Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Wang et al., J Biom Biostat 2017, 8:4
DOI: 10.4172/2155-6180.1000364

Review Article OMICS International

# Principal Component Regression Analysis of Nutrition Factors and Physical Activities with Diabetes

**Ke-Sheng Wang[1]*, Ying Liu[1], Xin Xie[2], Shaoqing Gong[1], Chun Xu[3] and Zhanxin Sha[4]**

[1]*Department of Biostatistics and Epidemiology, College of Public Health, East Tennessee State University, Johnson City, TN 37614, USA*
[2]*Department of Economics and Finance, College of Business and Technology, East Tennessee State University, Johnson City, TN 37614, USA*
[3]*Department of Health and Biomedical Sciences, College of Health Affairs, University of Texas Rio Grande Valley, Brownsville, TX 78520, USA*
[4]*School of Kinesiology, College of Health, University of Southern Mississippi, Hattiesburg, MS 39406, USA*

## Abstract

The associations of nutrition factors and physical activities with adult diabetes are inconsistent; while most of these factors are inter correlated. The aims of this study are to overcome the disturbance of the multicollinearity of the risk factors and examine the associations of these factors with diabetes using the principal component analysis (PCA) and regression analysis with principal component scores (PCS). Totally, 659 adults with diabetes and 2827 non-diabetic were selected from the 2012 Health Information National Trends Survey (HINTS 4, Cycle 2). PCA was utilized to deal with multicollinearity of the risk factors. Weighted univariate and multiple logistic regression analyses were used to estimate the associations of potential factors and PCS with diabetes. The odds ratios (ORs) with 95% confidence intervals (CIs) were estimated. The first 3 PCs for nutrition factors and physical activities could explain 70% variances. The first principal component ($PC_1$) is a measure of nutrition factors (fruit and vegetables consumption), $PC_2$ is a measure for physical activities (moderate exercise and strength training), and PC3 is about calorie information use and soda use. Weighted multiple logistic regression showed that African Americans, middle aged adults (45-64 years), elderly (65+), never married, and with lower education were associated with increased odds of diabetes. After adjusting for others factors, the $PC_1$ showed marginal association with diabetes (OR=0.84, 95% CI=0.70-1.01); while $PC_2$ and $PC_3$ revealed significant associations with diabetes (OR=0.73, 95% CI=0.61-0.86 and OR=0.85, 95% CI=0.74-0.99, respectively). In conclusion, PCA can be used to reduce the indicators in complex survey data. The first 3 PCs of nutrition factors and physical activities were associated with diabetes. Promotion of health food and physical activities should be encouraged to help decrease the prevalence of diabetes.

**Keywords:** Diabetes; Nutrition; Physical activities; Principal component analysis; Weighted logistic regression

## Introduction

Globally, there were 284.6 million of people with diabetes in 2010 and it was predicted to be 438.4 million in 2025 [1]. In the United Sates (US), it was reported that over 29 million people were living with diabetes and 37% of adults aged 20 years or older were pre-diabetic in 2012 [2,3]. The burden for diabetes will rise from 418 billion dollars to 490 billion dollars from 2010 to 2030 [4]. Several factors have been reported to be associated with diabetes such as family history, ethnic background, aging, being overweight, physical inactivity, alcohol use and smoking [5-8]; however, the impact of alcohol and smoking on diabetes has inconsistent findings.

It has been reported that regular consumption of fruit and vegetables, reduced consumption of saturated fats, sodium and sugary drinks, as well as increased physical activity and control of smoking habits could reduce the incidence of diabetes [9]. For example, dietary patterns characterized by high intakes of fruits and vegetables, whole grains, low-fat dairy products, and low glycemicoad have been associated with lower risk of type 2 diabetes [10-15]. A meta-analysis showed that increasing the amount of green leafy vegetables in an individual's diet could help to reduce the risk of type 2 diabetes [16]. Two-three servings/day of vegetable and 2 servings/day of fruit conferred a lower risk of type 2 diabetes than other levels of vegetable and fruit consumption, respectively [17]. However, it was found that vegetable but not fruit consumption reduced the risk of type 2 diabetes in Chinese women [18]; while another study showed that fruit or vegetables separately were not associated with diabetes, only green leafy vegetable intake was inversely associated with diabetes [19]. A recent study revealed that fruit and vegetable intake was not related to incidence of type 2 diabetes in older subjects [20]. Furthermore, only small differences were found in dietary behavior in comparison with cohort members without diabetes [21,22]. Another study found non-linear association of fruit intake with type 2 diabetes [23]. Muraki et al. concluded that there was heterogeneity in the associations between individual fruit consumption and risk of type 2 diabetes [24]. Previous study has suggested a correlation between drinking diet soda and glucose control in adults with diabetes [25]; while reduced sugar intake showed improvements in key risk factors for type 2 diabetes [26]. A recent study suggested that the impact of sugar on diabetes may be independent of sedentary behavior and alcohol use, and obesity [27]. A more recent study showed that consumptions of soft drinks, sweetened-milk beverages and energy from total sweet beverages were associated with increasing risk of type 2 diabetes [28].

Principal component analysis (PCA) is one of the most popular methods used for variable reduction, which can overcome the disturbance of the multicollinearity of the risk factors and has been used in social sciences, health service, and health sciences [29-32]. For example, PCA has been used to examine dietary patterns with diabetes

***Corresponding author:** Wang K, Department of Biostatistics and Epidemiology, College of Public Health, East Tennessee State University, Johnson City, TN 37614-1700, USA, Tel: +1 423 439 4481; E-mail: wangk@etsu.edu

**Received** July 31, 2017; **Accepted** August 27, 2017; **Published** August 31, 2017

**Citation:** Wang KS, Liu Y, Xie X, Gong S, Xu C, et al. (2017) Principal Component Regression Analysis of Nutrition Factors and Physical Activities with Diabetes. J Biom Biostat 8: 364. doi: 10.4172/2155-6180.1000364

in the US adults [33], Chinese population [34-37], and Japanese population [38,39]. It is also used to investigate the relationship between the physical activity and diabetes [38,40,41].

The associations of nutrition factors and physical activities with adult diabetes are inconsistently reported. For example, high levels of physical activities are associated with reduced risk of diabetes; however, some patients at risk for diabetes were inactive [40,41]. On the other hand, as shown previously, higher intakes of fruit, berries, and vegetables have been associated with reduced risk of diabetes in some observational studies; however, the evidence is limited and inconclusive [42]. Furthermore, most of these nutrition factors and physical activities are correlated. No study has been found to use PCA to extract PCs of these nutrition factors and physical activities followed by a logistic regression analysis to examine their associations with diabetes. In the present study, we collected data from the 2012 Health Information National Trends Survey (HINTS 4, Cycle 2). The aims of this study were to overcome the disturbance of the multicollinearity among the risk factors and examine the associations of these factors with diabetes using PCA and weighted logistic regression.

## Materials and Methods

### Participants

The data was drawn from the 2012 HINTS4, Cycle 2. The HINTS is a nationally-representative survey which has been administered by the US National Cancer Institute (NCI) since 2003. The HINTS target population includes adults aged 18 or older in the civilian non-institutionalized population of the US. The collection of the Cycle 2 data was conducted from October 2012 through January 2013. The sample design for the Cycle 2 survey is a two-stage design. In the first stage, a stratified sample of addresses was selected from a file of residential addresses. In the second-stage, one adult was selected within each sampled household. The respondent selection would be conducted uniformly for all households in Cycle 2 using the Next Birthday Method, in which one questionnaire was sent with each mailing so that the adult who would have the next birthday in the sampled household was asked to complete the questionnaire. Every sampled adult who completed a questionnaire in Cycle 2 received a full-sample weight and a set of 50 replicate weights. The full-sample weight is the weight which is used to calculate population and subpopulation estimates from the HINTS data collected in Cycle 2; while the replicate weights are used to compute standard errors for these estimates. More extensive background about the HINTS program and data collection efforts are available elsewhere [43,44]. The final HINTS 4 Cycle 2 sample consists of 3,630 respondents. The overall household response rate using the Next Birthday Method was 39.97%. This current study was approved by the IRB of East Tennessee State University.

### Outcome

Subjects were considered to have diabetes if they responded "yes" to the question "Has a doctor or other health professional ever told you that you had any of the following medical conditions: Diabetes or high blood sugar?" Controls were those if they responded "no" to the question. Of the 3,630 adults, 2,586 responded to the question including 659 with diabetes and 2,827 non-diabetic individuals (Table 1).

### Independent variables

Demographic characteristics included gender, age group (18-49 years, 50-64 years, 65+), race, marital status (married/living together, widowed/divorced/separated, and never married), and education. Race was recoded as Hispanic, Non-Hispanic White, Non-Hispanic Black or African American (AA), and other. Education was determined by asking whether he/she had a high school degree or not. Smoking status was classified as never smoking, current smoking, or past smoking.

Soda use was defined by the question "Not counting any diet soda

| Variable | Total (N) | Diabetes | Prevalence (%) | 95%CI | p-value |
|---|---|---|---|---|---|
| Gender | | | | | |
| Male | 1335 | 252 | 13.2 | 10.7-15.7 | 0.123 |
| Female | 2094 | 392 | 16.0 | 13.8-18.2 | |
| Age group | | | | | |
| 18-49 years | 1354 | 126 | 8.1 | 6.2-9.4 | <0.0001 |
| 50-64 years | 1124 | 881 | 20.8 | 17.9-23.8 | |
| 65+ years | 910 | 260 | 25.7 | 21.7-29.7 | |
| Race | | | | | |
| White | 1985 | 308 | 13.1 | 11.514.7 | 0.641 |
| AA | 475 | 118 | 16.3 | 12.3-20.3 | |
| Hispanic | 491 | 93 | 13.8 | 8.5-19.1 | |
| Other | 202 | 39 | 16.6 | 5.8-27.5 | |
| Marital status | | | | | |
| Married/living together | 1801 | 317 | 16.0 | 13.5-18.5 | <0.0001 |
| Widowed/Divorced/Separated | 988 | 242 | 23.3 | 19.4-27.1 | |
| Never Married | 610 | 81 | 5.9 | 3.9-8.0 | |
| Education | | | | | |
| ≤High school | 1050 | 295 | 21.6 | 17.9-25.2 | <0.0001 |
| >High school | 2363 | 345 | 10.8 | 9.1-12.5 | |
| Smoke status | | | | | |
| Current | 569 | 116 | 12.6 | 9.2-16.0 | 0.0009 |
| Former | 905 | 197 | 20.1 | 16.1-24.1 | |
| Never | 1969 | 337 | 13.1 | 11.2-14.9 | |
| Overall | 3486 | 659 | 14.6 | 13.1-16.1 | |

Abbreviations: AA: African American, CI: Confidence interval, p-value is based on $\chi^2$ test.

**Table 1:** Prevalence of diabetes in lifetime (%) within each group of exploratory variables.

or pop, about how often do you drink regular soda or pop in a typical week?" There are six ordinal levels (don't drink any regular soda or pop, less than 1 day a week, 1-2 days a week, 3-4 days a week, 5-6 days a week, and every day). Fruit consumption was defined by the question "About how many cups of fruit (including 100% pure fruit juice) do you eat or drink each day?" Seven levels were categorized such as none, ½ cup or less, ½ cup to 1 cup, 1 to 2 cups, 2-3 cups, 3-4 cups, and 4 or more cups. Vegetable consumption was defined by the question "About how many cups of vegetables (including 100% pure vegetable juice) do you eat or drink each day?" Seven levels were categorized such as none, ½ cup or less, ½ cup to 1 cup, 1 to 2 cups, 2-3 cups, 3-4 cups, and 4 or more cups. Calorie information use was defined by "When available, how often do you use menu information on calories in deciding what to order?" Five levels were categorized such as never, rarely, sometimes, often, and always" Moderate exercise was defined by the question "In a typical week, how many days do you do any physical activity of at least moderate intensity?" There are seven levels (none, 1 day a week, 2 days a week, 3 days a week, 4 days a week, 5 days a week, and 6-7 days a week). Strength training was defined by the question "In a typical week, how many days do you do leisure-time physical activities specifically designed to strengthen your muscles?" There are seven levels (none, 1 day a week, 2 days a week, 3 days a week, 4 days a week, 5 days a week, and 6-7 days a week).

### Descriptive statistics and prevalence

All the analyses were conducted using Statistical Analysis System (SAS) (version 9.4, SAS Inc., Cary, NC, USA). Data were weighted to produce overall and stratified estimates that would be nationally representative of the US population. Weights were derived initially from selection probabilities to compensate for planned oversampling procedures. The resulting weights were then calibrated using comparable population characteristics for sex, age, race, and education from data publicly available through the current population survey. A set of 50 replicate weights was used in order to generate an unbiased estimation of population variance. The PROC SURVEYFREQ procedure was used to weight and estimate population proportions in cases and controls and in different stratified demographics; while PROC SURVEYMEANS was used to estimate the overall prevalence. The chi-square test was used to compare the prevalence of diabetes across age, gender, and races.

### Principal component analysis

The PCA is an effective method to reduce the dimensionality of multivariate data. It is possible to account for the most information in the original data set with a relatively small number of PCs and there is no correlation among PCs [29]. Generally, the first Principal Component ($PC_1$) will be the linear combination of the variables that captures the maximum amount of information in the data and will be correlated with at least some of the observed variables. The general formula (1) is used to compute scores on the $PC_1$ extract in a PCA [31].

$$PC_1=b_{11}(X_1)+b_{12}(X_2)+\ldots b_{1k}(X_k) \quad (1)$$

Where,

$PC_1$=the participant's score on the first PC (the first component extracted)

$b_{1k=}$the coefficient (or weight) for observed variable k, as used in creating $PC_1$

$X_{k=}$the participant's score on the observed variable k, k=1,2,…k.

The second PC ($PC_2$) accounts for a maximum amount of variance in the data that was not accounted for by $PC_1$ and will be correlated with at least some of the observed variables that did not display strong correlations with $PC_1$. An eigenvalue reflects the amount of variance captured by a given PC. The eigenvalue-one criterion (eigenvalue≥1) is commonly used to decide how many PCs to be retained [45,46]. The proportion of variation explained by each PC can be calculated with formula (2). Any PC which accounts for at least 5% or 10% of the total variance can be retained.

$$Proportion = \frac{Eigenvalue\ for\ the\ component\ of\ interest}{Total\ eigenvalues\ of\ the\ correlation\ matrix} \quad (2)$$

A varimax rotation produces uncorrelated components and is the most commonly used orthogonal rotation in practice [47]. A factor loading of one independent variable is considered as large if its absolute value exceeds 0.40 [46]. Using ordinal and dichotomous indicators is a very common practice in social sciences and health sciences. It has been suggested that a polychoric correlation was created instead of Pearson's correlations for the categorical variable in PCA and other multivariate analyses [48]. A polychoric correlation and Pearson's correlation were calculated using PROC CORR for PCA; while PCA was performed with PROC FACTOR with SAS statistical software. A Scree diagram, a visual graphic display of the eigenvalues, was obtained using the SCREE option in the PROC FACTOR. The components in the steep curve before the first point that starts the flat line trend were retained.

### Weighted multiple logistic regression analysis

Multiple logistic regression analysis (3) with diabetes as a binary trait, adjusted for covariates, was performed using SAS software.

$$\text{logit}(p(Y_1=1))=\beta_0+\beta_1X_{p+1}+\beta_2X_{p+2}+\ldots\ \beta_mX_{p+m}+\beta_{pc1}PC_1+\beta_{pc2}PC_2+\ldots+\beta_{pcn}PC_n \quad (3)$$

where $Y_1$ is diabetes status (1 if diabetes) and $\beta_m$ is the slope for observed $m^{th}$ variable and $X_{p+m}$ is the value of observed variable m; while $\beta_{pcn}$ is the slope for the $n^{th}$ PC and $PC_n$ is the score of the $n^{th}$ PC.

The SURVEYLOGISTIC procedure fits logistic regression models for discrete response survey data by the method of maximum likelihood. The asymptotic p-values for this test were observed while the odds ratio (OR) and standard error (SE) of OR were estimated. Variances of the regression parameters and odds ratios were computed by using either the Taylor series (linearization) method or replication (resampling) methods to estimate sampling errors of estimators based on complex sample designs [49-52]. Two models were conducted to investigate the relationship between the occurrence of diabetes and its exploratory variables. In model one, simple logistic regression was used to examine the role of each potential risk factor including first several PCs on diabetes. In model two, multiple logistic regression models were used to adjust for all potential risk factors including PCs of diabetes.

## Results

### Prevalence of diabetes

Table 1 presents the prevalence of diabetes. The overall prevalence of diabetes was 14.6% (13.2% for males and 16.0% for females). There were no significant differences between males and females and among different race groups. The prevalence increased with age (8.1%, 20.8% and 25.7% for age groups 18-49, 50-64 and 65+ years, respectively). Higher prevalence was found for the individuals with lower education, being widowed/divorced/separated, and former smoking.

## Principal component analysis

The correlation coefficients among nutrition factors and physical activity are presented in Table 2. The fruit and vegetables consumption, moderate exercise and strength training, and calorie information use have significantly positive correlations using both polychoric correlation and Person's correlation ($p<0.0001$); whereas the regular soda use has significantly negative correlations with all other five factors ($p<0.0001$).

The first three PCs explained about 70% of total variation. The eigenvalues of first three PCs were 2.1009, 1.118 and 0.9786, respectively and the proportions of variation explained by these three PCs were 35%, 18.6% and 16.3%, respectively (Table 3). The Scree diagram in Figure 1 also revealed the first three PCs are appropriate to choose by considering proportion of variation. The rotated factor patterns of the first 3 PCs are presented in Table 4. The first $PC_1$ is strongly and positively correlated with fruit and vegetables consumption. More specifically, the $PC_1$ increases as the consumption of fruit and vegetables increases. This component can be viewed as a measure of nutrition with high loading values for fruits and vegetables (both loadings were 0.85). The $PC_2$ increases with increasing moderate exercise and strength training (loading values were 0.82 and 0.86, respectively); therefore, it can be treated as component for measuring of physical activities. The $PC_3$ increases with increasing soda use (loading value was 0.85), but decreasing calorie information use (loading value was -0.68).

| Variable | Calorie information use | Fruit consumption | Vegetable consumption | Soda use | Moderate exercise | Strength training |
|---|---|---|---|---|---|---|
| Calorie information use | 1.000 | 0.1927 | 0.1846 | -0.2163 | 0.1523 | 0.1282 |
| Fruit consumption | 0.2164 | 1.0000 | 0.4992 | -0.1365 | 0.1837 | 0.1427 |
| Vegetable consumption | 0.2048 | 0.5487 | 1.000 | -0.1560 | 0.2040 | 0.1471 |
| Soda use | -0.2565 | -0.1562 | -0.1965 | 1.000 | -0.1066 | -0.0697 |
| Moderate Exercise | 0.2060 | 0.2250 | 0.2418 | -0.1373 | 1.000 | 0.4308 |
| Strength Training | 0.2023 | 0.211 | 0.1986 | -0.1146 | 0.5739 | 1.000 |

Above diagonal is Person correlation coefficient; below the diagonal is polychoric correlation coefficient.
The p values of all correlation coefficients are smaller than 0.0001.

**Table 2:** Correlation of nutrition factors and physical activities.

| PC | Eigenvalue | Difference | Variance proportion | Cumulative variance proportion |
|---|---|---|---|---|
| 1 | 2.1009 | 0.9829 | 0.3501 | 0.3501 |
| 2 | 1.1180 | 0.1394 | 0.1863 | 0.5365 |
| 3 | 0.9786 | 0.2124 | 0.1631 | 0.6996 |
| 4 | 0.7661 | 0.2213 | 0.1277 | 0.8273 |
| 5 | 0.5448 | 0.0532 | 0.0908 | 0.9181 |
| 6 | 0.4916 | | 0.0819 | 1.0000 |

PC: Principal component.

**Table 3:** Eigenvalues and the proportion of variation explained by the principal components.

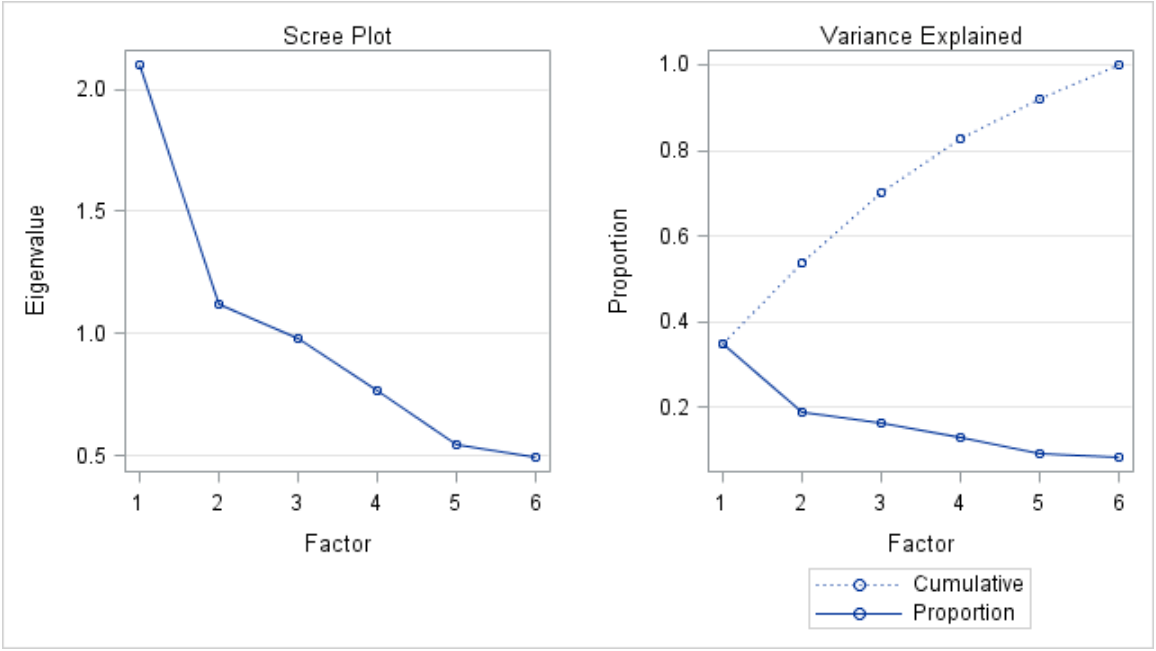


**Figure 1:** Scree plot of eigenvalues suggesting three components to the nutrition factors and physical activities.

| Variable | $PC_1$ | $PC_2$ | $PC_3$ |
|---|---|---|---|
| Calorie information use | 0.16 | 0.19 | -0.68* |
| Fruit consumption | 0.85* | 0.11 | -0.10 |
| Vegetable consumption | 0.85* | 0.11 | -0.12 |
| Soda use | -0.05 | 0.02 | 0.85* |
| Moderate exercise | 0.16 | 0.82* | -0.10 |
| Strength Training | 0.05 | 0.86* | -0.06 |

PC: Principal component.

*A factor loading of one independent variable is considered as large if its absolute value exceeds 0.40.

**Table 4:** Rotated factor pattern of nutrition factors and physical activities.

| Variable | Crude OR | 95% CI | p-value | Adjusted OR | 95% CI | p-value |
|---|---|---|---|---|---|---|
| Gender | | | | | | |
| Male | 1 | | | 1 | | |
| Female | 1.18 | 0.87-1.58 | 0.288 | 1.04 | 0.73-1.48 | 0.811 |
| Age group | | | | | | |
| 18-44 years | 1 | | | 1 | | |
| 45-64 years | 2.92 | 2.15-3.97 | <0.0001 | 2.19 | 1.53-3.14 | <0.0001 |
| 65 + | 4.04 | 2.86-5.70 | <0.0001 | 2.83 | 1.83-4.36 | <0.0001 |
| Race | | | | | | |
| White | 1 | | | 1 | | |
| AA | 1.34 | 0.97-1.85 | 0.0748 | 1.98 | 1.30-3.02 | 0.0015 |
| Hispanic | 1.10 | 0.70-1.74 | 0.684 | 1.26 | 0.79-1.98 | 0.330 |
| Other | 1.40 | 0.60-3.23 | 0.436 | 2.22 | 0.84-5.91 | 0.109 |
| Marital status | | | | | | |
| Married | 1 | | | 1 | | |
| Widowed/ Divorced/ Separated | 1.46 | 1.07-1.99 | 0.0174 | 1.06 | 0.77-1.45 | 0.725 |
| Never | 0.27 | 0.16-0.46 | <0.0001 | 0.38 | 0.20-0.70 | 0.0021 |
| Education | | | | | | |
| > High school | 1 | | | 1 | | |
| ≤ high school | 2.22 | 1.62-3.04 | <0.0001 | 1.80 | 1.27-2.54 | 0.001 |
| Smoking status | | | | | | |
| Never | 1 | | | 1 | | |
| Current | 0.93 | 0.64-1.36 | 0.697 | 0.89 | 0.57-1.33 | 0.511 |
| Former | 1.58 | 1.13-2.21 | 0.0069 | 1.18 | 0.78-1.78 | 0.439 |
| $PC_1$ | | | | | | |
| | 0.83 | 0.69-0.99 | 0.0453 | 0.84 | 0.70-1.01 | 0.066 |
| $PC_2$ | | | | | | |
| | 0.67 | 0.57-0.80 | <0.0001 | 0.73 | 0.61-0.86 | 0.0001 |
| $PC_3$ | | | | | | |
| | 0.85 | 0.75-0.94 | 0.0156 | 0.85 | 0.74-0.99 | 0.0328 |

Abbreviations: AA: African American; PC: Principal component; OR: Odds ratio; CI: Confidence interval.

**Table 5:** Univariate and multiple logistic regression analyses.

## Weighted logistic regression analyses

The results of univariate and multiple logistic regression analyses of independent factors including the first 3 PCs are presented in Table 5. By using univariate analysis, all factors except for gender and race were associated with diabetes ($p<0.05$). Multiple logistic regression analyses showed that lower education (OR=1.80, 95% CI=1.27-2.54), middle-aged adults (OR=2.18, 95% CI=1.53-3.14) and elderly adults (OR=2.83, 95% CI=1.83-4.36) were positively associated with diabetes. African Americans (AAs) (OR=1.98, 95% CI=1.30-3.02) were more likely to have diabetes compared to the Whites. Univariate logistic analysis revealed that the first 3 PCs were negatively associated with diabetes ($p<0.05$). After adjusted for others factors, the $PC_1$ showed a borderline association with diabetes (OR=0.84, 95% CI=0.70-1.01); while $PC_2$ and $PC_3$ revealed significant associations with diabetes (OR=0.73, 95% CI=0.61-0.86 and OR=0.85, 95% CI=0.74-0.99, respectively).

## Discussion

In this study, we found the prevalence of diabetes to be significantly higher in older adults, being widowed or divorced or separated, with low education, and being former smoking. The first 3 principal components ($PC_1$-$PC_3$) for nutrition factors and physical activities could explain 70% variances. The $PC_1$ is a measure of nutrition factors (fruit and vegetables consumption), the $PC_2$ is a factor for physical activity (moderate exercise and strength training), and the $PC_3$ is a measure of calorie information use and soda use. The results from weighted multiple logistic regressions showed that race, age, marital status and education were associated with diabetes. Univariate logistic analysis revealed that the first 3 PCs were negatively associated with diabetes ($p<0.05$). After adjusted for other factors, $PC_2$ and $PC_3$ were significantly associated with diabetes; however, the $PC_1$ showed a marginal association with diabetes.

Previous studies have shown that smoking is an independent risk factor for the development of diabetes [53-55]. Recently, a meta-analysis suggested that passive smoking is a risk factor of diabetes even in those who were not themselves active smokers [56]. However, both passive and active smoking is associated with diabetes in the elderly population [54]; whereas in men aged 25 years or over, morbid obesity and smoking were significantly associated with diabetes in Southern California American Indians [57]. In the present study, former smoking was a risk factor of diabetes in the univariate logistic analysis; however, after adjusting for other factors, the association disappeared. We speculated that smoking may have relationship with other factors. We further examined the polychoric correlation among these factors and found that smoking was correlated with age group ($p=0.0121$), education ($p<0.0001$), gender ($p<0.0001$) and marital status ($p=0.0281$).

Previous studies suggest that PCA can reduce recallable bias and the complexity of correlated data, which can be easily collected as single indicator variables in surveys [58,59]. For example, PCA has been used in dietary patterns with diabetes. It has been shown that fruits, green leafy vegetables, and regular soda were associated with lower risk of incident type 2 diabetes using the Multi-Ethnic Study of Atherosclerosis (MESA) [33]. Furthermore, the consumption of vegetables, fruits, soy and other legumes, whole grains, nuts, and seeds, likely decreases the risk of diabetes, while higher intake of processed meat, sweetened foods and beverages, fried foods, and refined grains increases the risk of developing type 2 diabetes in the Singapore Chinese health study [34]; while the dietary pattern of more vegetables, fruits and fish were associated with reduced risk and the dietary pattern of more meat and milk products were associated with an increased risk of diabetes in the Hong Kong Dietary Survey [35]. One Japanese study showed that consuming a healthy diet was associated with a lower risk for diabetes among the Japanese [38]. However, dietary patterns may not be appreciably associated with type 2 diabetes risk in Japanese [39]. In addition, one study suggested that consuming a healthy diet was associated with a lower risk for diabetes among the Japanese, particularly among those who eat regularly, habitually exercise are either non- or ex-smokers [38]. In the present study, we found that $PC_1$ was negatively associated with diabetes in univariate logistic analysis ($p=0.045$); however, after adjusting for others factors, the $PC_1$ showed marginal association with diabetes $p=0.066$); which indicated that diabetic individuals may have not realize the importance of nutrition on their health.

Another risk factor of diabetes is the lack of physical activity. Previous studies have shown that high levels of physical activity are associated with reduced risk of diabetes [60-63]. However, about 46% of primary care patients at risk for diabetes did not do physical activity per week [40]; while two-third of patients with diabetes remain inactive [64]. It has been recommended that moderate to vigorous physical activity can reduce the risk of chronic diseases such as diabetes and its complications [65-67]. Health counsellors should address these barriers to increase the patients' adherence to physical activity as the recommendations [41]. Our current results showed that physical activities ($PC_2$) were associated with a decreased risk of diabetes. To the best of our knowledge, few studies have used PCA to address the physical activity. For example, one study conducted exploratory principal components factor analyses of influences on physical activity instrument [40]; another study used PCA to extract the factors of barriers with physical activity level [41].

Previous study has suggested a correlation between drinking diet soda and glucose control in adults with diabetes [25]; while soda use was associated with greater risks of metabolic syndrome components and type 2 diabetes [28,68]; whereas reduced sugar intake showed improvements in key risk factors for type 2 diabetes [26]. A recent study suggested that the impact of consuming sugar on diabetes may be independent of sedentary behavior and alcohol use, and obesity [27]. In the present study, the $PC_3$ was negatively associated with diabetes, which suggested that diabetic individuals used less regular soda than non-diabetic. In addition, the calorie information use was negatively correlated to $PC_3$, and the logistic regression revealed that diabetic individuals used more calorie information than non-diabetic. The above results reflected the diabetic individuals pay more attention to their calorie intake to comply their physician's recommendation for diabetes treatment.

There are several important strengths in this study. First, new valuable variables were used, including strength training, regular soda use and calorie information use, which have not been intensively investigated in the past studies. Furthermore, the PCA was used to reduce variable dimension with keeping most of information followed by PCA. We are also aware certain limitations of this study, including the cross-sectional study design, which limits the ability to establish the causality as well as possible recallable, differential misclassification biases, and the effects of differences in how respondents interpreted survey questions.

## Conclusion

Our findings support the notion that PCA can be used to reduce the indicators in complex survey data. The PCs of nutrition factors and physical activities were associated with diabetes. Promotion of health food and physical activities should be encouraged to help decrease the prevalence of diabetes.

### Acknowledgements

The authors would like to thank the NCI for providing the Data from the 2012 Health Information National Trends Survey.

### References

1. IDF (2009) IDF Diabetes Atlas International Diabetes Federation Brussels.
2. Center for Disease Control and Prevention. National Diabetes Statistics Report, 2014 (2015). 2014-report-estimates-of-diabetes-and-its-burden-in-the-united-states.pdf. Accessed August 28, 2015.
3. Echouffo-Tcheugui JB, Caleyachetty R, Muennig PA, Narayan KM, Golden SH (2016) Cumulative social risk and Type 2 diabetes in US adults: the national health and nutrition examination survey (NHANES). Eur J Prev Cardiol 23: 128-228.
4. Zhang P, Zhang X, Brown J, Vistisen D, Sicree R, et al. (2010) Global healthcare expenditure on diabetes. Diabetes Res Clin Pract 87: 293-301.
5. Rosella LC, Manuel DG, Burchill C, Stukel TA, PHIAT-DM (2011) A population-based risk algorithm for the development of diabetes development and validation of the Diabetes Population Risk Tool (DPORT). J Epidemiol Community Health 65: 613-620.
6. Bozorgmanesh M, Hadaegh F, Azizi F (2011) Predictive performance of the visceral adiposity index for a visceral adiposity-related risk: type 2 diabetes. Lipids Health 10: 88.
7. Bi Y, Wang T, Xu M, Xu Y, Li M, et al. (2012) Advanced research on risk factors of type 2 diabetes. Diabetes Metab Res Rev 282: 32-39.
8. Schellenberg ES, Dryden DM, Vandermeer B, Ha C, Korownyk C (2013) Lifestyle interventions for patients with and at risk for type 2 diabetes a systematic review and meta-analysis. Ann Intern Med 159: 543-651.
9. World Health Organization (2005) Preventing chronic diseases a vital investment Geneva World Health Organization.
10. Elwood PC, Pickering JE, Fehily AM (2007) Milk and dairy consumption diabetes and the metabolic syndrome the Caerphilly prospective study. J Epidemiol Community Health 61: 695-808.
11. Ruidavets JB, Bongard V, Dallongeville J, Arveiler D, Ducimetiere P, et al. (2007) High consumptions of grain fish dairy products and combinations of these are associated with a low prevalence of metabolic syndrome. J Epidemiol Community Health 61: 810-817.
12. Cooper AJ, Sharp SJ, Lentjes MA, Luben RN, Khaw KT, et al. (2012) A prospective study of the association between quantity and variety of fruit and vegetable intake and incident type 2 diabetes. Diabetes Care 35: 1293-1300.
13. Hofe CR, Feng L, Zephyr D, Stromberg AJ, Hennig B, et al. (2014) Fruit and vegetable intake as reflected by serum carotenoid concentrations predicts reduced probability of polychlorinated biphenyl associated risk for type 2 diabetes National Health and Nutrition Examination Survey. Nutr Res 34: 285-393.
14. Li M, Fan Y, Zhang X, Hou W, Tang Z (2014) Fruit and vegetable intake and risk of type 2 diabetes mellitus meta-analysis of prospective cohort studies. BMJ Open 4: 105-497.
15. Maki KC, Nieman KM, Schild AL, Kaden VN, Lawless AL, et al. (2015) Sugar-sweetened product consumption alters glucose homeostasis compared with dairy product consumption in men and women at risk of type 2 diabetes mellitus. J Nutr 145: 459-666.
16. Carter P, Gray LJ, Troughton J, Khunti K, Davies MJ (2010) Fruit and vegetable intake and incidence of type 2 diabetes mellitus systematic review and meta-analysis. BMJ 341: 42-129.
17. Wu Y, Zhang D, Jiang X, Jiang W (2015) Fruit and vegetable consumption and risk of type 2 diabetes mellitus a dose-response meta-analysis of prospective cohort studies. Nutr Metab Cardiovasc Dis 25: 140-177.
18. Villegas R, Shu XO, Gao YT, Yang G, Elasy T, et al. (2008) Vegetable but not fruit consumption reduces the risk of type 2 diabetes in Chinese women. J Nutr 138: 574-580.
19. Cooper AJ, Forouhi NG, Ye Z, Buijsse B, Arriola L, et al. (2012) Fruit and vegetable intake and type 2 diabetes EPIC-InterAct prospective study and meta-analysis. Eur J Clin Nutr 66: 1082-1092.
20. Mamluk L, O'Doherty MG, Orfanos P, Saitakis G, Woodside JV, et al. (2017) Fruit and vegetable intake and risk of incident of type 2 diabetes results from the consortium on health and ageing network of cohorts in Europe and the United States . Eur J Clin Nutr 71: 83-91.
21. Nothlings U, Boeing H, Maskarinec G, Sluik D, Teucher B, et al. (2011) Food intake of individuals with and without diabetes across different countries and ethnic groups. Eur J Clin Nutr 65: 635-641.
22. Morton S, Saydah S, Cleary SD (2012) Consistency with the dietary approaches to stop hypertension diet among adults with diabetes. J Acad Nutr Diet 112: 1780-1805.
23. Li S, Miao S, Huang Y, Liu Z, Tian H, et al. (2015) Fruit intake decreases risk of incident type 2 diabetes an updated meta-analysis. Endocrine 48: 454-460.
24. Muraki I, Imamura F, Manson JE, Hu FB, Willett WC, et al. (2013) Fruit consumption and risk of type 2 diabetes results from three prospective longitudinal cohort studies. BMJ 347: 50-101.
25. Mackenzie T, Brooks B, O'Connor G (2006) Beverage intake diabetes and glucose control of adults in America. Ann Epidemiol 16: 688-691.

26. Ventura E, Davis J, Byrd-Williams C, Alexander K, McClain A, et al. (2009) Reduction in risk factors for type 2 diabetes mellitus in response to a low-sugar high-fiber dietary intervention in overweight Latino adolescents. Arch Pediatr Adolesc Med 163: 320-327.

27. Basu S, Yoffe P, Hills N, Lustig RH (2013) the relationship of sugar to population-level diabetes prevalence an econometric analysis of repeated cross-sectional data. PLoS One. 8: 57-873.

28. O'Connor L, Imamura F, Lentjes MA, Khaw KT, Wareham NJ, et al. (2015) Prospective associations and population impact of sweet beverage intake and type 2 diabetes and effects of substitutions with alternative beverages. Diabetologia 58: 1474-1483.

29. Lattin J, Carroll D, Green P (2002) Analyzing Multivariate Data 1st Edition, Cengage Learning.

30. Navarro Silvera SA, Mayne ST, Risch HA, Gammon MD, Vaughan T, et al. (2011) Principal component analysis of dietary and lifestyle patterns in relation to risk of subtypes of esophageal and gastric cancer. Ann Epidemiol 21: 543-550.

31. O'Rourke N, Larry Hatcher L (2013) A Step-by-Step Approach to Using SAS for Factor Analysis and Structural Equation Modeling, Second Edition.

32. Jackson EF, Siddiqui A, Gutierrez H, Kanté AM, Austin J, et al. (2015) Estimation of indices of health service readiness with a principal component analysis of the Tanzania Service Provision Assessment Survey. BMC Health Serv Res 15: 536.

33. Nettleton JA, Steffen LM, Ni H, Liu K, Jacobs DR (2008) Dietary patterns and risk of incident type 2 diabetes in the Multi-Ethnic Study of Atherosclerosis (MESA). Diabetes Care 31: 177-282.

34. Odegaard AO, Koh WP, Butler LM, Duval S, Gross MD, et al. (2011) Dietary patterns and incident type 2 diabetes in Chinese men and women: the Singapore Chinese health study. Diabetes Care 34: 880-905.

35. Yu R, Woo J, Chan R, Sham A, Ho S, et al. (2011) Relationship between dietary intake and the development of type 2 diabetes in a Chinese population the Hong Kong Dietary Survey. Public Health Nutr 14: 133-141.

36. Zuo H, Shi Z, Yuan B, Dai Y, Pan X, et al. (2013) Dietary patterns are associated with insulin resistance in Chinese adults without known diabetes. The British journal of nutrition 109: 162-199.

37. Batis C, Mendez MA, Gordon-Larsen P, Sotres-Alvarez D, Adair L, et al. (2016) Using both principal component analysis and reduced rank regression to study dietary patterns and diabetes in Chinese adults. Public Health Nutr 19: 195-203.

38. Morimoto A, Ohno Y, Tatsumi Y, Mizuno S, Watanabe S (2012). Effects of healthy dietary pattern and other lifestyle factors on incidence of diabetes in a rural Japanese population. Asia Pac J Clin Nutr 21: 601-608.

39. Nanri A, Shimazu T, Takachi R, Ishihara J, Mizoue T, et al. (2013) Dietary patterns and type 2 diabetes in Japanese men and women the Japan Public Health Center-based Prospective Study. Eur J Clin Nutr 67: 18-24.

40. Donahue KE, Mielenz TJ, Sloane PD, Callahan LF, Devellis RF (2006) Identifying supports and barriers to physical activity in patients at risk for diabetes. Prev Chronic Dis 3: 119.

41. Halali F, Mahdavi R, Asghari Jafarabadi M, Mobasseri M, Namazi N (2016) A cross-sectional study of barriers to physical activity among overweight and obese patients with type 2 diabetes in Iran. Health Soc Care Community 24: e92-e100.

42. Mursu J, Virtanen JK, Tuomainen TP, Nurmi T, Voutilainen S (2014) Intake of fruit, berries, and vegetables and risk of type 2 diabetes in Finnish men: the Kuopio Ischaemic Heart Disease Risk Factor Study. Am J Clin Nutr 99: 328-333.

43. Nelson DE, Kreps GL, Hesse BW, Croyle RT, Willis G, et al. (2004) The Health Information National Trends Survey (HINTS): development, design, and dissemination. J Health Commun 9: 443-460.

44. Finney Rutten LJ, Davis T, Beckjord EB, Blake K, Moser RP, et al. (2012) Picking Up the Pace: Changes in Method and Frame for the Health Information National Trends Survey (2011-2014). Journal of Health Communication 17: 979-989.

45. Kaiser H F (1960) The application of electronic computers to factor analysis. Educ Psycho Meas 20: 141-151.

46. Stevens J (2002) Applied multivariate statistics for the social sciences. Hillsdale, NJ: L. Erlbaum.

47. Chou PH, O'Rourke N (2012) Development and initial validation of the Therapeutic Misunderstanding Scale for use with clinical trials research participants. Aging Ment Health 16: 145-153.

48. Muthén B, Muthén LK (2000) Integrating person-centered and variable-centered analyses: growth mixture modeling with latent trajectory classes. Alcohol Clin Exp Res 24: 882-891.

49. Binder DA (1983) On the Variances of Asymptotically Normal Estimators from Complex Surveys. International Statistical Review 51: 279-292.

50. Särndal CE, Swensson B, Wretman J (1992) Model Assisted Survey Sampling. Springer-Verlag, New York.

51. Wolter KM (1985) Introduction to Variance Estimation. Springer-Verlag, New York.

52. Rao JNK, Wu CFJ, Yue K (1992) Some Recent Work on Resampling Methods for Complex Surveys. Survey Methodology 18: 209-217.

53. Hayashino Y, Fukuhara S, Okamura T, Yamato H, Tanaka H, et al. (2008) A prospective study of passive smoking and risk of diabetes in a cohort of workers: the high-risk and population strategy for occupational health promotion (HIPOP-OHP) study. Diabetes Care 31: 732-734.

54. Kowall B, Rathmann W, Strassburger K, Heier M, Holle R, et al. (2010) Association of passive and active smoking with incident type 2 diabetes mellitus in the elderly population: the KORA S4/F4 cohort study. Eur J Epidemiol 25: 393-402.

55. Yeh HC, Duncan BB, Schmidt MI, Wang NY, Brancati FL (2010) Smoking, Smoking Cessation, and Risk for Type 2 Diabetes Mellitus A Cohort Study. Annals of Internal Medicine 152: 10-17.

56. Wei X, E M, Yu S (2015) A meta-analysis of passive smoking and risk of developing Type 2 Diabetes Mellitus. Diabetes Res Clin Pract 107: 9-14.

57. Reid JL, Morton DJ, Wingard DL, Garrett MD, von Muhlen D, et al. (2010) Sex and age differences in the association of obesity and smoking with hypertension and type 2 diabetes in Southern California American Indians, 2002-2006. Ethn Dis 20: 231-238.

58. Yvas S, Kumaranayake L (2006) Constructing socio-economic status indices: how to use principal components analysis. Health Policy Plan 21: 459-468.

59. Krefis AC, Schwarz NG, Nkrumah B, Acquah S, Loag W, et al. (2010) Principal component analysis of socioeconomic factors and their association with malaria in children from the Ashanti Region, Ghana. Malar J 1: 201.

60. Fulton-Kehoe D, Hamman RF, Baxter J, Marshall J (2001) A case-control study of physical activity and non-insulin dependent diabetes mellitus (NIDDM). the San Luis Valley Diabetes Study. Ann Epidemiol 11: 320-327.

61. Wang L, Yamaguchi T, Yoshimine T, Katagiri A, Shirogane K, et al. (2002) A case-control study of risk factors for development of type 2 diabetes: emphasis on physical activity. J Epidemiol 12: 424-430.

62. Qin L, Corpeleijn E, Jiang C, Thomas GN, Schooling CM, et al. (2010) Physical activity, adiposity, and diabetes risk in middle-aged and older Chinese population: the Guangzhou Biobank Cohort Study. Diabetes Care 33: 2342-2348.

63. Fretts AM, Howard BV, McKnight B, Duncan GE, Beresford SA, et al. (2012) Modest levels of physical activity are associated with a lower incidence of diabetes in a population with a high rate of obesity: the strong heart family study. Diabetes Care 35: 1743-1745.

64. Palakodeti S, Uratsu CS, Schmittdiel JA, Grant RW (2015) Changes in physical activity among adults with diabetes: a longitudinal cohort study of inactive patients with Type 2 diabetes who become physically active. Diabet Med 32: 1051-1057.

65. Maeda S, Miyaki A, Kumagai H, Eto M, So R, et al. (2013) Lifestyle modification decreases arterial stiffness and plasma asymmetric dimethylarginine level in overweight and obese men. Coron Artery Dis 24: 583-588.

66. Lin CH, Chiang SL, Yates P, Lee MS, Hung YJ, et al. (2015) Moderate physical activity level as a protective factor against metabolic syndrome in middle-aged and older women. J Clin Nurs 24: 1234-1245.

67. Xiao J, Shen C, Chu MJ, Gao YX, Xu GF, et al. (2016) Physical Activity and Sedentary Behavior Associated with Components of Metabolic Syndrome among People in Rural China. PLoS One 11: e0147062.

68. Nettleton JA, Lutsey PL, Wang Y, Lima JA, Michos ED, et al. (2009) Diet soda intake and risk of incident metabolic syndrome and type 2 diabetes in the Multi-Ethnic Study of Atherosclerosis (MESA). Diabetes Care 32: 688-694.