Behavioral Cluster Analysis of Food Consumption: Associations with Comparatively Healthier Food Choices

J Dickerson, M Smith, R Rahn, M Ory

Citation

Abstract
Objectives – To identify demographic, behavioral, and health factors associated with food choices among community-dwelling adults.
Methods – A cross-sectional health assessment was used for the study. K-means cluster analysis identified natural groupings of individuals reporting similar food choices among four categories: fruits, vegetables, sugared beverages, and fast food. Multinomial logistic regression identified differences in comparatively healthier food choices between the clusters.
Results – Six unique cluster profiles of eating habits and food consumption were identified. Demographic, behavioral, and health factors were associated with variations in healthy food consumption. Compared to those in the cluster representing the least healthiest food choices, members of the cluster exhibiting the most healthy food choices were less likely to report physical illness (OR = 0.97, p<0.001). These cluster members were also 2.34 times as likely to be female (OR = 2.34, p<0.001), 2.78 times as likely to have earned more than a high school education (OR = 2.78, p<0.001), and 1.13 times as likely to spend more days per week engaging in at least 10 minutes of moderate physical activity (OR = 1.13, p<0.001) when compared to their counterparts in the cluster representing the least healthiest food choices.
Conclusion – Comparatively poorer food choices of community members remains an essential and modifiable indicator of existing health status. Interventions to achieve healthy food choices should utilize a multi-level approach that emphasizes the role of important covariates.

INTRODUCTION
Promoting the consumption of healthier foods (e.g., eating more fruits and vegetables and reducing sugared beverage and fast food consumption) is a focus of public health efforts that benefits overall community health (1, 2). Increased consumption of sugared beverages is associated with overweight and obesity in adults (3) and children (4). High levels of sugared beverage consumption in children as young as age five has been shown to negatively affect nutritional consumption across adolescence (5). Similarly, fast food consumption (which likely facilitates sugared beverage consumption) is also related to poor dietary consumption and propensity for overweight and obesity among both adults (6) and children (7).

Excessive sugared beverage and fast food consumption can result in negative health consequences, especially when such diets reduce or replace fruit and vegetable consumption. A diet low in fruit and vegetable consumption is associated with higher risk of ischemic stroke, lung cancer, stomach cancer, colorectal cancer, and cancers of the mouth, pharynx and esophagus (8). It has been shown that increasing consumption of fruits and vegetables lowers the chance of developing many of these diseases for individuals as young as age 15 (9). Higher fruit consumption alone has been shown to markedly lower the risk of developing hypertension (10), one of the leading causes of mortality (8). Despite these demonstrated benefits of higher fruit and vegetable consumption, many adults consume a diet with less than optimal fruit and vegetable consumption (11).

In addition to examining self-reported health status, researchers have counseled investigators to also examine a wide range of independent variables when trying to understand diet behaviors (12). This study examines the results of a community health assessment to identify differences in demographic, behavioral, and health characteristics associated with participants’ patterns of fruit, vegetable, sugared beverage and fast food consumption. The purposes of this study are to: 1) describe the factors associated with comparatively healthier food choices in the community, and 2) identify strategies to promote better health outcomes through healthier food choices.
METHODS

BRAZOS VALLEY HEALTH ASSESSMENT (BVHA)

The 2010 BVHA (n = 3,964) was conducted and funded by the Center for Community Health Development at the Texas A&M Health Science Center, School of Rural Public Health (13). In conjunction with community health partners, a voluntary questionnaire was disseminated to assess community health status and opportunities for health improvement in the Brazos Valley, an eight county region in central Texas (13). Data were collected using a random sampling of households. The instrument was 32 pages containing items from validated sources (13) such as the Centers for Disease Control and Prevention (CDC).

SAMPLE

All 2010 BVHA respondents were eligible for inclusion (n = 3,964). However, based on non-response to some questions from a small number of respondents, the total size of the analytic sample was reduced to n = 3,844. Those not answering the questions included in the study were subjected to list-wise deletion in Stata Version 11 (StataCorp, College Station, Texas).

INDEPENDENT VARIABLES

We identified several variables to analyze as independent variables throughout our analyses described below. Age of the respondent was categorized as follows: 18 to 34, 35 to 44, 45 to 64, and 65 and over. Sex was also a categorical variable. Race/ethnicity was categorized as non-Hispanic white, African-American, Hispanic, and Other. Educational attainment was categorized as less than high school, high school, and more than high school.

Several variables were included to measure health status of the respondent and history of chronic diseases. Physical and mental health – days not well, were measured as continuous variables on a scale from zero to 30 (14). Body mass index was categorized as underweight, normal weight, overweight, and obese. Days per week performing at least ten minutes of moderate physical activity (e.g., fast walking) was measured as a continuous variable from zero to seven. Categorical variables of diagnosed chronic disease were also included for: diabetes, hypertension, high cholesterol, previous heart attack or stroke, and previous cancer.

CLUSTER METHODOLOGY

Consistent with the research questions of our study, we selected variables for cluster analysis that would create distinct profiles of eating behavior and food consumption. Respondents were asked the following questions: 1) “How many times per month, week, or day did you eat fruit? Count any kind of fruit—fresh, canned, and frozen. Do not count juices. Include fruit you ate at mealtimes and for snacks.” 2) “How many times per month, week, or day did you eat vegetables? Count any kind of vegetables – fresh, canned, and frozen. Do not count lettuce, white potatoes, rice, vegetables in mixtures such as sandwiches, omelets, casseroles, Mexican dishes, stews, soups, etc. Include vegetables you ate at mealtimes and for snacks.” 3) “How many times per month, week, or day did you drink sugar sweetened beverages? This includes fruit drinks, regular soda, sweet tea, coffee with sugar and sports or energy drinks.” 4) “How many times per month, week, or day did you eat fast food meals or snacks?” All questions were scored using the same ordinal answer scale: 0 – Never, 1 = 1 to 3 times per month, 2 = 1 to 2 times per week, 3 = 3 to 4 times per week, 4 = 5 to 6 times per week, 5 = 1 time per day, 6 = 2 times per day, 7 = 3 or more times per day. Since these responses were normally distributed, no data transformations were performed. Also, since the questions were measured on the same scale and were of the same importance when addressing the research questions of this study, the variables were not weighted prior to clustering.

K-means partitioning was selected as the clustering method since it is uniquely designed for non-hierarchical data partitioning. Because the underlying variables of the cluster analysis were expressed in ordinal terms, Euclidean distance was used as the proximity function because of its utility in clustering such data (15).

Determining the number of clusters to extract was a key consideration. We created ten different cluster models. Changes in both the within group and between group sum of squared errors (SSE) were examined as additional clusters were added. Based on guidance from established literature (16, 17), our primary goal was to minimize the within group SSE. However, analysis was also performed to ensure the most parsimonious solution.

CLUSTER VISUALIZATION

A bubble chart was used to show the results of the cluster analysis on the dimension of sugared beverages and fast food consumption versus fruits and vegetables consumption. To create this visualization, sugared beverages and fast food were combined to form an index, and fruits and vegetables
were combined to form an index. This method was appropriate since no weighting was used for the variables in this study, and the sugared beverages, fast food, fruits and vegetables variables were measured on the same scale. Based on these indices, mean scores were calculated for the entire sample (n = 3,844). These means are identified on Figure 1 as horizontal and vertical thresholds. The bubble chart was then created to represent each cluster’s position relative to the means of both the fruits and vegetables index and the sugared beverages and fast food index. The size of the bubble represents the number of cluster members.

DESCRIPTIVE STATISTICS

The independent variables identified above were analyzed to provide context to the results of the cluster analysis. Chi-square statistics were used to identify differences between each of the clusters and the categorically measured independent variables. Kruskall-Wallis statistics were used to identify differences between each of the clusters and the continuously measured independent variables. This statistic was used instead of the t-statistic due to the number of clusters in the analysis. The Bonferroni correction method was used to establish the level of statistical significance by dividing α = 0.05 by the number of comparisons in the analysis (i.e., 12). This resulted in a level of statistical significance of α = 0.004.

MULTINOMIAL LOGISTIC REGRESSION

After clusters and descriptive variables were identified, multinomial logistic regression analysis was used to predict the likelihood of a respondent being a member of a cluster based on the independent variables identified above. The dependent variable of the multinomial logistic regression analysis was cluster membership of the respondent. Independent variables were then analyzed for statistical significance and tested in a series of multinomial logistic regression models with the goal of minimizing the Bayesian Information Criterion (BIC) statistic. To avoid misinterpretation of the results, we created two multinomial logistic regression models where the only difference in independent variables was physical health days not well versus mental health days not well. This was done because of inherent correlation between these two independent variables. Based on standards for interpreting BIC statistics (18), the independent variables generating the most parsimonious model were selected for inclusion in the final multinomial logistic regression model. The referent group for all analyses was the cluster representing the least desirable food choice: comparatively low consumption of fruits and vegetables and comparatively high consumption of sugared beverages and fast food. Relative risk ratios and 95% confidence intervals were then calculated in order to define the likelihood of certain events within the model.

RESULTS

CLUSTERS OF FOOD CHOICE BEHAVIORS

Table 1 identifies the results of the process used to designate the number of clusters based on the dataset. We determined six clusters to be the most parsimonious representation of our goal to minimize within group SSE. This is because adding additional clusters was not practical given the incremental reduction in within group SSE. Figure 1 illustrates the clusters according to the food choice behaviors of their members. Cluster 5 had a membership that practiced the most comparatively desirable food choice behavior; high fruit and vegetable consumption, low sugared beverage and fast food consumption. Conversely, Cluster 1 exhibited the comparatively least desirable food choice behavior, low fruit and vegetable consumption, high sugared beverage and fast food consumption. Both clusters were very similar in terms of the size of their membership. Cluster 4 was somewhat of an outlier in the analysis. It contained a relatively small membership, and also exhibited food choice behavior that was very different than other clusters (i.e., very high sugared beverage and fast food consumption, mean fruits and vegetables consumption).

DESCRIPTIVE STATISTICS OF CLUSTERS

Table 2 reports the results of the descriptive statistics analysis of the resulting clusters of food choice behavior. The overall sample was 70.4% female and 80.0% non-Hispanic white. The mean age of respondents was 58.52 (±15.26), and the majority (56.6%) had attained more than a high school education. On average, the sample reported more than 3 days per week of performing at least 10 minutes of moderate physical activity. Diabetes was not a prevalent condition among the sample (15.0%), but nearly half of the sample reported being diagnosed with hypertension (49.5%) and high cholesterol (45.1%). A relatively small number of sample respondents had a previous heart attack or stroke (9.2%) or reported having been previously diagnosed with cancer (18.5%). With the exception of diagnosed hypertension ($X^2 = 15.223$, p = 0.009) and a previous heart attack or stroke ($X^2 = 26.903$, p = 0.185), all variables demonstrated significant differences between clusters. This is an indication of how well the clustering methodology
partitioned the data. Since our goal was to have K-means partition the data in a manner minimizing within group SSE, it was expected that independent variables would have different values by cluster.

On a cluster by cluster basis, there was wide variation among the variables. Cluster 1 contained the youngest (25.8% under age 45) and least educated (8.1% less than high school attainment) proportion of members and generally had the lowest proportion of members who had experienced chronic disease. Cluster 2 contained the highest proportion of females (75.7%) and also had the lowest proportion of members with hypertension (45.4%) or previous diagnoses of heart attack or stroke (7.0%). Cluster 3 contained the highest proportion of non-Hispanic white members (89.4%), and also the highest proportion of members with diabetes (19.8%) and high cholesterol (51.4%). Cluster 4 contained the highest proportion of males (33.7%) and also the highest proportion of members who had experienced a previous heart attack or stroke (11.8%). Cluster 5 contained the highest proportion of members who attained more than a high school education (68.3%) as well as the most members who performed at least 10 minutes of moderate physical activity per day (3.78 ± 2.45). Cluster 5 also contained the highest proportion of members with previously diagnosed cancer (21.8%). Finally, Cluster 6 contained the highest proportion of members who were obese (41.4%) as well as the least members who performed at least 10 minutes of moderate physical activity per day (2.90 ± 2.44).

MULTINOMIAL LOGISTIC REGRESSION – PHYSICAL HEALTH

Table 3 examines the results of the multinomial logistic regression model used to associate cluster membership with the independent variables including number of the last 30 days where physical health was not good. Cluster 1 served as the referent group for all comparisons in this model (i.e., the cluster exhibiting the least desirable food choice behavior, low fruits and vegetables consumption, high sugared beverages and fast food consumption).

Compared to Cluster 1, members of Cluster 2 were less likely to report poor physical health (RRR = 0.99, p = 0.023). Members of Cluster 2 were also more likely to be older (RRR = 1.01, p < 0.001), female (RRR = 1.94, p < 0.001), better educated (RRR = 1.61, p < 0.001), and to perform more days per week of at least 10 minutes of moderate physical activity (RRR = 1.10, p < 0.001).

Compared to Cluster 1, members of Cluster 3 were less likely to report poor physical health (RRR = 0.98, p < 0.001). Members of Cluster 3 were also more likely to be older (RRR = 1.03, p < 0.001), female (RRR = 1.63, p < 0.001), better educated (RRR = 2.21, p < 0.001), and to have been diagnosed with diabetes (RRR = 0.34, p < 0.001).

Compared to Cluster 1, members of Cluster 4 were more likely to have been diagnosed with diabetes (RRR = 0.53, p = 0.006).

Compared to Cluster 1, members of Cluster 5 were less likely to report poor physical health (RRR = 0.97, p < 0.001). Members of Cluster 5 were also more likely to be older (RRR = 1.03, p < 0.001), female (RRR = 2.34, p < 0.001), better educated (RRR = 2.78, p < 0.001) perform more days per week of at least 10 minutes of moderate physical activity (RRR = 1.13, p < 0.001) and to have been diagnosed with diabetes (RRR = 0.33, p < 0.001).

Compared to Cluster 1, members of Cluster 6 were more likely to be older (RRR = 1.01, p < 0.001), better educated (RRR = 1.43, p < 0.001) and to have been diagnosed with diabetes (RRR = 0.42, p < 0.001).

MULTINOMIAL LOGISTIC REGRESSION – MENTAL HEALTH

Table 4 examines the results of the multinomial logistic regression model used to associate cluster membership with the independent variables including number of the last 30 days where mental health was not good. Cluster 1 served as the referent group for all comparisons in this model (i.e., the cluster exhibiting the least desirable food choice behavior, low fruits and vegetables consumption, high sugared beverages and fast food consumption).

Compared to Cluster 1, members of Cluster 2 were more likely to be older (RRR = 1.01, p = 0.002), female (RRR = 1.94, p < 0.001), better educated (RRR = 1.64, p < 0.001), and to perform more days per week of at least 10 minutes of moderate physical activity (RRR = 1.10, p < 0.001).

Compared to Cluster 1, members of Cluster 3 were less likely to report poor mental health (RRR = 0.97, p = 0.001). Members of Cluster 3 were also more likely to be older (RRR = 1.03, p < 0.001), female (RRR = 1.65, p < 0.001), better educated (RRR = 2.25, p < 0.001), perform more days per week of at least 10 minutes of moderate physical activity (RRR = 1.05, p = 0.039), and to have been diagnosed with diabetes (RRR = 0.35, p < 0.001).
Compared to Cluster 1, members of Cluster 4 were more likely to have been diagnosed with diabetes (RRR = 0.52, p = 0.004).

Compared to Cluster 1, members of Cluster 5 were less likely to report poor mental health (RRR = 0.96, p < 0.001). Members of Cluster 5 were also more likely to be older (RRR = 1.03, p < 0.001), female (RRR = 2.36, p < 0.001), better educated (RRR = 2.81, p < 0.001) perform more days per week of at least 10 minutes of moderate physical activity (RRR = 1.14, p < 0.001) and to have been diagnosed with diabetes (RRR = 0.34, p < 0.001).

Compared to Cluster 1, members of Cluster 6 were more likely to be older (RRR = 1.01, p < 0.001), better educated (RRR = 1.43, p < 0.001) and to have been diagnosed with diabetes (RRR = 0.43, p < 0.001).

**Figure 1**

**Table 1. Cluster Identification Statistics**

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>Within Group Sum of Squared Errors (GSE)</th>
<th>Between Group Sum of Squared Errors (BSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Cluster Model</td>
<td>41,647.75</td>
<td>19,954.12</td>
</tr>
<tr>
<td>3 Cluster Model</td>
<td>30,074.00</td>
<td>36,744.00</td>
</tr>
<tr>
<td>4 Cluster Model</td>
<td>20,126.10</td>
<td>41,320.79</td>
</tr>
<tr>
<td>5 Cluster Model</td>
<td>18,877.57</td>
<td>42,750.02</td>
</tr>
<tr>
<td>6 Cluster Model</td>
<td>16,844.95</td>
<td>44,782.59</td>
</tr>
<tr>
<td>7 Cluster Model</td>
<td>15,852.23</td>
<td>45,960.98</td>
</tr>
<tr>
<td>8 Cluster Model</td>
<td>15,217.80</td>
<td>46,641.85</td>
</tr>
<tr>
<td>9 Cluster Model</td>
<td>13,886.50</td>
<td>48,088.00</td>
</tr>
</tbody>
</table>

Figure 2

**Figure 3**

**Figure 1 - Cluster Composition Matrix**

Size of bubble indicates the number of observations within the cluster.
DISCUSSION

BEHAVIORAL CLUSTERING CAN BENEFIT CROSS-SECTIONAL FOOD CHOICE ANALYSIS

Many cross-sectional studies seek to identify specific populations or at-risk groups when examining health behavior. However, more often than not, the group is predetermined before any health behaviors are studied. Cluster methods represent a complementary tool in validating such approaches and may be used in place of these methods when the behavior under study is complex and the group practicing such behavior is not readily identifiable. At a community level, clustering is useful since the aggregation of behaviors into cluster membership takes place prior to any partitioning based on individual characteristics such as sex, race, or ethnicity. As a result, any pre-conceived notion that...
would lead the researcher to select the group before examining the behavior can be thoroughly examined and validated before data is excluded from the analysis.

Finally, behavioral clustering may represent an opportunity seldom afforded to researchers of cross-sectional data, the ability to study data in the most naturalistic sense possible. By utilizing clustering methods when examining behaviors, pre-conceived beliefs about the way cross-sectional samples should be selected are trumped by a more naturalistic approach in setting parameters for cluster models to select cases that most naturally exhibit the behaviors in question. This is a powerful opportunity for cross-sectional researchers to better contextualize their findings.

FOOD CHOICE BEHAVIOR ASSOCIATED WITH A MYRIAD OF FACTORS

Goodwin (12) counseled researchers to examine factors other than health status when attempting to understand dietary behaviors among adolescents. While our results are not confined to one population, this counsel appeared justified in our findings. As described above, Cluster 1 represented the poorest food choice behaviors in our study. Conversely, Cluster 5 represented the best food choice behaviors in our study. When examining differences between these clusters in Table 3 and Table 4, Goodwin’s (12) recommendation seems justified. Table 3 illustrates those in Cluster 5 were somewhat less likely to experience physical health days not well than those in Cluster 1. However, this was a relatively inconsequential finding considering those in Cluster 5 were substantially more likely to be female, to have attained more education, and to have performed more days per week of at least 10 minutes of moderate physical activity. The results in Table 4 are nearly identical with the exception that physical health days not well are replaced by mental health days not well. Again, those in Cluster 5 were only slightly more likely to report fewer mental health days not well compared to their Cluster 1 counterparts.

These findings suggest that being mentally and physically healthy relative to peers in the community is important in developing positive food choice behaviors. However, the picture would be incomplete without considering the important associations of demographic factors such as sex and educational attainment, as well as variables measuring physical activity.

MULTI-LEVEL INTERVENTIONS ARE NEEDED TO ADDRESS FOOD CHOICE BEHAVIORS

Our findings provide support for improving food choice behaviors through a multi-level intervention approach. Single focus interventions do not seem appropriate since negative food choice behaviors affect a multitude of community subgroups. In addition to demographic associations with the consumption of healthier foods (i.e., female and more highly educated members of the community), our findings suggest community members are associated with healthier food choice behaviors based on how they feel (i.e., physical and mental health status along with whether the community member has been diagnosed with diabetes) and how they behave (i.e., level of physical activity). This suggests the need for an intervention that engages community members on multiple dimensions. This is similar to recent findings which have argued for addressing nutrition through individual, social and environmental variables simultaneously (19).

There are several examples of multi-level intervention approaches to improve community nutrition. The Central California Regional Obesity Prevention Program has experienced success using community-based strategies to change the local food environment, motivate neighborhoods to promote physical activity, and take advantage of local policymaking resources to effect positive health behavior change (20). Multi-level interventions in diverse settings such as the workplace (21) and religious organizations (22) have also been successful in promoting consumption of fruits and vegetables. Our work further substantiates the need to work creatively within the community to leverage opportunities for multi-level interventions to address food choice behavior. Based on our findings, we believe interventions accounting for sex, education, physical activity, and health history of the community member would be appropriately matched to begin the process of moving the community toward better food choices.

LIMITATIONS

While our study is innovative in its usage of clustering methods to describe food choice behavior, it is not without its limitations. First, our study was conducted in a relatively small geographic area in Texas. As such, the ability to generalize the findings is limited. Second, our data is self-reported and from a cross-sectional sample. As a result, the findings of our study are susceptible to criticisms of validity. Third, we relied on multinomial logistic regression to define...
Behavioral Cluster Analysis of Food Consumption: Associations with Comparatively Healthier Food Choices

the strength of relationships between our independent variables and cluster membership. However, there is no equivalent of the coefficient of determination for multinomial logistic regression analysis. Thus, we do not know how much of the variation in cluster membership was explained by our independent variables on a cumulative basis.

CONCLUSION
This study utilized the power of clustering methodology to best reflect the naturalistic state of the community regarding food choice behaviors with no a priori assumptions about which populations were likely to exhibit certain food choice behaviors. This is important because it allows researchers to understand how the community naturally operates and how diverse collections of community members come together around certain behavioral themes, even the theme of food choice. This is the knowledge that fosters the creation of multi-level community-based interventions that can directly address the many facets of poor food choice behavior. This approach should be a key consideration for any practitioner looking for ways to understand and address community nutrition needs.

References
Author Information

Justin B. Dickerson, MBA
PhD Student, Health Services Research, Department of Health Policy & Management School of Rural Public Health Texas A&M Health Science Center 1266 TAMU College Station, Texas 77843 USA

Matthew Lee Smith, PhD, MPH, CHES
Assistant Professor, Department of Health Behavior and Promotion College of Public Health The University of Georgia 315 Ramsey Center Athens, Georgia 30602 USA

Rhonda Rahn, MS, CHES
Health Education Coordinator, Department of Student Health Services Texas A&M University Beutel Health Center College Station, Texas 77843 USA

Marcia G. Ory, PhD, MPH
Regents Professor, Department of Social & Behavioral Health School of Rural Public Health Texas A&M Health Science Center 1266 TAMU College Station, Texas 77843 USA