 doi:10.20944/preprints202007.0106.v1

*Type of the Paper: Article*

# Sociodemographic and Genetic Influences on Dietary Patterns and Their Influence on Health Outcomes in the Atlanta Center for Health Discovery and Well Being Cohort

**Jingheng Chen[1], Xueying Huang[1], Thomas R. Ziegler[2], Dean P. Jones[2], Michelle Lampl[2], Arshed A. Quyyumi[2], Jane Clark[2], Gregory S. Martin[2], Kenneth L. Brigham[2], and Greg Gibson[1,*]**

[1] Georgia Institute of Technology, Atlanta GA 30332, USA; jingheng.chen1995@gmail.com; Sonia.xyh@gmail.com

[2] Emory University, Atlanta GA 30322, USA; tzieg01@emory.edu, dpjones@emory.edu, mlampl@emory.edu, aquyyum@emory.edu, jbehclark@emory.edu, greg.martin@emory.edu, KBrigha@emory.edu

* Correspondence: greg.gibson@biology.gatech,edu; Tel.: +1-404-385-2343

**Abstract:** Diet influences, and is influenced by, a wide range of socioeconomic, cultural, geographic, and genetic variables. Here we survey a matrix of such interactions as well as their connection to a variety of health outcomes, in a cohort of 689 diverse adults employed at Emory University and enrolled in the Center for Health Discovery and Well-Being (CHDWB) study. Principal component analysis (PCA) of the Block Food Frequency Questionnaire revealed seven PC cumulatively explaining 25.8% and each individually at least 2% of the proportional consumption of 110 food items. PC1 is strongly correlated with the Healthy Eating Index-2015 measure, and accordingly healthier scores associate with multiple measures of physical and mental health. It, as well as PC2 (likely a measure of food expense) and PC3 (carbohydrate versus protein consumption) show significant geographic structure across the Atlanta metropolitan area, correlating with race and ethnicity, income level, age and sex. Notably, a polygenic score for body mass index (BMI) consisting of 281 SNPs explains 2.8% of the variance in PC5, which is as strong as its association with BMI itself. PC5 appears to differentiate participants with respect to conscious eating behavior related to the choice of diet or comfort foods. Our analysis adds to the growing literature on factor analysis of socio-demographic influences on nutrition and health.

**Keywords:** polygenic risk; wellness; food frequency; principal component analysis; healthy eating index; obesity; food desert

## 1. Introduction

While it is well established that obesity and metabolic disease are mediated in part by total food intake, and the basic components of a healthy diet are well-known [1,2], rates of conformity to healthy diet recommendations differ widely across populations. Variation in diet has also been suspected as one of the leading mechanisms mediating the relationship between socioeconomic status and health outcomes [3,4]. Reciprocally, genetics and health-related factors also contribute to dietary choice [5,6]. Much remains to be learned about the distribution of dietary patterns across different socio-demographic, genetic and health spectra as well as the relative effect of these variables on dietary preference.

Large scale community surveys that include food frequency questionnaires (FFQ) provide exciting opportunities to link estimates of dietary consumption and choice to genetics, health,

 doi:10.20944/preprints202007.0106.v1



socioeconomic, and other cultural variables. Typical dietary epidemiological analyses start by extracting a subset of the information from a food consumption survey, such as the average daily or weekly intake amount of whole grains, fruits, certain types of meats, or nutrients like dietary fibers and saturated fats. One tool for assessing food consumption is the Block FFQ, which self-reports the frequency and qualitative amount of consumption of over one hundred food items [7,8].

Another widely used measure of diet quality computed from food intake is a summary "healthy eating index" (HEI). A series of versions of such indices have been developed, including the HEI-2005 [9], HEI-2015 [10], and Alternative HEI [11]. These scores quantify an individual's conformance to the Dietary Guidelines for Americans developed by the U.S. Department of Agriculture (USDA). Individuals who meet all aspects of these USDA dietary guidelines receive a maximum score, whereas a diet consisting of highly saturated fats, added sugar and insufficient nutrients will receive a poor score. While useful for assessment of overall diet quality, cumulative summary scores such as HEI are unable to capture aspects of dietary variation across a cohort, since the sources of the variation are not captured and hence two individuals with quite different diets may have similar scores.

An alternative approach is to use factor analysis or machine learning to identify dimensions underlying common dietary trends from the full matrix of FFQ information [12]. Summary scores along the principal components of the variation can then be correlated with health-related variables that might either influence dietary choices, or be influenced by them [13]. This approach evaluates the entire dietary pattern instead of individual nutrients or foods, but it emphasizes items that are the most variable across the sample cohort, and provides a more comprehensive view of food and nutrient consumption in the study cohort [14-16].

The Center for Health Discovery and Wellbeing (CHDWB) at the Predictive Health Institute of Emory University and Georgia Tech was established in 2008 with the aim of ascertaining whether targeted health interventions based on detailed self-knowledge of sub-clinical disease could have tangible health benefits [17]. Highly significant but modest improvements in multiple aspects of well-being were documented over a three-year period and were particularly notable in those with baseline poor health [18,19]. The cohort has a broad sociodemographic intake, and is drawn from the broader Atlanta metropolitan area. Previous studies from our group have investigated metabolomic correlates of diet-related obesity [20,21]. Here we report on an analysis of self-reported food frequency surveys and their relationship to health outcomes in the CHDWB cohort consisting of 689 adults aged 25-75.

We are interested in two major questions. First, how are dietary patterns structured across the cohort? To address this, we performed factor analysis on the dietary data for 110 food items, and used the identified principal components as individual surrogates for traditional healthy eating indices. These were contrasted with the HEI-2015 computed from the same data. Second, how are dietary tendencies correlated with demographic, health, and genetic factors? We used multivariate statistics to evaluate the influences of self-reported ethnicity, gender, age, education, marital status, income and geographic location on measures of dietary health, and evaluated whether polygenic risk scores for obesity differentiate individuals by diet. Finally, results are presented confirming general preconceptions about the influence of proportional food consumption on physical and mental health.

## 2. Materials and Methods

### *2.1. Data collection and classification*

The dietary information and personal health profiles analyzed in this study were generated by the Center for Health Discovery and Well Being (CHDWB), within the Predictive Health Institute of Emory University and Georgia Tech in Atlanta, Georgia. The full data set includes social, physical,

 doi:10.20944/preprints202007.0106.v1



physiological, psychological and lifestyle profiles for all participants of a longitudinal health promotion program [22]. The participants were generally healthy and active employees without uncontrolled chronic disease conditions, drawn at random from all sectors of Emory University, including a breadth of social backgrounds and ethnicities as indicated in Table 1. Sociodemographic information was collected at recruitment using an electronic Personal Information Form. This included gender, race/ethnicity, age (computed from date of birth and visit date), household income in ordinal levels, educational attainment in years of schooling completed, marital status, and zip-code of residence.

Associated with each visit over a four-year period, with 6-month intervals between the first three visits, and 12-months thereafter, participants were asked to complete a web-based Block FFQ. We analyze data for a total of 689 participants who reported total daily caloric intake in the range of 700-4200 kcal. They also had body composition measurements, blood was drawn for a comprehensive metabolic and immunological profile, and a range of surveys of health-related behavior facilitated computation of the Beck Depression Index (BDI)[23], General Anxiety Disorder-7 score (GAD-7)[24], Perceived Stress Scale (PSS-14)[25], Epworth Sleepiness Scale (ESS)[26], and the SF36 Quality of Life Survey [27]. Here we report on only the first visit since completion of the survey was variable at subsequent time-points, though the major PC of diet remain similar in a dataset including 2552 surveys. Additional details and analysis of health outcomes are provided in our previous publications [18,19,22].

**Table 1**. Characteristics of the CHDWB FFQ sample included in the study

| Characteristic | | Analytic Data Set |
|---|---|---|
| n | | 689 |
| Females, n (%) | | 453 (65.9%) |
| Age, years, mean ± SD | | 48.0 ± 11.0 |
| Race/Ethnicity, n (%) | | |
| | White, non-Hispanic | 487 (70.9%) |
| | Black, non-Hispanic | 151 (22.0%) |
| | Asian | 35 (5.1%) |
| | Hispanic | 9 (1.3%) |
| | American Indian/ Alaska Native | 5 (0.7%) |
| Education years, mean ± SD | | 18.9 ± 4.5 |
| Household Income, median group | | $100,000 to $150,000 |
| Marital Status, n (%) | | |
| | Single | 157 (23.1%) |
| | Married | 436 (64.0%) |
| | Divorced | 88 (12.9%) |

### 2.2 *Dietary pattern analysis*

The dietary questionnaire used in the study was a version of the semi-quantitative FFQ-2005 administered over the internet by NutritionQuest. It included 110 food items with specified serving sizes described in natural portions (e.g. 1 banana) or standard weight and volume measures of common servings. For each food item, participants indicated the intake frequency and number of portions per intake based on 7-day recall. Daily consumption of each food was calculated by

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 July 2020 doi:10.20944/preprints202007.0106.v1



multiplying weekly intake frequency with number of portions consumed, divided by 7 days. The questionnaire thus returned a matrix of food consumption data, along with software-generated dietary and nutritional measurements that align with the 2015-2020 USDA Dietary Guidelines for Americans [28]. Some examples of the dietary and nutritional measurements are: caloric intake per day, cup equivalents of whole fruits consumed per 1,000 kcal, and percent of energy that comes from saturated fats and from added sugar [9].

We computed HEI-2015 [10] from the dietary and nutrition summary data generated by NutritionQuest. Some variables required for the computation of HEI were not present in our FFQ dataset and were thus excluded or replaced by other variables. "Milk, including soy milk (cups)" was not present and the total dairy category was represented by "total cup equivalents of milk, yogurt and cheese" only. Similarly, "non-juice fruits" was replaced by "solid fruits", representing whole fruits. "Lean meat from soy products, excluding soy milk" was replaced by "lean meat from total soy products". Additional details of HEI calculation and FFQ variables used can be found on the NCI (https://epi.grants.cancer.gov/hei/) and WHI (https://www.whi.org/researchers/data/Pages/Available%20Data.aspx) and websites.

The proportional food consumption for the entire study cohort is summarized in Figure 1 in which the 110 food items are grouped into 23 categories based on an established food classification method [16] which reduces the total number of items to be analyzed while retaining much of the variety. As might be expected, the largest consumption was observed for fruits, sweets and deserts, refined grains, and meats (processed and red).

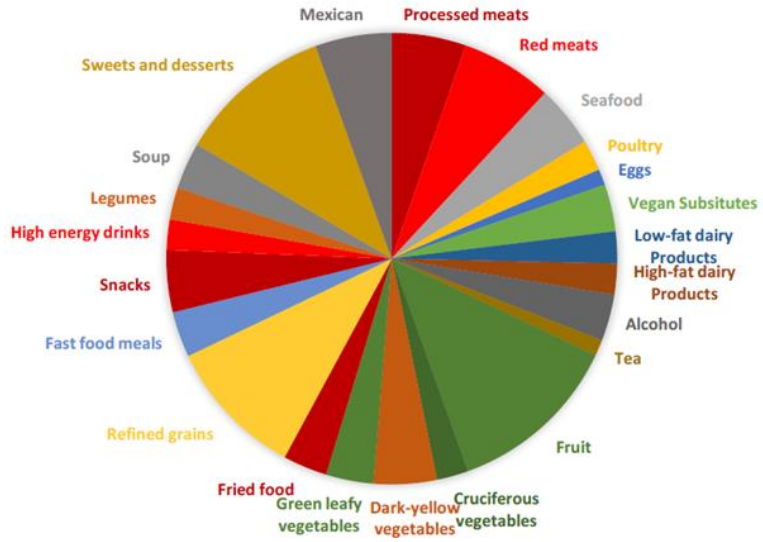


**Figure 1**. Overall dietary intake of the total study cohort (2552 surveys) at a glance.

Weekly consumed amounts of each food item were calculated by multiplying the intake frequency with number of portion sizes per intake. The missing food frequency and quantity data was imputed using the expectation-maximization method [29]. Initial analysis indicated that the dietary variation was mainly driven by the total amount of food consumed, rather than the proportions of each food. While interesting and relevant to the influence of psychosocial stress on health, for the purposes of this study we considered it a bias to be overcome. Consequently, we transformed the food amounts into their relative proportions in each person's diet by and dividing each food amount by the sum of all the food amounts. These values, after standardization into z-scores, were used as the entries into Principal Component Analysis (scikit-learn version 0.19.1, Python). Although the percent variance explained by each the major PC decreased slightly relative to PC generated with non-proportionalized data, the contributions of different items to each factor was more spread out and the scores more normally distributed. The Kaiser criterion suggested 36

 doi:10.20944/preprints202007.0106.v1



significant components, however, since most of these showed no obvious dietary associations and were thus not helpful for later analysis, we instead examined the scree plot, which suggested a cutoff with variance explained > 2%, and seven principal components were retained. These cumulatively explained 25.8% of the variation in inferred dietary proportions in the cohort.

Geographic projection was performed using the "leaflet" package (https://rstudio.github.io/leaflet/) in R. Zip codes were combined into 23 zones based on geographical proximity, division of census tracts, similarity of neighborhoods in terms of sociodemographic profile and grocery store density, so that the number of individuals in each zone was roughly equal.

### *2.3 Polygenic score assessment*

Genotyping of 423 individuals in our sample was performed on genomic DNA extracted from whole blood samples using either the HumanCoreExome-12 v.1.1 or HumanOmniExpress-12 v1.1 genotyping Illumina arrays [30]. Imputation was performed using IMPUTE v2. software [31] with 1000 Genomes data, resulting in 8,242,192 imputed SNPs. A polygenic score for BMI (PGSBMI) was calculated using the linear scoring function in PLINK v2.0 [32] using reference GWAS data accessed from the EBI GWAS catalog (https://www.ebi.ac.uk/gwas/publications/30108127). The score includes 281 of the 289 SNPs reported in [33] with association p-values ranging from $2\times10^{-210}$ to $9\times10^{-6}$. BMI and prevalence of obesity defined as BMI >= 30 were plotted against polygenic risk scores for 404 participants with all necessary data available, to see if these SNPs correlate with the BMI trait in our cohort. Simple linear regression was performed to test whether PRSBMI has significant effects on the major Principal Components of dietary variation. Similarly, PGSWHR for 402 people was derived with 307 of the 316 independent SNPs in [34], each with p-values ranging from $5 \times 10^{-183}$ to $5 \times 10^{-9}$ and accessed at https://www.ebi.ac.uk/gwas/studies/GCST008996. Obesity in this case was defined as WHR >= 0.9 for males, or WHR >= 0.85 for females.

### *2.4 Statistical Analyses*

Statistical analyses were performed in JMP Pro 14.3 (SAS Institute, Cary, NC). The distributions of the first 3 diet principal components were assessed and described by the sociodemographic characteristics (i.e. gender, age, race/ethnicity, education, income, marital status and zone of residence). Differences in the means of the PCs between genders were assessed by 2-tailed t-test assuming equal variance. Differences in PC distributions among levels of other categorical/ordinal variables were first assessed by one-way ANOVA. For variables that have an intrinsic linear nature (i.e. age group, household income level, education level), we then used the orders of the categories as "dummy numerical variables" to perform linear regression. In order to further investigate the relative effect sizes of the sociodemographic characteristics on the PCs, we then performed multivariable linear regression with the 6 variables described above. Associations between health and diet were measured by Pearson correlation between each PC and each continuous measure of physical, metabolic or mental health. ANOVA was used to evaluate associations with clinical illness by categorizing the participants into 6 health groups (obese, hypertensive, diabetic, and combinations thereof, as well as controls.)

## 3. Results

### *3.1 Principal Components of Dietary Proportions*

The food items that load most strongly onto the first seven principle components are listed in Table 2. Each PC captures different aspects of overall diet that we subjectively classify into vegetarian (PC1), expensive (PC2), high carb (PC3), soups (PC4), juices or typical diet foods (PC5), and fish based diets (PC6), with corresponding negative loadings for unhealthy Western food items, inexpensive processed goods, high protein, breakfast, commonly consumed items, and red meat, respectively, while PC7 is more difficult to categorize. PC1 in particular might alternatively be conceptualized as capturing a healthy diet including a large proportion of fruits and vegetables. PC1 also has a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 July 2020 doi:10.20944/preprints202007.0106.v1



significant linear relationship with Healthy Eating Index-2015 (R2 = 0.354, p <0.0001), further corroborating PC1's representation of the healthy versus fast food eating axis.

**Table 2.** Food items contributing to top seven principal components of diet proportions

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
|---|---|---|---|---|---|---|---|
| | Vegetarian vs. Western | Expensive vs. Processed | High carb vs. Ketogenic | Soups vs Breakfast | Diet vs Satiety | Fish/seafood vs. Meat | Unclear vs. Grilled |
| 1 | Pea Soup | Wine | Meat subst. | Other noodles | Pumpkin pie | Fried fish | Not fried fish |
| 2 | Other veggies | Beer | Other bread | Potatoes | Menudo | Tuna | Mixed chicken |
| 3 | Spinach cooked | Coffee | Jelly | Pinto beans | Real juice | Meat subst. | Tuna |
| 4 | Apples pears | Veal | Cold cereal | Veggie stew | Oysters | Greens | Other noodles |
| 5 | Veggie stew | Liquor | Bagel | Rice | Fried fish | Shellfish | Veal |
| 6 | Oranges | Oysters | Pizza | Pea soup | Chicken feet | Salty snacks | Other pie |
| 7 | Tofu | Cheese | Cookies | Refried beans | Orange juice | Butter | Wine |
| 8 | Peaches | Bologna | Refried beans | Mix beef/pork | Some juice | Mustard | Liver |
| 9 | Strawberries | Steak | Spaghetti | Veg soup | Break sand | Cracker | Other soup |
| 10 | Carrots | Tacos | Ice cream | Tofu | Liver | Not fried fish | Not fried chick |
| 11 | Other fruit | Salsa | Milk | Other soup | Tomato juice | Salad dressing | Cookies |
| 12 | Not fried fish | Other eggs | Choco Candy | Coleslaw cab | Water | Nuts | Cracker |
| 13 | Tomatoes | Tomatoes | Peanut butter | Spaghetti | Other eggs | Mayo | Coffee |
| 14 | Broccoli | Not fried fish | Breakfast bars | Other veggies | Diet shakes | Cookies | Yogurt |
| 15 | Water | Tomato juice | Power bars | Tacos | Tofu | Spinach | Pumpkin pie |
| | | | | …….. | | | |
| 96 | Salty snacks | Carrots | Beans peas | Diet shakes | Mixed chick | Ribs | Salty snacks |
| 97 | Macaroni | Some juice | Shellfish | Breakfast bars | Margarine | Veal | Sausage |
| 98 | Sausage | Potatoes | Fried fish | Nuts | Iced tea | Cold cereal | Beer |
| 99 | Cake | Greens | Hotdog | Break sand | Broccoli | Other fruit | Watermelon |
| 100 | Mix beef/pork | Cooked cereal | Not fried chick | Not fried chick | Other fruit | Rice | Tacos |
| 101 | Pork | Other fruit | Steak | Cold cereal | Salad dressing | Burger | Break sand |
| 102 | Steak | Cake | Fried chicken | Cantaloupe | Potatoes | Menudo | Refried beans |
| 103 | Hotdog | Cornbread | Sausage | Bacon | Tacos | Steak | Burger |
| 104 | Biscuits | Corn | Liver | Yogurt | Ice cream | Pork | Buns |
| 105 | Bacon | Cookies | Feet | Peanut butter | Beans peas | Mix beef/pork | Hotdog |
| 106 | Donut | Hic | Greens | Banana | Bologna | Oranges | Greens |
| 107 | Fried chicken | Fried fish | Pork | Water | Strawberries | Meatloaf | Fries |
| 108 | Buns | Watermelon | Bacon | Salad dressing | Soft drinks | Apples pears | Pinto beans |
| 109 | Burger | Real juice | Ribs | Other eggs | Not fried chick | Milk | Mustard |
| 110 | Fries | Canned fruit | Coleslaw | Green salad | Green salad | Banana | Salsa |

*3.2 Geographic and cultural associations with the principal components*

We next explored how these dietary components vary with respect to geographic and socioeconomic factors characteristic of Atlanta. Figure 2 shows very strong geographic structure to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 July 2020 doi:10.20944/preprints202007.0106.v1



food eating tendencies, with PC1 and PC2 as exemplars. Broadly speaking, PC1 tracks with wealth, being higher in the more affluent regions of Midtown, Decatur (near-east of Atlanta), and the upper and middle class suburbs of Roswell and Marietta. PC2 is markedly divided between north and south Atlanta, likely reflecting access to fresh and more expensive foods in the north, and higher prevalence of food deserts in the south. This distinction also tracks with the historical segregation of Atlantans by ancestry. All five PC are highly significantly differentiated by region (ANOVA, $p<10^{-5}$).

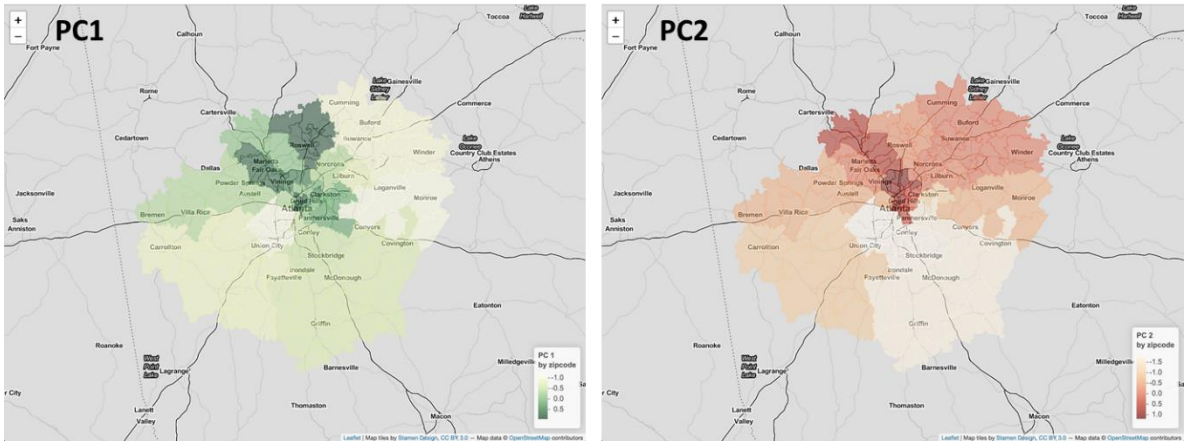


**Figure 2**. Geographic distribution of principal components of food consumption in Atlanta. Regions of greater metropolitan Atlanta are colored with respect to PC scores (positive values have stronger colors) according to the mean value in regions of the city. Midtown is at the center of each Figure, where the major highways converge.

Orthogonally, we also performed regression analysis to evaluate dependence of the three largest PC, each of which explains over 2% of the food item variance, for each of the social factors gender, self-reported race and ethnicity, age group, education level, household income, and marital status. Age was grouped by 10-year intervals, education level was categorized as high school or less (6-12 yrs), some college or college graduates (13-16 yrs), graduate school (17-22 yrs) or post graduate school (>22 yrs), wealth was binned in $25,000 or $50,000 increments as shown, and marital status was categorized as single, married or divorced by excluding 8 widowed individuals. The bins were assigned increasing numerical values for evaluation of the significance of the regression, with the exception of race/ethnicity which was evaluated by ANOVA. Salient results are presented in Table 3.

Overall, dietary patterns varied along the socioeconomic gradient and self-reported race was the most strongly portioned among the 7 tested variables ($P < 0.0001$ for PC 1-5). The healthy-eating PC1 was observed to be higher in females (mean $PC1_{Female} = 0.41$, mean $PC1_{Male} = -0.78$), Asians (2.58 vs 0.01 for European and -0.62 for African American), and generally increased for participants with higher education level or higher household income. A particularly strong gradient by income was also observed for PC2, confirming inference from the geographic analysis.

Multivariate analysis indicated that gender dominates the association with PC1, but age as well as race and ethnicity independently contribute as well ($p<0.0001$ each category). Furthermore, race/ ethnicity, gender and income level have significant independent influences on PC2, whereas only race/ethnicity and age are associated with PC3 in the multivariate analysis. Neither education level nor marital status were significant when analyzed alongside the other variables.

**Table 3.** Association of dietary PC with geographic and socioeconomic variables

| Characteristic | PC1 | PC2 | PC3 |
|---|---|---|---|

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 July 2020 doi:10.20944/preprints202007.0106.v1



| | N (%) | Mean (95% CI) | p | Mean (95% CI) | p | Mean (95% CI) | p |
|---|---|---|---|---|---|---|---|
| **Gender** | | | | | | | |
| Male | 234 (34) | -0.78 (-1.17,-0.39) | *** | 0.66 (0.41, 0.92) | *** | 0.19 (-0.08, 0.47) | 0.03* |
| Female | 453 (66) | 0.41 (0.13, 0.69) | | -0.35 (-0.53, 0.16) | | -0.10 (-0.28, 0.07) | |
| **Race/Ethnicity** | | | | | | | |
| White, non-Hispanic | 487 (71) | 0.01 (-0.25, 0.27) | *** | 0.68 (0.52, 0.83) | *** | 0.35 (0.19, 0.52) | *** |
| Black, non-Hispanic | 151 (22) | -0.62 (-1.15,-0.10) | | -2.09 (-2.35, -1.84) | | -1.19 (-1.47, -0.91) | |
| Asian | 35 (5) | 2.58 (1.42, 3.73) | | -0.61 (-1.16, -0.07) | | -0.06 (-1.05, 0.92) | |
| Hispanic | 9 (1) | 0.39 (-2.57, 3.36) | | 0.39 (-1.34, 2.12) | | 0.96 (-1.14, 3.06) | |
| Amer. Indian/Alaska Native | 5 (1) | -0.76 (-4.49, 2.98) | | 0.16 (-1.43, 1.76) | | -0.04 (-1.87, 1.80) | |
| **Age group** | | | | | | | |
| < 35 yrs | 100 (15) | -0.80 (-1.37, -0.22) | *** | -0.01 (-0.42, 0.39) | 0.37 | 0.62 (0.23, 1.01) | ** |
| 36~45 yrs | 174 (25) | -0.71 (-1.15, -0.26) | | -0.02 (-0.34, 0.29) | | 0.23 (-0.08, 0.53) | |
| 46~55 yrs | 227 (33) | 0.45 (0.02, 0.88) | | -0.09 (-0.36, 0.18) | | -0.30 (-0.57, -0.03) | |
| 56~65 yrs | 164 (24) | 0.49 (0.04, 0.94) | | 0.04 (-0.27, 0.35) | | -0.09 (-0.37, 0.20) | |
| > 66 yrs | 24 (3) | 0.87 (-0.20, 1.95) | | 0.81 (-0.15, 1.78) | | -0.83 (-1.61, -0.05) | |
| **Education level** | | | | | | | |
| High School or Less | 19 (3) | -0.65 (-1.70, 0.39) | 0.04* | -1.82 (-2.99, -0.66) | *** | -0.98 (-1.97, 0.01) | 0.01* |
| College | 263 (38) | -0.23 (-0.60, 0.14) | | -0.48 (-0.73, -0.23) | | -0.23 (-0.47, 0.00) | |
| Graduate | 257 (37) | -0.07 (-0.44, 0.31) | | 0.26 (0.02, 0.49) | | 0.20 (-0.04, 0.44) | |
| Post-Graduate | 148 (22) | 0.62 (0.08, 1.16) | | 0.62 (0.30, 0.93) | | 0.17 (-0.17, 0.52) | |
| **Household Income level** | | | | | | | |
| $0 to $50,000 | 72 (11) | -0.99 (-1.75, -0.23) | 0.01* | -1.18 (-1.66, -0.70) | *** | -0.37 (-0.85, 0.10) | 0.22 |
| $50,000 to $75,000 | 93 (14) | -0.13 (-0.79, 0.53) | | -0.65 (-1.04, -0.25) | | -0.10 (-0.53, 0.32) | |
| $75,000 to $100,000 | 91 (14) | -0.07 (-0.62, 0.49) | | -0.46 (-0.89, -0.03) | | -0.26 (-0.70, 0.18) | |
| $100,000 to $150,000 | 154 (24) | -0.12 (-0.55, 0.31) | | 0.10 (-0.17, 0.37) | | 0.17 (-0.12, 0.47) | |
| $150,000 to $200,000 | 80 (12) | 0.67 (-0.02, 1.35) | | 0.74 (0.32, 1.15) | | 0.22 (-0.22, 0.67) | |
| $200,000 to $250,000 | 44 (7) | 0.06 (-0.96, 1.08) | | 0.28 (-0.35, 0.92) | | -0.34 (-0.95, 0.27) | |
| $250,000 to $300,000 | 26 (4) | -1.02 (-2.25, 0.20) | | 0.23 (-0.53, 0.99) | | 0.08 (-0.68, 0.84) | |
| Above $300,000 | 85 (13) | 0.67 (-0.03, 1.37) | | 1.34 (0.89, 1.79) | | 0.27 (-0.15, 0.69) | |
| **Marital Status** | | | | | | | |
| Single | 157 (23) | -0.72 (-1.17, -0.26) | 0.01* | -0.06 (-0.35, 0.22) | 0.01* | 0.20 (-0.11, 0.51) | 0.22 |
| Married | 436 (64) | 0.22 (-0.07, 0.51) | | 0.13 (-0.06, 0.33) | | -0.02 (-0.21, 0.16) | |
| Divorced | 88 (13) | 0.03 (-0.71, 0.77) | | -0.60 (-1.00, -0.11) | | -0.26 (-0.72, 0.21) | |

Note. ***P <.0001, **P < .001, *P <.05

### 3.3 *Health correlates with dietary principal components*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 July 2020 doi:10.20944/preprints202007.0106.v1



The relationship between dietary patterns and health was investigated by correlating the main dietary principal components with the participants' clinical profiles. A total of 45 health related traits were examined, including 3 physical traits, 19 items from a comprehensive metabolic panel (CMP), 8 items from the complete blood count (CBC) test, 4 mental health related measures and 8 overall summary scores from the SF36 Quality of Life Survey. Pearson linear and Spearman rank correlations gave very similar results, and we report the Pearson's correlations for 15 traits with particularly strong correlations to specific PC in Table 4, in which significant positive correlations are shaded red and negative blue.

**Table 4.** Association of dietary principal components with health outcomes

| | PC 1 | PC 2 | PC 3 | PC 4 | PC 5 | PC 6 | PC 7 |
|---|---|---|---|---|---|---|---|
| Basal Metabolic Rate | -0.40 | 0.11 | -0.06 | 0.04 | -0.06 | 0.02 | -0.26 |
| BMI | -0.29 | -0.08 | -0.24 | -0.04 | -0.09 | 0.16 | -0.19 |
| Waist-to-Hip Ratio | -0.22 | 0.13 | -0.05 | 0.09 | 0.03 | -0.04 | -0.07 |
| Beck Depression Index (BDI) | -0.19 | 0.04 | 0.04 | 0.00 | 0.04 | 0.07 | 0.02 |
| Systolic Blood Pressure | -0.15 | -0.04 | -0.15 | 0.01 | -0.04 | 0.05 | -0.06 |
| Diastolic Blood Pressure | -0.13 | 0.01 | -0.09 | 0.03 | -0.05 | 0.00 | -0.03 |
| Body Fat Percent | -0.13 | -0.18 | -0.30 | -0.10 | -0.11 | 0.26 | -0.10 |
| Perceived Stress Scale Score | -0.12 | 0.04 | 0.09 | 0.02 | 0.00 | 0.00 | 0.00 |
| General Anxiety Survey-7 Score | -0.11 | 0.12 | 0.08 | -0.05 | 0.02 | 0.04 | 0.04 |
| Epworth Sleepiness Scale Score | -0.11 | -0.20 | -0.04 | -0.02 | 0.01 | 0.05 | 0.01 |
| Fasting Blood Glucose | -0.10 | 0.10 | -0.14 | 0.05 | -0.02 | 0.02 | 0.00 |
| SF-36: Physical Health Score | 0.11 | 0.09 | 0.15 | 0.05 | 0.01 | -0.08 | 0.02 |
| SF-36: Mental Health Score | 0.15 | -0.12 | -0.07 | -0.01 | -0.01 | -0.02 | -0.05 |
| SF-36: Vitality Score | 0.23 | -0.02 | -0.02 | 0.08 | 0.03 | -0.07 | 0.00 |
| SF-36: General Health Score | 0.24 | -0.05 | 0.00 | -0.05 | 0.01 | -0.04 | 0.08 |

Most notable in this analysis is the strong correlation between PC1 and most health measures, indicating the expected positive impact of a healthier diet in general as well as specific aspects of physical and mental well-being including vitality. Note that negative correlations are due to the association of larger values of sleep, anxiety and depression scores, weight and blood pressure, and basal metabolic rate, with poor health. Perhaps surprisingly, PC2 which captures a more expensive diet, is also negatively correlated with BMI and body fat percent, and an association of a high carb diet (PC3) with reduced body weight was seen. Consumption of inexpensive processed foods implied by high values of PC2 is very clearly associated with elevated waist-to-hip ratio, and mildly with mental health concerns. Notably, and also unexpectedly, the high fish diet implied by PC6 strongly correlates with high body fat percent and BMI

These results are consistent with diet being a major contributor to chronic disease. To further investigate this, we next performed a categorical analysis designed to evaluate whether obese individuals (BMI >= 30 kg/m2), hypertensives (blood pressure greater than 140/90 mmHg), and diabetics (mean blood sugar over 126 mg/dL) had abnormal PC scores. These clinical conditions are all pharmacologically controlled in the study participants. Joint incidence of all three

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 July 2020 doi:10.20944/preprints202007.0106.v1



conditions was observed in 15 individuals, and for one or two of the conditions in one third of the samples, leaving approximately two-thirds of the CHDWB cohort classified as relatively healthy controls. Figure 3 shows marked differences in the first three PC with respect to these three chronic conditions, with effects in the expected direction. Healthy status associates with high values of PC1 and PC3, but there was surprisingly little differentiation with respect to PC2. Individuals with all three conditions tend to have the most extreme dietary consumption patterns. We caution against over-interpretation of individual comparisons due to small sample size of some categories and presence of confounding variables.

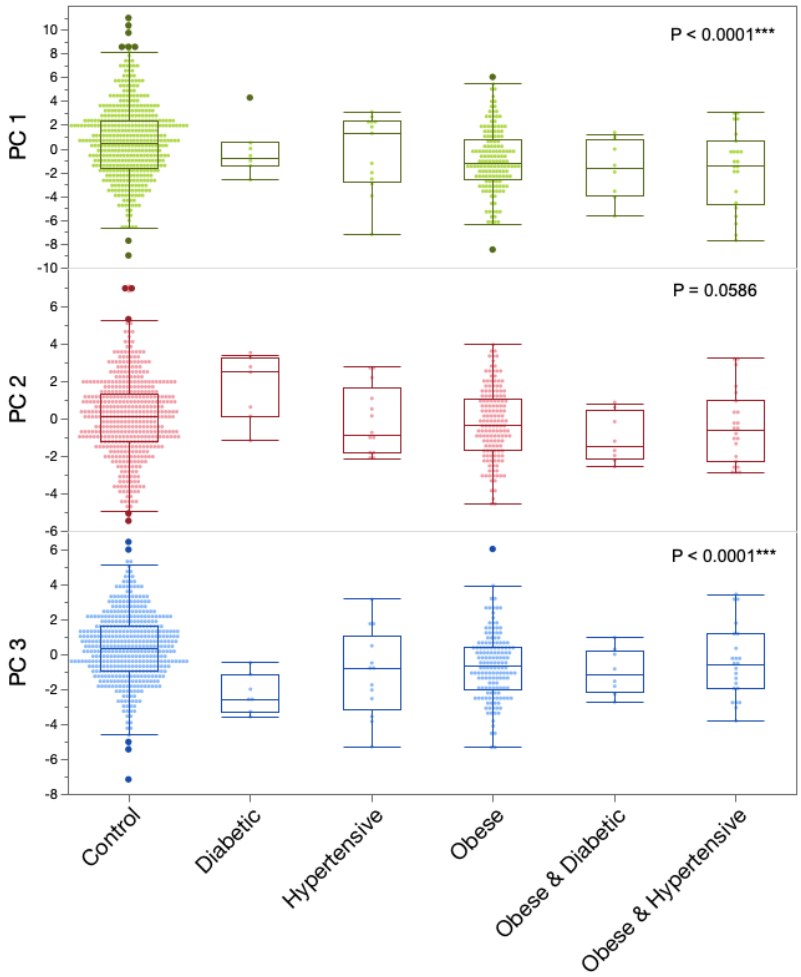


**Figure 3**. Association of dietary principal components with chronic health conditions.

### *3.4 Polygenic association of BMI with health-conscious dietary preferences*

In order to assess whether genetic variation for body weight might act through dietary preference, we evaluated polygenic scores (PGS) for BMI and WHR using independent genotype weights for 281 and 307 SNPs respectively and weights ascertained by the contributing studies [33,34]. Imputed whole genome genotypes were available for 410 individuals, and despite the small sample size, the expected positive correlations between PGSBMI and BMI (Figure 4A) and prevalence of obesity (Figure 4B) were clearly observed. The polygenic score explains approximately 3% of the variance for BMI after excluding a handful of individuals with extreme BMI over 40. Each point in Figure 4B represents 41 individuals and there is a clear trend for increasing proportion of obese individuals as the PGS decile increases.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 July 2020 doi:10.20944/preprints202007.0106.v1



There does not appear to be any association between polygenic predisposition to BMI and dietary PC1 (Figure 4C) or any of the other PC, with the marked exception of PC5 (Figure 4D) where the regression explains 2.8% of the variance (p=0.0005), being approximately as strong as the association with BMI. Consideration of the loadings on PC5 suggests that high values reflect consumption of fresh juices, diet shakes and other items typically consumed by dieting persons, whereas positive ones might be related to more filling foods like ice cream, soft drinks and tacos. The negative association of PC5 with BMIPRS is consistent with the interpretation that polygenic risk for obesity is mediated through health-conscious eating behaviors and satiety. No associations with PGSWHR were observed.

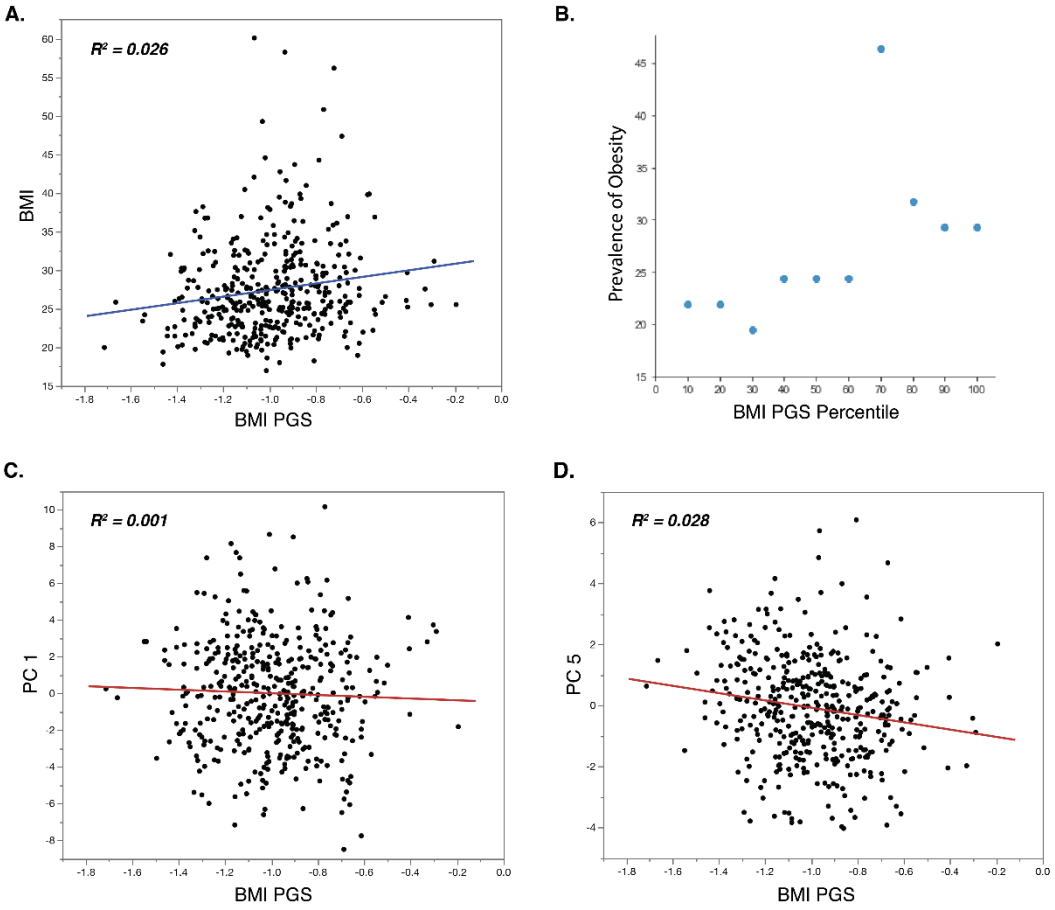


Figure 4. Association between dietary categories, body weight, and polygenic scores (A) BMI as a function of BMI polgenic score. (B) prevalence of obesity (BMI>=30) in ten decile binds from low to high. (C) Lack of association between PC1 and the BMI polygenic score, cf (D) strong association with PC5 (p<0.0001).

## 4. Discussion

This study investigated the correlation of four major types of non-dietary factors with people's diet, namely geography, socioeconomic standing, health status, and genetics. Few studies have considered combinations of these subjects together, despite the general acceptance of the concept that diet is one of the major factors that connects SES with health. Our results are entirely consistent with the supposition that a healthy diet strongly associates with education, income, and access to quality food, with highly significant health outcomes as argued by [35,36]. In the context of Atlanta, a cosmopolitan city with a large African American population, it is also evident that these cultural factors are confounded with race and ethnicity. Our results regarding the geographic distribution of dietary patterns are also consistent with descriptions of unique features related to so-called food deserts, where access to food is dominated by dollar stores, gas station food marts or fast food

establishments. It is evident that food access and environment has a major effect on the resident's diet choices, but confounded by disparities due to racial and SES factors makes it difficult to parse specific contributions.

In contrast to the conventional dietary analytic approach focusing on a single nutrient or a summary score, in this study a regression-based methodology based on principal components was applied to food proportions computed from dietary surveys. We show that this allows association of specific aspects of dietary choice with other variables. The PCs identified with our approach represent the most significant dimensions of eating behavior, and were specifically designed to capture proportions of consumed food items rather than overall consumption amounts. The reproducibility and validity of this approach was initially discussed by [7]. We note that recent machine learning approaches may reveal stronger and more consistent clusters of dietary patterns, and that their utility in nutrition research is just beginning to be tapped [37].

The present data reveal significant relationships between dietary patterns and health status, supporting a link between healthy eating habits and well-being. Intuitively, what we eat can affect our physical health, and this is apparent in the trend for people having a balanced diet with plenty of fruits and vegetables being at lower risk for diet-related disease and "healthier". However, health status, or the perception of health status, is likely to reciprocally impact dietary choice as well: one example is that people with diabetes tended to consume less sugary food than the remainder of the cohort. It is in general difficult to ascribe the direction of causality to any of the described relationships.

One caveat of the study was that the Block FFQ is only semi-quantitative, is biased by self-recall of eating habits, and does not survey subtle but important distinctions such as types of salad dressing and kinds of cooking oil. It does include calibration questions needed for computation of nutritional intake calculations, but we elected instead to focus on food group proportions and so these did not contribute to the analyses. Nevertheless, the approach is used widely in the nutrition literature and is accepted to capture broad trends in food consumption.

Recently, a number of large-scale genome-wide association studies have begun to attribute genetic factors to BMI, WHR and obesity [33,34] as well as to patterns of dietary consumption. A GWAS on 85 single food intake and 85 principal components of diet in FFQ data for the UK Biobank [38], identifying 136 associations specific to dietary choices such as white versus wholemeal/wholegrain bread consumption. Many of these link to olfactory receptor associations for example with fruit and tea intake, but Mendelian randomization failed to adduce strong evidence for a causal role in coronary artery disease or diabetes. We provide preliminary evidence that one component of dietary intake, PC5, which possibly captures a measure of health-conscious eating, is significantly correlated with a polygenic score for BMI. This finding is consistent with the enrichment of neuronally expressed genes in the BMI GWAS loci and the notion that this PGS mediates its effect in part through the propensity to diet. Our results also indicate how important it will be to control genetic analyses for cultural and socioeconomic confounders which are major mediators of dietary behavior.

**Author Contributions:** Conceptualization of the CHDWB program, K.L.B., T.R.Z, M.L., D.P.J and A.A.Q.; statistical methodology, J.C. and X.H.; project administration, J.C.; program directorship, K.L.B and G.S.M.; manuscript preparation, G.G., J.C. and X.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding. The CHDWB is funded by the Predictive Health Institute of Georgia Institute of Technology and Emory University.

**Acknowledgments:** We are particularly grateful for the participation of all members of this study and their consent for publication of this data.

**Conflicts of Interest:** "The authors declare no conflict of interest."

## References


1. Randall, E.; Marshall, J.; Graham, S.; Brasure, J. Frequency of food use data and the multidimensionality of diet. J. Am. Dietetic Assoc. 1989, 89, 1070-1075.
2. National Research Council. Diet and Health: Implications for Reducing Chronic Disease Risk. 1989. National Academies Press. http://www.nap.edu/catalog/1222.html
3. James, W.P.T.; Nelson, M.; Ralph, A.; Leather, S. Socioeconomic determinants of health: the contribution of nutrition to inequalities in health. BMJ, 1997, 314, 1545-1549.
4. Psaltopoulou, T.; Hatzis, G.; Papageorgiou, N.; Androulakis, E.; Briasoulis, A.; Tousoulis, D. Socioeconomic status and risk factors for cardiovascular disease: impact of dietary mediators. Hellenic J Cardiol. 2017, 58, 32-42.
5. Wardle, J.; Cooke, L. Genetic and environmental determinants of children's food preferences. Brit. J. Nutr. 2008, 99, S15-S21.
6. Kaur, A.; Scarborough, P.; Rayner, M. A systematic review, and meta-analyses, of the impact of health-related claims on dietary choices. Int J. Behav. Nutr. Phys. Act. 2017, 14, 93.
7. Block. G.; Hartman, A.M.; Dresser, C.M.; Carroll, M.D.; Gannon, J.; Gardner L. A data-based approach to diet questionnaire design and testing. Am J Epidemiol. 1986, 124, 453-469.
8. Harlan LC, Block G. Use of adjustment factors with a brief food frequency questionnaire to obtain nutrient values. Epidemiology. 1990, 1, 224-231.
9. Guenther, P.M.; Reedy, J.; Krebs-Smith, S.M. Development of the Healthy Eating Index-2005. J. Am. Dietetic Assoc. 2008, 108, 1896-1901.
10. Krebs-Smith, S.M.; Pannucci, T.E.; Subar, A.F.; Kirkpatrick, S.I.; Lerman, J.L.; Tooze, J.A.; Wilson, M.M.; Reedy, J. Update of the Healthy Eating Index: HEI-2015. J Acad Nutr Diet, 2018, 118, 1591-1602.
11. Chiuve, S.E.; Fung, T.T.; Rimm, E.B.; Hu, F.B.; McCullough, M.L.; Wang, M.; Stampfer, M.J.; Willett, W.C. Alternative dietary indices both strongly predict risk of chronic disease. J. Nutr. 2012, 142, 1009-1018.
12. Hu, F.B. Dietary pattern analysis: a new direction in nutritional epidemiology. Curr. Opin. Lipidol. 2002, 13, 3-9.
13. McCann, S.E.; Marshall, J.R.; Brasure, J.R.; Graham, S.; Freudenheim, J.L. Analysis of patterns of food intake in nutritional epidemiology: food classification in principal components analysis and the subsequent impact on estimates for endometrial cancer." Public Health Nutr. 2001, 4, 989-997.
14. Hu, F.B.; Rimm, E.; Smith-Warner, S.A.; Feskanich, D.; Stampfer, M.J.; Ascherio, A.; Sampson, L.; Willett WC. Reproducibility and validity of dietary patterns assessed with a food frequency questionnaire. American J. Clin. Nutr. 1999, 69, 243-249.
15. Millen, B.E.; Quatromoni, P.A.; Copenhafer, D.L.; Demissie, S.; O'Horo, C.E.; D'Agostino, R.B. Validation of a dietary pattern approach for evaluating nutritional risk: the Framingham Nutrition Studies. J Am Diet Assoc. 2001, 101, 187-194.
16. Quatromoni, P.A.; Copenhafer, D.L.; Demissie, S.; D'Agostino, R.B.; O'Horo, C.E.; Nam, B.H.; Millen, B.E. The internal validity of a dietary pattern analysis. The Framingham Nutrition Studies. J. Epidemiol. Community Health. 2002, 56, 381-388.
17. Brigham KL. Predictive health: the imminent revolution in health care. J Am Geriatr Soc. 2010, 58 Suppl 2, S298-302.
18. Tabassum, R.; Cunningham, L.; Stephens, E.H.; Sturdivant, K.; Martin, G.S.; Brigham, K.L.; Gibson, G. A longitudinal study of health improvement in the Atlanta CHDWB wellness cohort. J. Pers. Med. 2014, 4, 489-507.
19. Al Mheid, I.; Kelli, H.M.; Ko, Y.A.; Hammadah, M.; Ahmed, H.; Hayek, S.; Vaccarino, V.; Ziegler, T.R.; Gibson, G.; Lampl, M.; Alexander, R.W. Brigham, K.L., Martin, G,S.; Quyyumi, A.A. Effects of a health-partner intervention on cardiovascular risk. J Am Heart Assoc. 2016, 5, e004217.
20. Bettermann, E.L.; Hartman, T.J.; Easley, K.A.; Ferranti, E.P.; Jones, D.P.; Quyyumi, A.A.; Vaccarino, V.; Ziegler, T.R.; Alvarez, J.A. Higher Mediterranean diet quality scores and lower body mass index are associated with a less-oxidized plasma glutathione and cysteine redox status in adults. *J. Nutr.* **2018**, *148*, 245-253.
21. Bellissimo, M.P.; Cai, Q.; Ziegler, T.R.; Liu, K.H.; Tran, P.H.; Vos, M.B.; Martin, G.S.; Jones, D.P.; Yu, T.; Alvarez, J.A. Plasma high-resolution metabolomics differentiates adults with normal weight obesity from lean individuals. *Obesity* **2019**, *27*, 1729-1737.

22. Rask, K.J.; Brigham, K.L.; Johns, M.M.E. Integrating comparative effectiveness research programs into predictive health: a unique role for academic health centers. Acad. Med. 2011, 86, 718-723.
23. Beck, A.T.; Steer, R.A. Internal consistencies of the original and revised Beck depression inventory. J. Clin. Psych. 1984, 40, 1365-1367.
24. Spitzer, R.L.; Kroenke, K.; Williams, J.B.W.; Löwe, B. A brief measure for assessing generalized anxiety: the GAD-7. Arch Intern Med. 2006, 166, 1092-1097.
25. Cohen, S.; Kamarch, T.; Mermelstein, R. A global measure of perceived stress. J. Health Soc. Behav. 1983, 24, 385-396.
26. Johns, M.W. A new method for measuring daytime sleepiness: the Epworth Sleepiness Scale. Sleep 1991, 14, 540-545.
27. Ware, J.E.; Gandek, B. Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project. J. Clin. Epidem. 1998, 51, 903-912.
28. US Department of Health and Human Services; US Department of Agriculture. 2015-2020 Dietary Guidelines for Americans. 8th ed. Washington, DC: US Dept of Health and Human Services; 2015. http://www.health.gov/DietaryGuidelines.
29. Malan, L.; Smuts, C.M.; Baumgartner, J.; Ricci, C. Missing data imputation via the expectation-maximization algorithm can improve principal component analysis aimed at deriving biomarker profiles and dietary patterns. Nutrition Research. 2020, 75, 67-76.
30. Wingo, A.P.; Gibson, G. Blood gene expression profiles suggest altered immune function associated with symptoms of generalized anxiety disorder. Brain Behav Immun. 2015, 43, 184-191.
31. Howie, B.N.; Donnelly, P.; Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet., 2009, 5, e1000529.
32. Chang, C.C.; Chow, C.C.; Tellier, L.C.; Vattikuti, S.; Purcell, S.M.; Lee, J.J. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2015, 4, 7.
33. Hoffmann, T.J.; Choquet, H.; Yin, J.; Banda, Y.; Kvale, M.N.; Glymour, M.; Schaefer, C.; Risch, N.; Jorgenson, E. Multiethnic genome-wide association study of adult body mass index identifies novel loci. Genetics 2018, 210, 499-515.
34. Pulit, S.L.; Stoneman, C.; Morris, A.P.; Wood, A.R.; Glastonbury, C.A.; et al. Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. Hum Mol Genet. 2019, 28, 166-174.
35. Patrick, H.; Nicklas, T.A. A review of family and social determinants of children's eating patterns and diet quality. J. Am. Coll. Nutrition 2005, 24, 83-92.
36. Link, B.G.; Phelan, J. Social conditions as fundamental causes of disease. J. Health Soc. Behav. 1995, Spec. No., 80-94.
37. Reis, R.; Peixoto, H.; Machado, J.; Abelha, A. Machine-learning in nutritional follow-up research. Open Comput. Sci. 2017, 7, 41-45.
38. Cole, J.B., Florez, J.C., Hirschhorn, J.N. Comprehensive genomic analysis of dietary habits in UK Biobank identifies hundreds of genetic associations. Nat. Commun. 2020, 11, 1467.