status. Regression analysis showed that general health is most influenced by physical health, employment status, education, diabetes, and high blood pressure. The study also shed light on what behaviors women exhibit in lifestyle and preventative health screening. Lastly, the variables most crucial in PCA's results are someone's mental and physical health, as these are key to two influential principal components.

In conclusion, it was determined that many different risk factors influence different aspects of a person's health, and health in turn influences a person's quality of life. It is important to conduct these kinds of studies to better educate the general population on actions they can take to prevent avoidable illness and disease.

References and Appendices

References

1. Centers for Disease Control and Prevention (2015). *Behavioral Risk Factor Surveillance System*. Centers for Disease Control and Prevention. https://www.cdc.gov/brfss/annual_data/annual_2015.html
2. Conner, M. & Norman, P. (2002). Health Behaviors. *Comprehensive Clinical Psychology, 8.* 10.1016/B978-0-08-097086-8.14154-6

3. Kontogianni, et al. (2010). Journal of the American Dietetic Association. *Associations between Lifestyle Patterns and Body Mass Index in a Sample of Greek Children and Adolescents.* https://doi.org/10.1016/j.jada.2009.10.035
4. Megari, K. (2013). Quality of life in chronic disease patients. *Health Psychology Research*, *1*(3), 27. https://doi.org/10.4081/hpr.2013.e27
5. Raghupathi, W., & Raghupathi, V. (2018). An empirical study of chronic diseases in the United States: A visual analytics approach to Public Health. *International Journal of Environmental Research and Public Health*, *15*(3), 431. https://doi.org/10.3390/ijerph15030431
6. United States Census Bureau (2023, June 9). 2020 Census Demographic Profile. Retrieved August 23, 2023, from https://www.census.gov/data/tables/2023/dec/2020-census-demographic-profile.html
7. Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building risk prediction models for type 2 diabetes using machine learning techniques. *Preventing Chronic Disease*, *16*. https://doi.org/10.5888/pcd16.190109

Appendix C - Clustering Variable Codes and Description

Variable Code	Variable Description and Original Scale
AGEGROUP	Categorical variable assigning age group (5-year groupings)
ACTIVLVL	Categorical variable determining if respondent has performed moderate, vigorous, or no physical activity in the past 30 days. Aggregated and calculated from several survey answers, first by BRFSS staff, then by author
AEROREC S	Calculated binary variable determining if respondent has met aerobic exercise requirements in past 30 days. Aggregated and calculated from several survey answers first by BRFSS staff, then by author.
CAREGIV1	Binary variable - 'have you provided regular care or assistance to someone with a disability in the last 30 days?'
CHECKUP1	Time since last checkup within last X years, maximum of 6 (where 6 is 'more than 5 years ago')
CHILDREN	Number of children in household
EDUCA	Categorical variable representing highest level of schooling completed
EMPLOY1	Categorical variable representing how wages were earned, including out of work, disabled, or retired responses
FLUSHOT6	Binary variable - 'have you received a flu vaccine in the last 12 months?'
FRUITNORM	Normalized variable derived from calculated quantitative variable derived from several survey answers by BRFSS staff indicating the average number of fruits consumed per day. Normalized via min-max scaling.
FUNCCAP	Estimated functional capacity (calculated by BRFSS staff)
GENHLTH	Categorical variable with 5 levels indicating overall health
HADHPVORPAP	Combination of HPVTEST and HADPAP2, indicating if respondent has had either a Pap smear or stand-alone HPV test
HADHYST2	Binary - 'have you had a hysterectomy?'
HADMAM	Binary variable - 'have you ever had a mammogram?'
HEIGHTNORM	Normalized height in inches; height normalized through z-score scaling
HIVTST6	Binary variable - 'have you ever been tested for HIV?'
HLTHPLN1	Binary variable - 'do you have any kind of health care coverage?'
HOWLONG	Time since last mammogram within last X years, maximum of 6 (where 6 is 'more than 5 years ago')
INCOMG	Categorical variable representing annual household income brackets; maximum is $50k+

INTERNET	Binary variable - ‘have you used the internet in the past 30 days?’
LASTPAP2	Time since last Pap smear within last X years, maximum of 6 (where 6 is ‘more than 5 years ago’)
MAXV02NORM	Normalized variable derived from calculated quantitative variable describing estimated maximum oxygen consumption. Calculated by BRFSS staff, normalized va min-max scaling
MEDCOST	Binary variable - ‘was there a time in the past 12 months when you needed to see a doctor but could not because of cost?’
NORMALC	Normalized variable indicating X days in last 30 days where alcohol was consumed; 3 is max. Normalized via decimal scaling.
NORMEXER	Normalized variable indicating X days in last 30 days where mental health was not good; 3 is max. Normalized via decimal scaling.
NORMMENT	Normalized variable indicating X days in last 30 days where mental health was not good; 3 is max. Normalized via decimal scaling.
NORMPHYS	Normalized variable indicating X days in last 30 days where physical health was not good; 3 is max. Normalized via decimal scaling.
NORMPOOR	Normalized variable indicating X days in last 30 days where poor health detracted from doing activities; 3 is max. Normalized via decimal scaling.
NORMSTREN	Normalized variable indicating X days in last 30 days where mental health was not good; 3 is max. Normalized via decimal scaling.
SEATBELT	Categorical variable classifying frequency of seatbelt use when riding in a vehicle
SMOKE100	Binary variable - ‘have you smoked at least 100 cigarettes/5 packs in your entire life?’
VEGNORM	Normalized variable derived from calculated quantitative variable derived from several survey answers by BRFSS staff indicating the average number of vegetables consumed per day. Normalized via min-max scaling.
WEIGHTNORM	Normalized weight in pounds; weight normalized through z-score scaling

Appendix D: Scree Plots for Principal Component Analysis

Appendix E: Full PCA Plot Set

Appendix F: PCA Summary and Output

Appendix G: PCA Variable Inputs

Variable Code	Variable Description and Original Scale
ALCDAY5	During the past 30 days, how many days per week or per month did you have at least one drink of any alcoholic beverage such as beer, wine, a malt beverage or liquor?
CHILDREN	How many children less than 18 years of age live in your household?
FRUIT1	During the past month, not counting juice, how many times per day, week, or month did you eat fruit? Count fresh, frozen, or canned fruit.
FRUITJU1	During the past month, how many times per day, week or month did you drink 100 percent PURE fruit juices? Do not include fruit-flavored drinks with added sugar or fruit juice you made at home and added sugar to. Only include 100 percent juice.
FVBEANS	During the past month, how many times per day, week, or month did you eat cooked or canned beans, such as refried, baked, black, garbanzo beans, beans in soup, soybeans, edamame, tofu or lentils. Do NOT include long green beans.
FVGREEN	During the past month, how many times per day, week, or month did you eat dark green vegetables for example broccoli or dark leafy greens including romaine, chard, collard greens or spinach?
FVORANG	During the past month, how many times per day, week, or month did you eat orange-colored vegetables such as sweet potatoes, pumpkin, winter squash, or carrots?
HEIGHT3	About how tall are you without shoes?
MENTHLTH	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?
NUMADULT	Number of Adults in Household
PHYSHLTH	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?
STRENGTH	During the past month, how many times per week or per month did you do physical activities or exercises to STRENGTHEN your muscles? [Do NOT count aerobic activities like walking, running, or bicycling. Count activities using your own body weight like yoga, sit-ups or push-ups and those using weight machines, free weights, or elastic bands.]
VEGETAB1	Not counting what you just told me about, during the past month, about how many times per day, week, or month did you eat OTHER vegetables? Examples of other vegetables include tomatoes, tomato juice or V-8 juice, corn, eggplant, peas, lettuce, cabbage, and white potatoes that are not fried such as baked or mashed potatoes.
WEIGHT2	About how much do you weigh without shoes?

[1] 2015 is the first year since the study's inception that both cellular and landline telephone numbers were used to collect the sample.