Journal of Data Science xx (xx), 1–20 DOI: 0000
July 2022

HEI Analysis of NHANES Dietary Data: Exploring the Diet Quality of Americans with R Package heiscore

Berkeley Ho[1,†], Vijetha Ramdas[1,†], and Abhra Sarkar[1]

[1]*Department of Statistics and Data Sciences, The University of Texas at Austin, TX 78705, USA*

Abstract

Connections between subpar dietary choices and negative health consequences are well established in the field of nutritional epidemiology. Consequently, in the United States, there is a standard practice of conducting regular surveys to evaluate dietary habits. One notable example is the National Health and Nutrition Examination Survey (NHANES) conducted every two years by the Center for Disease Control (CDC). Several scoring methods have been developed to assess the quality of diet in the overall population as well as different pertinent subgroups using dietary recall data collected in these surveys. The Healthy Eating Index (HEI) is one such metric, developed based on recommendations from the United States Department of Health and Human Services (HHS) and Department of Agriculture (USDA) and widely used by nutritionists. Presently, there is a scarcity of user-friendly statistical software tools implementing the scoring of these standard scoring metrics. Herein, we develop an R package heiscore to address this need. Our carefully designed package, with its many user-friendly features, increases the accessibility of the HEI scoring using three different methods outlined by the National Cancer Institute (NCI). Additionally, we provide functions to visualize multidimensional diet quality data via various graphing techniques, including bar charts and radar charts. Its utility is illustrated with many examples, including comparisons between different demographic groups.

Keywords *diet quality scores; dietary intake; Healthy Eating Index (HEI); National Health and Nutrition Examination Survey (NHANES); visualization tools.*

1 Introduction

The evaluation of diet quality is crucial in assessing the nutritional health of individuals and populations, as well as for comparing the nutritional trends across various sub-populations of interest such as different ethnicities and age groups. One way to do this is by assigning scores to different dietary components and then combining them to get an overall diet quality score. The United States Department of Health and Human Services (HHS) and the Department of Agriculture (USDA) suggest various scoring methods for individuals and populations. They update the Dietary Guidelines for Americans (DGA) every five years to encourage better and healthier eating habits. The Healthy Eating Index (HEI) is a scoring method meticulously crafted to assess diet quality based on adherence to the DGA by utilizing dietary information from National Health and Nutrition Examination Survey (NHANES) data. Despite being widely used in nutritional research, there is a need for sophisticated software tools facilitating the calculation, visualization

*Corresponding author Email: abhra.sarkar@utexas.edu.
†Equal contributions

and comparison of the HEI scores across populations, subgroups, and individuals. The focus of this article is a new `R` package `heiscore` that aims to address some of these needs.

Dietary Health: A healthy eating habit is generally characterized by the consumption of different dietary components and nutrients in appropriate proportions. Carbohydrates, proteins, and unsaturated fats support energy needs. Whole grains and fresh fruit are preferred to processed grains and fruit juices to fulfill such physiological needs. Additionally, the consumption of some components such as saturated fats should be moderated, as their excessive intake can pose health risks. Unhealthy dietary patterns have been linked with increased risk for developing conditions such as hypertension, high cholesterol, obesity, and inflammation which in turn increase the risk for cardiovascular disease, diabetes, and cancer (Cena and Calder, 2020). Older populations are particularly vulnerable to health-related complications when their dietary habits fall short of essential nutritional standards (Sharkey, 2008). Studies of dietary trends are therefore crucial for understanding how they may impact overall health (Hu, 2002).

Table 1: HEI Components

Adequacy Components	Moderation Components
Total Fruits	Refined Grains
Whole Fruits	Sodium
Total Vegetables	Added Sugars
Greens and Beans	Saturated Fats
Whole Grains	
Dairy	
Total Protein Foods	
Seafood and Plant Proteins	
Fatty Acids	

HEI: In summary, The HEI organizes an individual's diet into different components and assigns each component a score relative to overall caloric intake. The individual component scores are then combined to provide a measure of overall dietary health. The total HEI score thus obtained ranges from 0 to 100. The standards to earn the maximum amount of points are based on a diet that is quite lenient towards the subjects while adhering to nutritional standards. Individual components include both adequacy and moderation scores (Table 1), as described in Kennedy et al. (1995) and Kirkpatrick et al. (2018). Adequacy scores gauge whether essential food groups or components, such as whole grains or fruits and vegetables, are sufficiently consumed. Individuals get the highest adequacy score when they meet the minimum criterion outlined by the DGA, and exceeding this criterion does not result in reduced scores. In contrast, moderation scores assess the consumption of components that may be detrimental to health when consumed in large quantities, such as saturated fats or sodium. Consuming these elements excessively therefore lowers the component score to a minimum of zero when surpassing the limits outlined by the DGA. When the intake of moderation or adequacy components falls within the maximum and minimum criteria, scores are assigned proportionally.

Existing Packages: Software resources available to evaluate diet quality using the HEI are presently fairly limited. The `R` package `NHANES` (Pruim, 2015), intended mainly for educational purposes, allows users to retrieve adapted (but not actual) NHANES data and does not focus on dietary data. The package `nhanesA` (Endres et al., 2024) allows users to retrieve data directly

from the NHANES website but does not include the Food Patterns Equivalents Database (FPED) that converts NHANES dietary recalls into data compatible with the HEI scoring methods. The package `hei` (Nagraj and Folsom, 2017) was designed to calculate scores on an individual level, not accounting for survey weights. However, it was archived from the Comprehensive R Archive Network (CRAN) repository. While working on `heiscore` (Ramdas et al., 2024), we also became aware of the package `dietaryindex` (Zhan et al., 2023a) developed recently by Zhan et al. (2023b) which allows for the retrieval of NHANES data and the calculation of various dietary indices, including the HEI, but lacks support for different HEI scoring methods as well as tools for visualizing the multidimensional scores and comparing them between sub-populations.

Our Contributions: The `heiscore` package contributes to the small but expanding collection of software tools available for HEI analysis. Its aims and contributions are threefold. Firstly, it implements three different scoring methods prescribed by the DGA for the overall analysis of the HEI directly from preloaded NHANES data, appropriately incorporating survey weights associated with the sampled individuals. Secondly, for each of these scoring methods, it offers mechanisms to segment the overall sample into different demographic subgroups, enabling a straightforward comparison of the diet quality and patterns between these sub-populations. Finally, it provides user-friendly functions and shiny apps to visualize multidimensional HEI scores via various graphing techniques, including histograms, bar charts, and radar charts.

The package and its documentation are available on CRAN at `https://CRAN.R-project.org/package=heiscore` and GitHub at `https://github.com/abhrastat/heiscore`.

Outline of the Article: The rest of the paper is organized as follows. Section 2 presents the methodology for scoring diets without adjustments to estimate usual intake. Section 3 introduces and illustrates the key functions of the `heiscore` package. Section 4 concludes with a discussion and considerations for future work.

2 HEI Methodology

2.1 NHANES Dietary Data

The data used for the HEI analysis originate primarily from the What We Eat in America (WWEIA) interview component of the NHANES. They comprise two days of participants' detailed dietary intakes and demographic information. The dietary intake data are collected via two 24-hour recalls. In these recalls, which occur between three and ten days of each other, the participants report the types and quantities of food they consumed in the past 24 hours. Per NHANES guidelines, in subsequent analysis we only include individuals who participated in both recalls.

While NHANES surveys occur practically on a continuous basis, the data are grouped into time periods called cycles. Since 1999, NHANES has used two-year cycles so that survey results most accurately reflect the health and eating habits of the U.S. population. The CDC also recommends combining multiple adjacent cycles, such as the 2011-12 and 2013-14 periods, when analyzing groups of small sample sizes such as Hispanic groups or subjects with specific health conditions. Doing so allows for a larger sample coming from 4 or more years of data, enhancing the reliability of the estimates (Chen et al., 2020).

The NHANES sample is intended to be representative of the civilian, non-institutionalized resident U.S. population. Underrepresented cohorts of the population are however oversampled during the survey process to provide more accurate estimates for these subgroups. Sample weights accounting for such oversampling as well as other factors such as nonresponse are available for

each participant's dietary responses in the NHANES data set. The weights are formulated by initially assigning each respondent a base sample weight, determined by the proportion of the population sharing the respondent's demographic characteristics. Subsequent adjustments to these base weights account for the proportion of homes screened in each sampling area and address non-response. Finally, the weights are fine-tuned to ensure that the weighted estimates of demographic groups align with population data sourced from the U.S. Census Bureau. Proper application of these weights is important to create estimates that are representative of the U.S. population and to accurately evaluate patterns in diet quality estimated from the NHANES data.

Table 2: HEI-2020 Scoring Standards

Component	Points	Standard for Maximum Score	Standard for Minimum Score
Adequacy			
Total Fruits	5	$\geqslant$ 0.8 cups per 1000 kcal	0 cups per 1000 kcal
Whole Fruits	5	$\geqslant$ 0.4 cups per 1000 kcal	0 cups per 1000 kcal
Total Vegetables	5	$\geqslant$ 1.1 cups per 1000 kcal	0 cups per 1000 kcal
Greens and Beans	5	$\geqslant$ 0.2 cups per 1000 kcal	0 cups per 1000 kcal
Whole Grains	10	$\geqslant$ 1.5 oz per 1000 kcal	0 oz per 1000 kcal
Dairy	10	$\geqslant$ 1.3 cups per 1000 kcal	0 cups per 1000 kcal
Total Protein Foods	5	$\geqslant$ 2.5 oz per 1000 kcal	0 oz per 1000 kcal
Seafood and Plant Proteins	5	$\geqslant$ 0.8 oz per 1000 kcal	0 oz per 1000 kcal
Fatty Acids	10	$\frac{\text{PUFAs + MUFAs}}{\text{SFAs}} \geqslant 2.5$	$\frac{\text{PUFAs + MUFAs}}{\text{SFAs}} \leqslant 1.2$
Moderation			
Refined Grains	10	$\leqslant$ 1.8 oz per 1000 kcal	$\geqslant$ 4.3 oz per 1000 kcal
Sodium	10	$\leqslant$ 1.1 grams per 1000 kcal	$\geqslant$ 2.0 grams per 1000 kcal
Added Sugars	10	$<$ 6.5% of total kcal	$\geqslant$ 26% of total kcal
Saturated Fats	10	$\leqslant$ 8% of total kcal	$\geqslant$ 16% of total kcal

2.2 HEI Basics

The HEI is a measure of overall diet quality with scoring criteria based on 13 dietary components (Table 1). The FPED, available since the 2005-06 NHANES cycle, converts the raw intakes of foods and beverages from the NHANES to 37 USDA 'Food Patterns' components. Some of these individual Food Patterns components and some combinations of multiple of these Food Patterns components then make up the 13 HEI components. In all our HEI analysis, we obtained the survey weights, demographic information of the participants, and intakes of total kilocalories, Sodium, and Saturated Fat directly from the NHANES data. Additionally, we retrieved the total monounsaturated (MUFA) and polyunsaturated fatty acids (PUFA) to calculate the Fatty Acids HEI component also directly from the NHANES data. The dietary data for all other HEI components are however obtained from the FPED. The subjects were matched between these two data sets using their respondent sequence numbers.

Nine of the HEI food categories are adequacy components that the DGA encourages individuals to consume in sufficient amounts. For these constituents, higher scores correspond to

more substantial relative intake. The remaining four elements are moderation components since the DGA recommends them to only be consumed in moderation. For these components, higher scores indicate better adherence to their recommended limits.

The HEI is revised every five years in accordance with the DGA (Shams-White et al., 2023). The most recent HEI scoring standards, the HEI-2020, summarized in Table 2, delineate the criteria for achieving the maximum score in each of the 13 components for individuals aged 2 years or older. All areas of the diet are considered equally important in the HEI, but certain components only represent half of a dietary category. As a result, these half-groups are assigned a maximum of five points compared the other categories' ten full points. For example, the Total Vegetables and Greens and Beans components constitute the vegetables category. Accordingly, each component is allotted a maximum score of five points. Additionally, a standard to earn a minimum score of zero is defined for each component. Since the HEI measures diet quality rather than quantity, it uses a density approach to determine scores. Consequently, the standards for almost all components are given with respect to 1000 calories of overall intake. For example, the standard for a maximum score for Total Fruits is the consumption of at least 0.8 cups of fruit per 1,000 kcal. The remaining standards are defined by a percent of total calories consumed (for Added Sugars and Saturated Fats) or by the ratio of dietary elements within the category (for Fatty Acids). The maximum scores of all 13 components sum to 100 — the highest possible overall HEI score.

Note that to address the unique dietary needs of toddlers, in 2020, in a departure from past practices, a separate index, namely the HEI-Toddlers-2020, was developed for children between 12 and 23 months (Pannucci et al., 2023). The HEI-Toddlers-2020 includes the same 13 components and their respective maximum point values as the HEI-2020. Additionally, the scoring rules for both indices are the same and therefore are not separately described below. The difference lies in the consumption standards for the maximum and minimum scores. We describe these standards in Table S.1 in the supplementary material. Our package `heiscore` uses these revised scoring standards for individuals under 2 years old.

2.3 HEI Scoring Methods

The following methods have been prescribed by the National Cancer Institute (NCI) for the purpose of scoring the raw intakes (Herrick et al., 2018).

2.3.1 Simple Scoring

The Simple scoring method assigns a score to an individual's diet by first calculating the total intake of each dietary component over two recalls, then constructing ratios using these values, and finally comparing the ratios to the correct sets of HEI standards, either HEI-2020 or HEI-Toddlers-2020 depending on the age of the individual.

When calculating the ratios for the Simple scoring method, there are two groups of components to consider. For the first group, which includes all of the adequacy components but Fatty Acids and also the moderation component Refined Grains, the ratios are determined as the raw amount of intake of a component per one thousand kilocalories consumed without any other adjustments or modifications.

$$\text{Basic Simple Scoring Ratio} = \frac{\text{Raw Amount of Component Consumed}}{\text{Total Kilocalories Consumed}} \times 1000 \text{ Kilocalories.} \quad (1)$$

For the second group, which includes the adequacy component Fatty Acids and the remaining three moderation components, some adjustments are made. Specifically, for the Fatty Acids ratio, the consumption of total MUFA and PUFA is compared to the total consumption of all saturated fatty acids. The resulting calculation is

$$\text{Fatty Acids Simple Scoring Ratio} = \frac{\text{MUFA + PUFA Consumed}}{\text{Saturated Fatty Acids Consumed}}. \tag{2}$$

This is to encourage the consumption of more beneficial fatty acids (PUFAs and MUFAs) than saturated fatty acids. The Sodium ratio requires a conversion from the raw NHANES data in milligrams to grams first before being adjusted for each 1000 kilocalories consumed, and is calculated as

$$\text{Sodium Simple Scoring Ratio} = \frac{\text{Total Sodium Consumed (in grams)}}{\text{Total Kilocalories Consumed}} \times 1000. \tag{3}$$

The calculations of Added Sugars and Saturated Fat ratios require converting their raw measurements in teaspoons or grams to kilocalories to express them as percentages of total energy or kilocalories consumed. Reported Added Sugars are multiplied by a conversion factor of 16 as

$$\text{Added Sugars Simple Scoring Ratio} = \frac{\text{Total Sugar Consumed} \times 16}{\text{Total Kilocalories Consumed}} \times 100. \tag{4}$$

Likewise, reported Saturated Fats are multiplied by a factor of 9 as

$$\text{Saturated Fat Simple Scoring Ratio} = \frac{\text{Total Saturated Fats Consumed} \times 9}{\text{Total Kilocalories Consumed}} \times 100. \tag{5}$$

Table 3 provides a summary of these Simple ratio calculation rules.

Finally, for each component, the ratios are adjusted to make them proportional to the point values for the maximum and minimum standards prescribed in the DGA as

$$\text{Simple Score} = \frac{\text{Calculated Ratio} - \text{Minimum Score Standard}}{\text{Maximum Score Standard} - \text{Minimum Score Standard}} \times \text{Maximum Points}, \tag{6}$$

where 'Calculated Ratio' corresponds to any of the ratios calculated in (1) through (5).

2.3.2 Mean Ratio Scoring

The Simple scoring method generates scores for each individual for each component but ignores the survey weights associated with the individuals. To generate scores representative of the population, the weights should however be utilized in the scoring mechanism. The Mean Ratio scoring method provides an average score for a sub-population of interest (such as a specific sex, race/ethnicity, age, or income) incorporating the survey weights. Each individual is first scored using the Simple scoring method. For each HEI component, the score for the subgroup is then calculated as the weighted average across all individuals of the subgroup. The formula can be summarized as

$$\text{Mean Ratio Score} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \tag{7}$$

where w_i is an individual's assigned survey weight and x_i is their Simple score. The result in (7) is the Mean Ratio score for the corresponding group.

2.3.3 Population Ratio Scoring

The population ratio method is another way of scoring a demographic subgroup of interest prescribed by Freedman et al. (2008), Guenther et al. (2013) and Kirkpatrick et al. (2021). For individual i, let c_i denote the average recall amount, k_i the average raw kilocalorie consumption and w_i the survey weight, as before. The weighted mean of raw consumption of each component is then calculated across all individuals in the subgroup as

$$\text{Weighted Mean of Raw Consumption} = \frac{\sum_{i=1}^{n} w_i c_i}{\sum_{i=1}^{n} w_i}. \tag{8}$$

Likewise, the weighted mean of kilocalories is calculated as

$$\text{Weighted Mean of Total Kilocalories} = \frac{\sum_{i=1}^{n} w_i k_i}{\sum_{i=1}^{n} w_i}. \tag{9}$$

These means are then converted to ratios as

$$\text{Population Ratio} = \frac{\text{Weighted Mean of Raw Consumption}}{\text{Weighted Mean of Total Kilocalories}} \times 1000. \tag{10}$$

The simple scoring is then applied to these ratios to calculate the Population Ratio scores.

Along the lines of the Simple scoring method, scoring certain components requires adjustments to be made. The Fatty Acids Population Ratio score does not require dividing by the weighted mean of total kilocalories. Instead, the weighted means of MUFAs and PUFAs are summed and divided by the weighted mean of saturated fatty acids, as before. Once again, Added Sugars and Saturated Fats are evaluated based on the percentage of total daily calories they represent. Both of these components begin with the typical base ratio of the weighted mean of the raw intakes divided by the weighted mean of the total kilocalories consumed. Then, Added Sugars are converted from teaspoons to kilocalories by multiplying by 16, and Saturated Fats are converted from grams to kilocalories by multiplying the ratio by 9. These two ratios are finally multiplied by 100 to obtain the percent of daily calories that saturated fat and added sugars represent. Sodium recalls, originally reported in the NHANES in milligrams, are first converted to grams by dividing by 1000, and then adjusted for each 1000 kilocalories consumed, as before.

3 The `heiscore` Package

The `heiscore` package aims to increase the accessibility and use of the HEI, estimating it directly from publicly available preloaded NHANES data using the three aforementioned scoring methods via user-friendly functions and features that can calculate and plot the HEI scores across different demographic groups as well as different NHANES cylces.

3.1 Loading Data

The `selectDataset()` function fetches a tibble containing demographic information and unprocessed 24-hour recall data sourced from the WWEIA segment of the NHANES. Valid input years range from the 2005-2006 NHANES cycle through the 2017-2018 cycle. The function returns a raw dataset that includes respondents' subject numbers ('SEQN'), survey weights for the first and second dietary recalls ('WTDRD1' and 'WTDRD2'), and demographic information, including sex ('SEX'), race/ethnicity ('RACE_ETH'), age ('AGE'), and family income ('FAMINC').

Table 3: HEI Components and Their Simple Ratio Calculation Rules

Component	Rules
Adequacy	
Total Fruits	Basic Simple Scoring Ratio
Whole Fruits	Basic Simple Scoring Ratio
Total Vegetables	Basic Simple Scoring Ratio
Greens and Beans	Basic Simple Scoring Ratio
Whole Grains	Basic Simple Scoring Ratio
Dairy	Basic Simple Scoring Ratio
Total Protein Foods	Basic Simple Scoring Ratio
Seafood and Plant Proteins	Basic Simple Scoring Ratio
Fatty Acids	Add MUFAs and PUFAs together before dividing by SFAs.
Moderation	
Refined Grains	Basic Simple Scoring Ratio
Sodium	Total Sodium (in grams) to Total Kilocalories quotient multiplied by 1000.
Added Sugars	Total Sugar to Total Kilocalories quotient multiplied by 16×100.
Saturated Fats	Total Saturated Fats to Total Kilocalories quotient multiplied by 9×100.

This function is also used by `score()` and `plotScore()` to access the NHANES data for scoring and graphing. Day one and two of raw component intake for individuals are also in the dataset, and they begin with 'DR1' and 'DR2', respectively. Note that since the NHANES methodology has changed over the years, some demographic groups are missing from older cycles. For example, earlier survey cycles, such as the 2005-2006 data, do not have a separate category for subjects who identify as Asian. Additionally, responses were not required for demographic information such as family income, resulting in null values for some respondents.

To access a dataset using `selectDataset()`, input the last two digits of each year within the survey cycle of interest as a string for the `years` argument. For example, to retrieve the data for the 2017-2018 NHANES cycle, we call the cycle's years:

```
NHANES_2017 <- selectDataset(years = "1718")
print(head(NHANES_2017)[, 1:8])
```

```
   SEQN    WTDRD1    WTDR2D    SEX RACE_ETH AGE         FAMINC DR1TKCAL
1 93703     0.000        NA Female    Asian   2        >100000       NA
2 93704 81714.005 82442.869   Male    White   2        >100000     1230
3 93705  7185.561  5640.391 Female    Black  66 [10000, 15000)     1202
4 93706  6463.883     0.000   Male    Asian  18           <NA>     1987
5 93707 15333.777 22707.067   Male    Other  13 [65000, 75000)     1775
6 93708 10825.545 22481.854 Female    Asian  66 [25000, 35000)     1251
```

3.2 Scoring Data

The `score()` function is a versatile tool that simplifies the process of applying the three scoring methods to raw nutritional data. Simple (`"simple"`), Mean Ratio (`"mean ratio"`), and Population Ratio (`"pop ratio"`) scoring are all valid inputs for the required scoring method (`method`) argument. The `years` argument, inputted the same way as in `selectDataset()` to use data from a particular NHANES cycle, is also required. The function generates individual component or total HEI scores depending on the required argument `component` that accepts specific component names such as `"whole grains"` or `"total score"`. Additionally, a subset of the population can be selected with the `score()` function. By default, the function will calculate scores for all members of the population, but the `sex`, `race`, `age`, and `income` arguments allow users to investigate a sub-population of interest by inputting the desired demographic groups for these arguments. Table S.2 in the supplementary material provide a summary of the names and descriptions of the variables in the `score()` output.

To calculate the simple scores for the Total Vegetable component for all individuals in the 2017-2018 cycle, we supply `score()` with the scoring method, cycle years, and component of interest:

```
simple_score_veg <- score(method = "simple", years = "1718",
    component = "total vegetables")
print(head(simple_score_veg))
```

```
# A tibble: 6 x 7
   SEQN WTDR2D SEX    AGE   RACE_ETH         FAMINC         score
  <int>  <dbl> <fct>  <fct> <fct>            <fct>          <dbl>
1 93704 82443. Male   2     White            >100000         1.48
2 93705  5640. Female 66    Black            [10000, 15000)  4.05
3 93707 22707. Male   13    Other            [65000, 75000)  2.09
4 93708 22482. Female 66    Asian            [25000, 35000)  5
5 93711  8230. Male   56    Asian            >100000         5
6 93712 89066. Male   18    Mexican American [15000, 20000)  2.17
```

For the Mean Ratio and Population Ratio scoring methods, a single score is calculated for a population subgroup. Therefore, users can specify which demographic category they would like to group the population by (`"sex"`, `"race"`, `"age"`, or `"income"`) in the `demo` argument. For example, if a user wanted to compare the scores or males and females, they would input `"sex"` as the `demo`. Note that `NULL` must be inputted for this argument when Simple scoring is the chosen `method` since it generates a score for each individual.

The arguments necessary to calculate the Mean Ratio scores for the Total Vegetables component grouped by sex in the 2017-2018 cycle are:

```
mean_ratio_veg <- score(method = "mean ratio", years = "1718",
    component = "total vegetables", demo = "sex")
print(mean_ratio_veg)
```

```
# A tibble: 2 x 2
  SEX    score
```

```
  <fct>  <dbl>
1 Female  3.83
2 Male    3.04
```

Calculating the population ratio scores for the same group is similar, as we only have to change the `method`:

```
population_ratio_veg <- score(method = "pop ratio", years = "1718",
    component = "total vegetables", demo = "sex")
print(population_ratio_veg)
```

```
# A tibble: 2 x 2
  SEX    score
  <fct>  <dbl>
1 Female  3.66
2 Male    2.99
```

To access the Total Simple Scores for all individuals in the 2017-2018 cycle, the arguments are:

```
simple_total <- score(method = "simple", years = "1718", component = "total
↪  score")
print(head(simple_total)[, 1:7])
```

```
   SEQN    WTDR2D    SEX AGE         RACE_ETH         FAMINC    F_TOTAL
1 93704 82442.869   Male   2            White        >100000 5.00000000
2 93705  5640.391 Female  66            Black [10000, 15000) 0.07693886
3 93707 22707.067   Male  13            Other [65000, 75000) 0.00000000
4 93708 22481.854 Female  66            Asian [25000, 35000) 1.46321070
5 93711  8229.960   Male  56            Asian        >100000 3.69985863
6 93712 89066.416   Male  18 Mexican American [15000, 20000) 4.40385685
```

Calculating the Total Mean Ratio scores for all individuals in the 2017-2018 cycle grouped by sex uses similar arguments. However, we change the `method` to `"mean ratio"` and include the `demo` argument `"sex"`:

```
mean_ratio_total <- score(method = "mean ratio", years = "1718",
    component = "total score", demo = "sex")
print(head(mean_ratio_total)[, 1:5])
```

```
     SEX score_F_TOTAL score_FWHOLEFRT score_VTOTALLEG score_VDRKGRLEG
1   Male      2.966643        4.389309        3.043977        2.646224
2 Female      3.410473        5.000000        3.826086        3.678896
```

Calculating scores for the Total Fruit component using the Population Ratio method across different races and ethnicities in the 2005-2006 cycle with specific subsets for each demographic group would include the following arguments:

```
population_ratio_full <- score(method = "pop ratio", years = "0506",
    component = "total fruit", demo = "race", sex = "male", race = c("White",
        "Black", "Mexican American", "Other Hispanic"), age = c(2,
        100), income = c("[45000, 55000)", "[55000, 65000)",
        "[65000, 75000)", "[75000, 100000)", ">100000", ">20000",
        "<20000", "Refused", "Don't know", "NA"))
print(population_ratio_full)
```

```
# A tibble: 4 x 2
  RACE_ETH         score
  <fct>            <dbl>
1 Black             3.10
2 Mexican American  2.76
3 Other Hispanic    3.06
4 White             2.49
```

3.3 Visualizing Data

The `plotScore()` function allows users to effortlessly create visualizations from the HEI scoring data. The graphs produced by `plotScore()` make it easier to interpret and communicate the multidimensional data. Its arguments are consistent with the `score()` function, the only additional argument being the type of graph to display (`graph`). When Simple scoring is the chosen `method`, the only valid `graph` option is `"histogram"` since Simple scoring generates a score for each individual, and a histogram shows the distribution of these values. The only accepted `graph` input is `"bar"` for the Mean Ratio and Population Ratio scoring methods when plotting a single component score. When plotting total Mean or Population Ratio scores, both `"bar"` and `"radar"` plots are allowed.

The `plotScore()` function is compatible with the `ggplot2` package (Wickham, 2016). Users can add additional layers to the graphs generated by `plotScore()` such as adding titles and axes labels.

To create a histogram of the Total Vegetables component Simple Scores from the 2005-2006 data with a title and label on the x-axis, we use the following code:

```
plotScore(graph = "histogram", method = "simple", years = "0506",
    component = "total vegetables") + ggtitle("Total Vegetables Simple Scores
    ↪  for 2005-2006") +
    xlab("Total Vegetable Score")
```

Histograms provide a clear visual representation of a population's individual component or total HEI score distribution. In Figure 1, the mode of the distribution seems to be at the maximum of five points allotted to this category. The graph depicts more than 20% of the distribution being at the maximum score of 5 for the HEI component Total Vegetables, indicating that at least 20% of the total population in the selected year, 2005-2006, successfully met the recommended threshold set by the DGA. The remaining scores are concentrated around 2.5, meaning that people who did not eat the suggested amount of vegetables only consumed half of the recommendation on average. Similar trends described by Hurley et al. (2009) indicate lower

Figure 1: Histogram of Simple scores for the Total Vegetables component in 2005-2006. The greatest proportion of scores is located at the maximum score of 5.

scores to be related to disproportional development of chronic diseases, especially in minority populations.

To create a histogram of the total scores using the Simple scoring method for individuals in 2005-2006, we start with the same arguments as in the `score()` function and then add the `graph` argument:

```
plotScore(graph = "histogram", method = "simple", years = "0506",
    component = "total score") + ggtitle("Total Simple Scores for 2005-2006") +
    xlab("Total Score")
```

Unlike the Total Vegetable component, the Total Score histogram in Figure 2 is less skewed. However, the distribution is centered around nearly half of the total possible points, indicating that much of the population for the 2005-2006 cycle did not meet nutritional standards.

To create a bar chart of the Mean Ratio scores for the Total Vegetables grouped by the sex demographic in the 2005-2006 cycle, the following arguments are used:

```
plotScore(graph = "bar", method = "mean ratio", years = "0506",
    component = "total vegetables", demo = "sex") + ggtitle("Total Vegetable
    ↪  Mean Ratio for 2005-2006") +
    xlab("Sex")
```

Bar charts provide a method of comparison between relevant subgroups. In Figure 3, females score higher than males in the total vegetable component, indicating that they consume more vegetables on average compared to males.

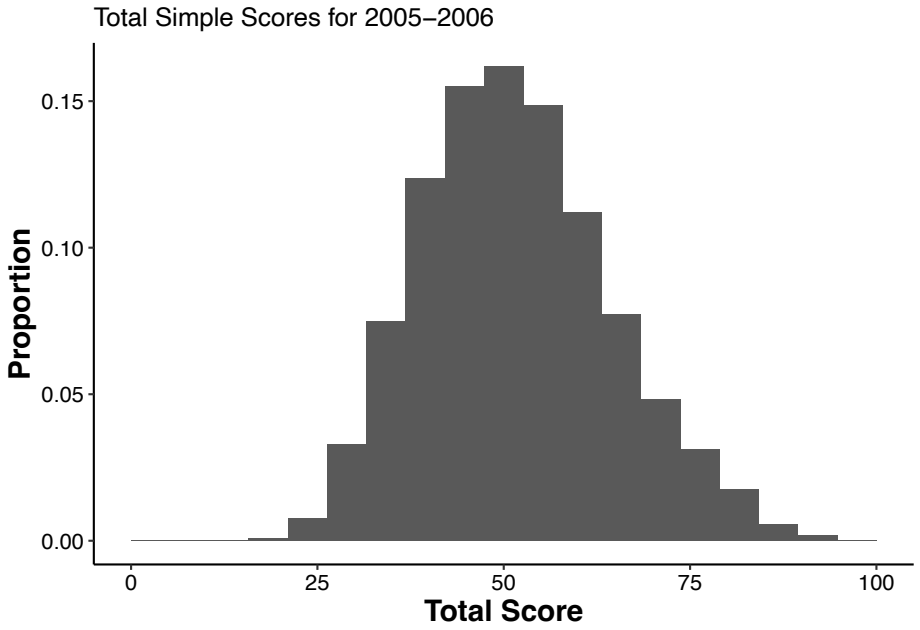


Figure 2: Histogram of Total Simple scores for the 2005-2006 cycle data. The distribution of scores is centered just above 50.

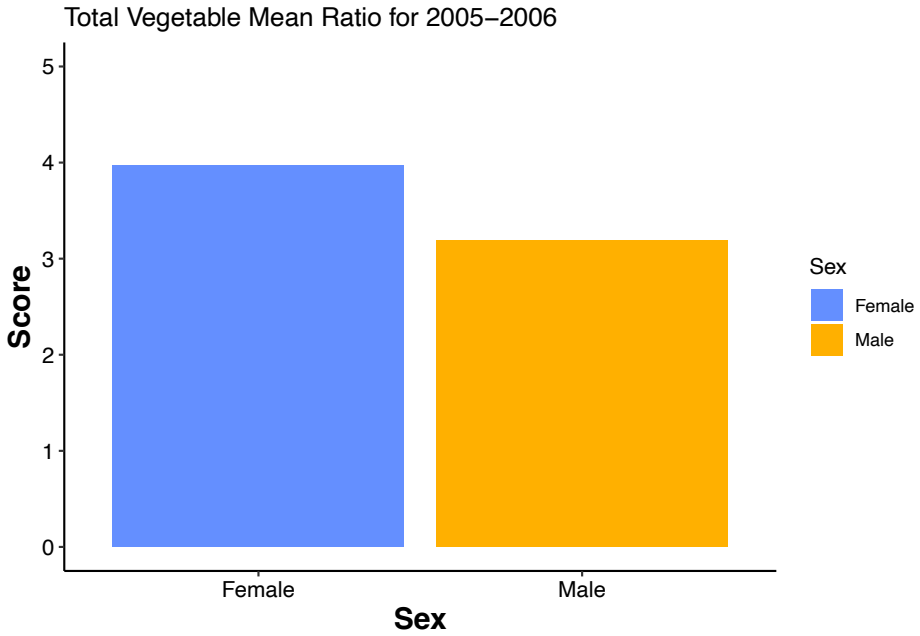


Figure 3: Bar chart of the Mean Ratio scores for the Total Vegetable component for the 2005-2006 cycle data. The scores are grouped by sex, with females scoring almost a full point higher than males.

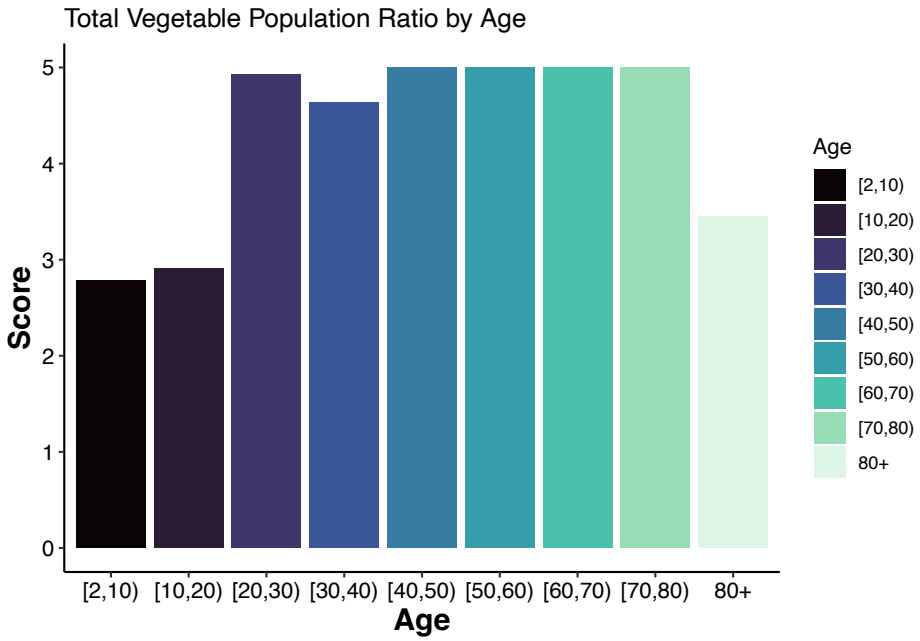


Figure 4: Bar chart of Population Ratio scores for the Total Vegetable component grouped by age. The data are from the 2011-2012 cycle and have been subset to Asian females.

To create a bar chart of the Population Ratio scores for the Total Vegetables component grouped by age in the 2011-2012 cycle, the code would be as follows:

```
plotScore(graph = "bar", method = "pop ratio", years = "1112",
    component = "total vegetables", demo = "age", sex = "female",
    race = "asian") + ggtitle("Total Vegetable Population Ratio by Age") +
    xlab("Age")
```

The bar chart in Figure 4 shows the distribution of scores across age groups for Asian females in 2011 and 2012. It is apparent that Asian women between the ages of 20 and 80 consumed more vegetables than younger generations of the same race and gender at this time. In Au et al. (2023) the dietary patterns of younger generations were observed to have an impact on the development of health issues later in life. Thus, it has been suggested that the eating habits developed at younger ages can improve quality of health as individuals age.

A bar chart of the Total Mean Ratio scores grouped by sex for the 2005-2006 data cycle is created with the following code:

```
plotScore(graph = "bar", method = "mean ratio", years = "0506",
    component = "total score", demo = "sex") + ggtitle("Mean Ratio Scores by
    ↪  Sex") +
    xlab("Sex")
```

The bar chart in Figure 5 depicts the difference in the Mean Ratio scores between females and males in 2005 and 2006. The female population again exhibits a higher Mean Ratio score.

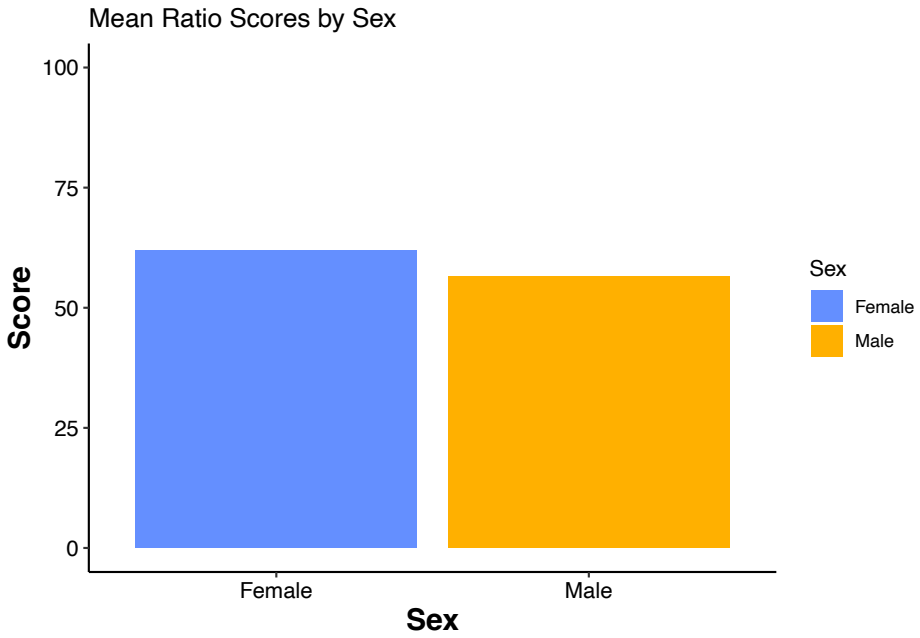


Figure 5: Bar chart of Mean Ratio scores by sex for the 2005-2006 cycle. Females appear to have higher scores compared to males.

This difference suggests that women achieved higher total HEI scores on average. Therefore, females generally adhered more closely to dietary recommendations than males, leading to a comparatively healthier and more balanced overall diet.

As noted, `heiscore`'s plotting functions are compatible with the package `ggplot2`. Users are able to add additional plot layers, such as subtitles or labels, or make adjustments, such as centering the titles, to the graphs generated by `plotScore()`. Additionally, the output from `score()` can be used as an argument for `ggplot2` functions. For example, scoring data generated by `score()` can be used as the data argument for `geom_text()` to add labels over a bar chart for clarity. There are several functions that can facilitate placing multiple graphs generated by `plotScore()` next to each other in the desired layout. Examples include the `grid.arrange()` function from `gridExtra` (Auguie, 2017), `plot_grid()` from `cowplot` (Wilke, 2024), and `plot_layout()` from `patchwork` (Pedersen, 2024). We show an example of this compatibility in Figure S.1 in the supplementary material.

3.3.1 Radar Plots

Radar charts can effectively display multiple variables on a single diagram, where each variable is represented by an axis extending outward from the center, making it easy to compare the patterns of two or more groups (Saary, 2008). They are therefore particularly useful for viewing multidimensional HEI data. On the radar plots drawn by our `heiscore` package, a point on the innermost or outermost end of an axis represents a minimum or maximum component score, respectively. Since not all component scores have the same maximum value, the axes range from 0% to 100% of the component score. Our package allows the use of distinctly colored and shaded lines to display the 13 component scores for different demographic subgroups.

To create a radar plot of the total scores using the Population Ratio method grouped by sex

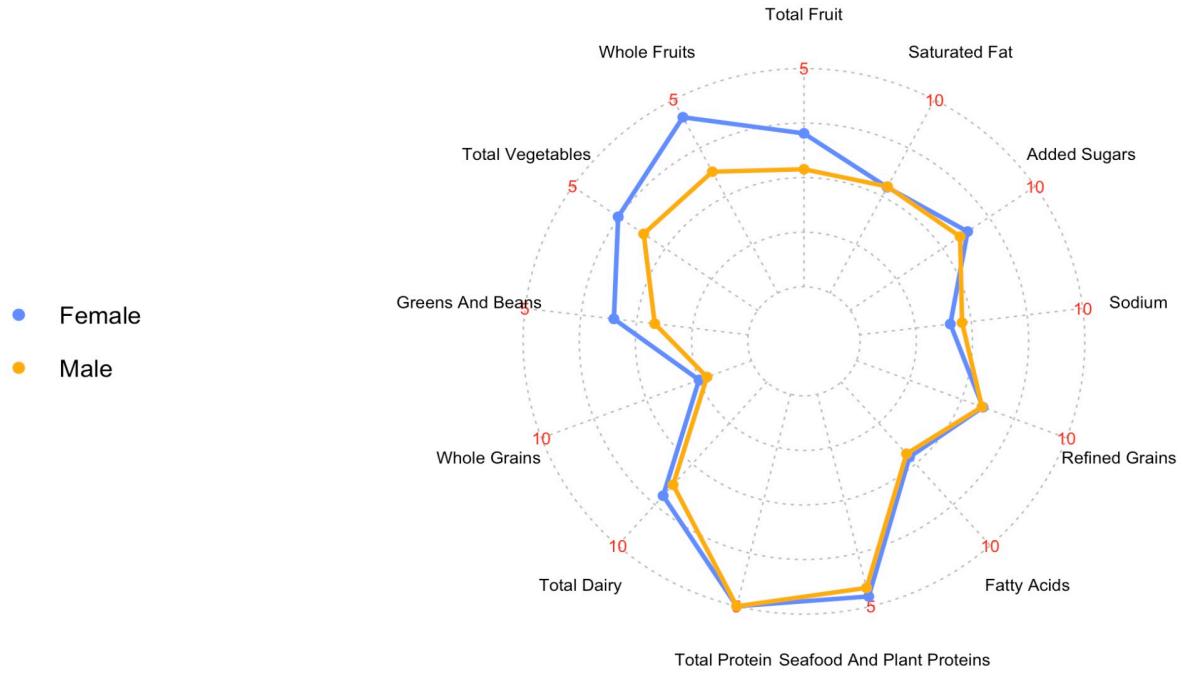


Figure 6: Radar plot showing the Population Ratio scores of men and women in 2005 and 2006. It appears that the female population exhibited higher scores than males in majority dietary elements, including especially Total Fruit, Whole Fruits, Total Vegetables, and Greens and Beans.

for the 2005-2006 data, users can change the `graph` argument to `"radar"` and use the following code:

```
plotScore(graph = "radar", method = "pop ratio", years = "0506",
    component = "total score", demo = "sex")
```

Cohort studies to observe eating habits by gender have been performed to determine the effects gender possibly has on dietary patterns (George et al., 2021). The radar plot in Figure 6 shows the Population Ratio scores of men and women in 2005 and 2006. It appears that the female population exhibited higher scores than males in majority dietary elements, including especially Total Fruit, Whole Fruits, Total Vegetables, and Greens and Beans. The radar plot highlights the female population's superior performance in achieving higher HEI scores across multiple dietary components, indicating that females generally had a healthier and more well-rounded dietary profile than males in 2005 and 2006. The heightened scores for females, particularly in the Whole Fruits component, reveal their better compliance with recommended dietary guidelines than males'. Similar patterns were consistently observed in the radar plots generated by our package across all NHANES cycles, reinforcing the significance of gender as a notable factor influencing dietary health.

To create a radar plot of the Total Population Ratio score grouped by family income for the 2017-2018 data, the following code can be used:

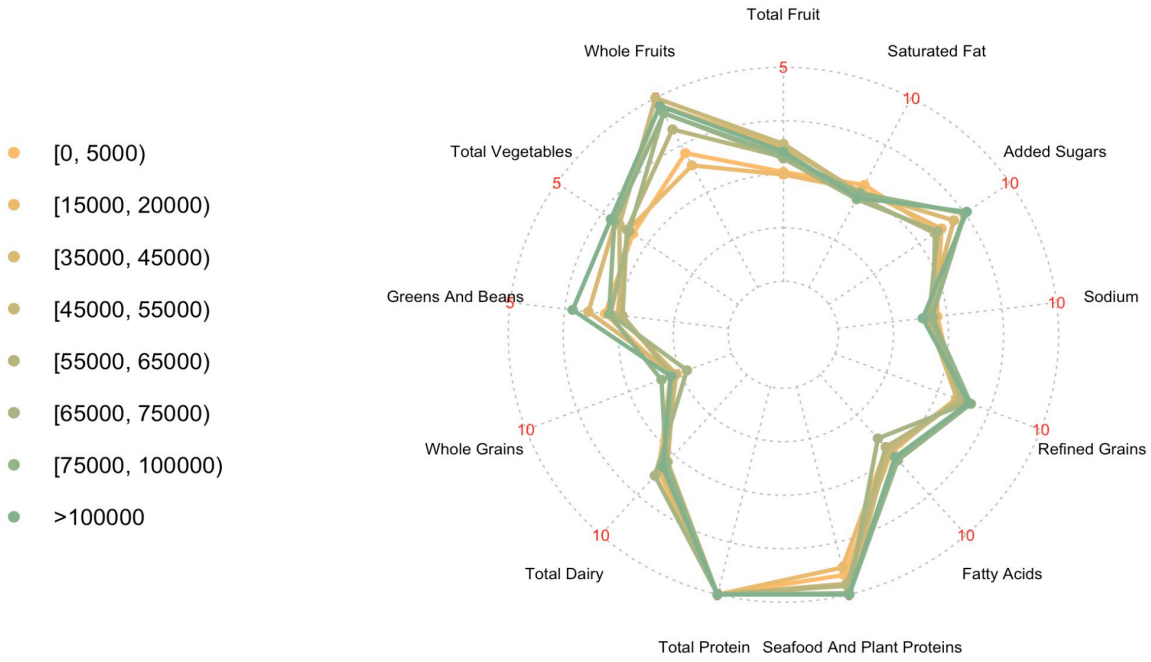


Figure 7: Radar plot comparing the dietary habits of different family income groups for the 2017-2018 cycle, revealing a trend where higher income brackets achieve greater HEI scores, particularly in categories such as Whole Fruits. Each income level is represented by a different color on the plot.

```
plotScore(graph = "radar", method = "pop ratio", years = "1718",
    component = "total score", demo = "income", income = c("[0, 5000)",
        "[15000, 20000)", "[35000, 45000)", "[45000, 55000)",
        "[55000, 65000)", "[65000, 75000)", "[75000, 100000)",
        ">100000"))
```

Studies by Aggarwal et al. (2012) and Vinyard et al. (2021) have found substantial influence of the socioeconomic environment on individuals' dietary habits. The radar plot in Figure 7 visualizes the association between higher income levels and more favorable dietary patterns, specifically the increased consumption of Whole Fruits. This suggests a positive association between socioeconomic status and healthier dietary choices, as reflected in the elevated scores for families with higher incomes, particularly for Whole Fruits. Similar patterns were observed in the radar plots in all NHANES cycles, reinforcing socioeconomic status to be an important predictor of dietary health.

3.3.2 Shiny App

Figure 8 illustrates the `runShinyApp()` function included in our package, providing a user-friendly visualization tool that does not require much technical expertise or experience. The app interface allows users to navigate through tabs to view raw data distributions, visualizations from

Figure 8: The shiny app contains a dashboard for users to navigate through dietary data for the cycle of their choice. Users can manipulate which dietary components or constituents to display, as well as the population demographics whose dietary trends they wish to see.

the `plotScore()` function, demographic distributions, and background information on the HEI components. Users can select their population group of interest and the HEI component to plot by simply clicking on the correct boxes in the side panel. Additionally, two graphs using data from different NHANES cycles can be compared side by side. Overall, `runShinyApp()` allows users with limited technical expertise to visualize nutritional epidemiology data with ease.

4 Discussion

In this article, we described a new `R` package, **`heiscore`**, which aims to enhance the accessibility and usability of the HEI by providing user-friendly functions that implement three distinct scoring methods to estimate it directly from preloaded NHANES dietary recall data. Our functions allow users to easily compare the HEI scores across different demographic groups, aiding in the study of contrasts and disparities between these sub-populations. Furthermore, **`heiscore`** goes beyond just calculating the HEI scores by including user-friendly functions and apps to visualize multidimensional HEI scores via histograms, bar charts, and radar plots.

While **`heiscore`** offers a robust set of tools and features for estimating the HEI and comparing its values across demographic subgroups, one limitation of the current version is that its functions do not adjust for the presence of measurement errors in the recalls. Such corrections can result in more accurate estimates of the dietary intake values used in the scoring methods (Schap et al., 2017) but requires substantial additional effort implementing sophisticated statistical methods (Sarkar et al., 2021). Our plans for further software development include potential integration with these methods to improve the efficacy of HEI estimation.

5 Acknowledgements

We thank Ms. Rimli Sengupta for help in understanding and implementing the NCI HEI SAS codes used to validate `heiscore`'s results.

References

Aggarwal A, Monsivais P, Drewnowski A (2012). Nutrient intakes linked to better health outcomes are associated with higher diet costs in the us. *PLOS ONE*, 7.5: e37533.

Au LE, Arnold CD, Ritchie LD, Lin SK, Frongillo EA (2023). Differences in infant diet quality index by race and ethnicity predict differences in later diet quality. *The Journal of Nutrition*, 153: 3498–3505.

Auguie B (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3.

Cena H, Calder PC (2020). Defining a healthy diet: Evidence for the role of contemporary dietary patterns in health and disease. *Nutrients*, 12: 334.

Chen T, Clark J, Riddles M, Mohadjer L, Fakhouri T (2020). National health and nutrition examination survey, 2015-2018: Sample design and estimation procedures. *Vital Health Stat*, 2: 1–35.

Endres C, Ale L, Gentleman R, Sarkar D (2024). *nhanesA: NHANES Data Retrieval*.

Freedman LS, Guenther PM, Krebs-Smith SM, Kott PS (2008). A Population's Mean Healthy Eating Index-2005 Scores Are Best Estimated by the Score of the Population Ratio when One 24-Hour Recall Is Available. *The Journal of Nutrition*, 138: 1725–1729.

George SM, Reedy J, Cespedes Feliciano EM, Aragaki A, Caan BJ, Kahle L, et al. (2021). Alignment of dietary patterns with the dietary guidelines for americans 2015-2020 and risk of all-cause and cause-specific mortality in the women's health initiative observational study. *American Journal of Epidemiology*, 190: 886–892.

Guenther PM, Casavale KO, Reedy J, Kirkpatrick SI, Hiza HA, Kuczynski KJ, et al. (2013). Update of the healthy eating index: Hei-2010. *Journal of the Academy of Nutrition and Dietetics*, 113: 569–580.

Herrick KA, Roseen LM, Parsons R, Dodd KW (2018). Estimating usual dietary intake from national health and nutrition examination survey data using the national cancer institute method. *Vital Health Stat*, 2(1): 63.

Hu FB (2002). Dietary pattern analysis: A new direction in nutritional epidemiology. *Current Opinion in Lipidology*, 13: 3–9.

Hurley KM, Oberlander SE, Merry BC, Wrobleski MM, Klassen AC, Black MM (2009). The healthy eating index and youth healthy eating index are unique, nonredundant measures of diet quality among low-income, african american adolescents. *The Journal of Nutrition*, 139: 359–364.

Kennedy ET, Ohls J, Carlson S, Fleming K (1995). The healthy eating index: Design and applications. *Journal of the American Dietetic Association*, 95: 1103–1108.

Kirkpatrick SI, Dodd KW, Potischman N, Zimmerman TP, Douglass D, Guenther PM, et al. (2021). Healthy eating index-2015 scores among adults based on observed vs recalled dietary intake. *Journal of the Academy of Nutrition and Dietetics*, 121: 2233–2241.e1.

Kirkpatrick SI, Reedy J, Krebs-Smith SM, Pannucci TE, Subar AF, Wilson MM, et al. (2018). Applications of the healthy eating index for surveillance, epidemiology, and intervention re-

search: Considerations and caveats. *Journal of the Academy of Nutrition and Dietetics*, 118: 1603–1621.

Nagraj V, Folsom T (2017). *hei*.

Pannucci TE, Lerman JL, Herrick KA, Shams-White MM, Zimmer M, Meyers Mathieu K, et al. (2023). Development of the healthy eating index-toddlers-2020. *Journal of the Academy of Nutrition and Dietetics*, 123: 1289–1297.

Pedersen TL (2024). *patchwork: The Composer of Plots*. R package version 1.2.0.

Pruim R (2015). *NHANES: Data from the US National Health and Nutrition Examination Study*.

Ramdas V, Ho B, Sarkar A (2024). *heiscore: Score and Plot the Healthy Eating Index from NHANES Data*.

Saary MJ (2008). Radar plots: a useful way for presenting multivariate health care data. *Journal of Clinical Epidemiology*, 61: 311–317.

Sarkar A, Pati D, Mallick BK, Carroll RJ (2021). Bayesian copula density deconvolution for zero-inflated data in nutritional epidemiology. *Journal of the American Statistical Association*, 116: 1075–1087.

Schap T, Kuczynski K, Hiza H (2017). Healthy eating index—beyond the score. *Journal of the Academy of Nutrition and Dietetics*, 4: 519–521.

Shams-White MM, Pannucci TE, Lerman JL, Herrick KA, Zimmer M, Mathieu KM, et al. (2023). Healthy eating index-2020: Review and update process to reflect the dietary guidelines for americans, 2020-2025. *Journal of the Academy of Nutrition and Dietetics*, 123: 1280–1288.

Sharkey JR (2008). Diet and health outcomes in vulnerable populations. *Annals of the New York Academy of Sciences*, 1136: 210–217.

Vinyard M, Zimmer M, Herrick KA, Story M, Juan W, Reedy J (2021). Healthy eating index-2015 scores vary by types of food outlets in the united states. *Nutrients*, 13: 2717.

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wilke CO (2024). *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 1.1.3.

Zhan JJ, Hodge RA, Dunlop A (2023a). *dietaryindex*.

Zhan JJ, Hodge RA, Dunlop AL, Lee MM, Bui L, Liang D, et al. (2023b). Dietaryindex: A user-friendly and versatile R package for standardizing dietary pattern analysis in epidemiological and clinical studies. *bioRxiv*.

Journal of Data Science xx (xx), 1–S.10 DOI: 0000
July 2022

Supplementary Material for HEI Analysis of NHANES Dietary Data: Exploring the Diet Quality of Americans with R Package heiscore

Berkeley Ho[1,†], Vijetha Ramdas[1,†], and Abhra Sarkar[1]

[1]*Department of Statistics and Data Sciences, The University of Texas at Austin, TX 78705, USA*

Contents

*Corresponding author Email: abhra.sarkar@utexas.edu.
†Equal contributions

S.1 Additional Illustration of HEI Exploratory Analysis Using `heiscore`

Figure S.1 visualizes the decrease in Saturated Fat Mean Ratio Scores from the 2011-12 to the 2017-18 NHANES cycle for all of the five primary race/ethnicity groups. The race/ethnicity group `"Other"` is excluded from the graphs. Only the NHANES cycles from 2011-2018 are included since the cycles prior to 2011 did not have the `"Asian"` category. It seems that the saturated fat scores of almost all of the five groups decreased between each cycle in the eight-year time period. The exception to this pattern was the Mexican-American group which exhibited a slight increase from the 2011-12 cycle to the 2013-14 cycle.

```
library(heiscore)
library(gridExtra)
library(tidyverse)

data_1112 <- score(method = "Mean Ratio",
                   years = "1112",
                   component = "Saturated Fat",
                   demo = "Race",
                   race = c("Asian", "White", "Black", "Mexican American",
                   ↪  "Other Hispanic"))
plot_1112 <- heiscore::plotScore(graph = "Bar",
                   method = "Mean Ratio",
                   years = "1112",
                   component = "Saturated Fat",
                   demo = "Race",
                   race = c("Asian", "White", "Black", "Mexican American",
                   ↪  "Other Hispanic")) +
  geom_text(data = data_1112, aes(x = RACE_ETH, y = score, label=round(score,
  ↪  2), vjust = -1)) +
  labs(subtitle = "2011-12") +
  theme(
  plot.title = element_text(hjust = 0.5),
  plot.subtitle = element_text(hjust = 0.5)
)

data_1314 <- score(method = "Mean Ratio",
                   years = "1314",
                   component = "Saturated Fat",
                   demo = "Race",
                   race = c("Asian", "White", "Black", "Mexican American",
                   ↪  "Other Hispanic"))
plot_1314 <- heiscore::plotScore(graph = "Bar",
                   method = "Mean Ratio",
                   years = "1314",
                   component = "Saturated Fat",
```

```
                    demo = "Race",
                    race = c("Asian", "White", "Black", "Mexican American",
                    ↪  "Other Hispanic")) +
  geom_text(data = data_1314, aes(x = RACE_ETH, y = score, label=round(score,
  ↪  2), vjust = -1)) +
  labs(subtitle = "2013-14") +
  theme(
  plot.title = element_text(hjust = 0.5),
  plot.subtitle = element_text(hjust = 0.5)
  )

data_1516 <- score(method = "Mean Ratio",
                    years = "1516",
                    component = "Saturated Fat",
                    demo = "Race",
                    race = c("Asian", "White", "Black", "Mexican American",
                    ↪  "Other Hispanic"))
plot_1516 <- heiscore::plotScore(graph = "Bar",
                    method = "Mean Ratio",
                    years = "1516",
                    component = "Saturated Fat",
                    demo = "Race",
                    race = c("Asian", "White", "Black", "Mexican American",
                    ↪  "Other Hispanic")) +
  geom_text(data = data_1516, aes(x = RACE_ETH, y = score, label=round(score,
  ↪  2), vjust = -1)) +
  labs(subtitle = "2015-16") +
  theme(
  plot.title = element_text(hjust = 0.5),
  plot.subtitle = element_text(hjust = 0.5)
  )

data_1718 <- score(method = "Mean Ratio",
                    years = "1718",
                    component = "Saturated Fat",
                    demo = "Race",
                    race = c("Asian", "White", "Black", "Mexican American",
                    ↪  "Other Hispanic"))
plot_1718 <- heiscore::plotScore(graph = "Bar",
                    method = "Mean Ratio",
                    years = "1718",
                    component = "Saturated Fat",
                    demo = "Race",
                    race = c("Asian", "White", "Black", "Mexican American",
                    ↪  "Other Hispanic")) +
  geom_text(data = data_1718, aes(x = RACE_ETH, y = score, label=round(score,
  ↪  2), vjust = -1)) +
```

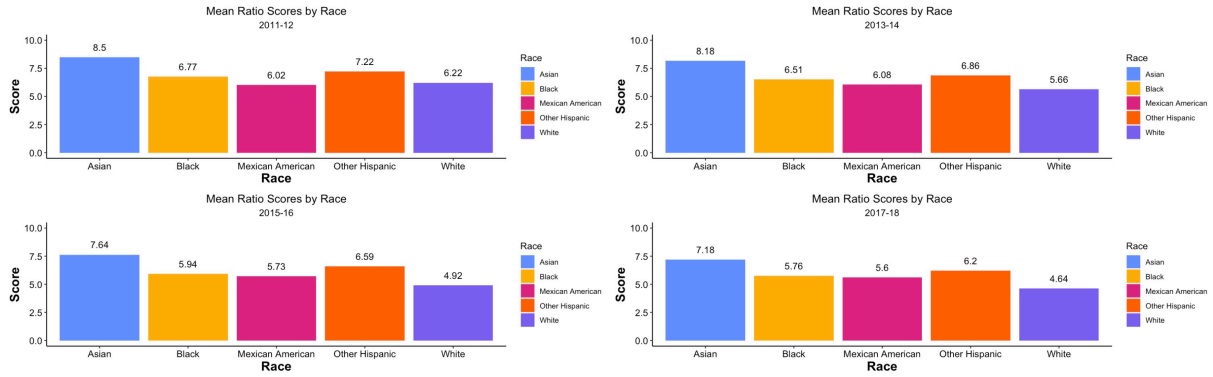


Figure S.1: Bar plots showing the Saturated Fat Mean Ratio scores across 5 primary race/ethnicity categories for each of the four NHANES cycles between 2011 and 2018. Each of the four charts displays the data for one NHANES cycle. Each bar represents the score for the respective race/ethnicity group, and the numeric score is labeled above the bar. Generally, the saturated fat scores of each race/ethnicity category decreased across the four cycles.

```
  labs(subtitle = "2017-18") +
  theme(
  plot.title = element_text(hjust = 0.5),
  plot.subtitle = element_text(hjust = 0.5)
  )

layout <- matrix(c(1, 2,
                   3, 4,
                   5, 6,
                   7, 8), ncol = 2, byrow = TRUE)
gridExtra::grid.arrange(plot_1112, plot_1314, plot_1516, plot_1718,
↪  layout_matrix = layout)
```

S.2 HEI-Toddlers-2020 Scoring Standards

Table S.1: HEI-2020 Toddler Scoring Standards

Component	Points	Standard for Maximum Score	Standard for Minimum Score
Adequacy			
Total Fruits	5	$\geqslant$ 0.7 cups per 1000 kcal	0 cups per 1000 kcal
Whole Fruits	5	$\geqslant$ 0.3 cups per 1000 kcal	0 cups per 1000 kcal
Total Vegetables	5	$\geqslant$ 0.9 cups per 1000 kcal	0 cups per 1000 kcal
Greens and Beans	5	$\geqslant$ 0.1 cups per 1000 kcal	0 cups per 1000 kcal
Whole Grains	10	$\geqslant$ 1.5 oz per 1000 kcal	0 oz per 1000 kcal
Dairy	10	$\geqslant$ 2.0 cups per 1000 kcal	0 cups per 1000 kcal
Total Protein Foods	5	$\geqslant$ 2.0 oz per 1000 kcal	0 oz per 1000 kcal
Seafood and Plant Proteins	5	$\geqslant$ 0.5 oz per 1000 kcal	0 oz per 1000 kcal
Fatty Acids	10	$\frac{\text{PUFAs + MUFAs}}{\text{SFAs}} \geqslant 1.5$	$\frac{\text{PUFAs + MUFAs}}{\text{SFAs}} \leqslant 0.9$
Moderation			
Refined Grains	10	$\leqslant$ 1.5 oz per 1000 kcal	$\geqslant$ 3.4 oz per 1000 kcal
Sodium	10	$\leqslant$ 1.1 grams per 1000 kcal	$\geqslant$ 1.7 grams per 1000 kcal
Added Sugars	10	0% of total kcal	$\geqslant$ 13.8% of total kcal
Saturated Fats	10	$\leqslant$ 12.2% of total kcal	$\geqslant$ 18.2% of total kcal

S.3 Variable Names and Descriptions for `score()` output

Table S.2: Variable Names and Descriptions

Variable Names	Description
SEQN	Respondent Sequence Number
WTDR2D	Dietary Two-Day Sample Weight
SEX	Sex of Respondent
RACE_ETH	Race/Ethnicity of Respondent
AGE	Age of Respondent
FAMINC	Self-Reported Annual Family Income of Respondent
F_TOTAL	Total Fruits
FWHOLEFRT	Whole Fruits
VTOTALLEG	Total Vegetables
VDRKGRLEG	Greens and Beans
G_WHOLE	Whole Grains
D_TOTAL	Dairy
PFALLPROTLEG	Total Protein Foods
PFSEAPLANTLEG	Seafood and Plant Proteins
TFACIDS	Fatty Acids
G_REFINED	Refined Grains
TSODI	Sodium
ADD_SUGARS	Added Sugars
TSFAT	Saturated Fat

S.4 Validation of heiscore

The code and figure below validate the Simple Scoring calculations performed by heiscore's score() function. The score() output is compared to the results from the NCI HEI Sample SAS Code - Simple HEI Scoring Algorithm Per Person for the 2011-12 cycle.

Load Data

```
# Load the NCI simple scoring results
validation_data <- read_csv("NCI_results.csv", show_col_types = FALSE)
# Retrieve raw data from heiscore
heiscore_raw <- selectDataset("1112")
# Retrieve heiscore's score() results
heiscore_score <- score(method = "simple", years = "1112", component = "total
↪  score")
```

Compare datasets

There are observations in the NCI results that are missing from the heiscore::score() results. We find that these missing observations lack a day 2 sample weight (WTDR2D). These results are therefore intentionally excluded from the heiscore::score() output because our package only includes individuals that participated in both days of the NHANES dietary recall process and are therefore assigned day 2 sample weights.

```
# Find observations missing from heiscore's score() output
missing_from_heiscore <- anti_join(x = validation_data,
                                   y = heiscore_score,
                                   by = "SEQN")
nrow(missing_from_heiscore)
```

```
## [1] 833
```

```
# Get more information about these missing observations
missing_info <- heiscore_raw[heiscore_raw$SEQN %in% missing_from_heiscore$SEQN,]
# These observations are missing because they all have NA as the day 2 sample
↪  weight values
sum(is.na(missing_info$WTDR2D))
```

```
## [1] 833
```

There are no observations in the heiscore::score() output that are missing from the NCI results.

```
# Find observations missing from validation data
missing_from_validation <- anti_join(x = heiscore_score,
```

```
                                    y = validation_data,
                                    by = "SEQN")
nrow(missing_from_validation)
```

```
## [1] 0
```

Join data

```
joined_data <- left_join(x = heiscore_score,
                         y = validation_data,
                         by = "SEQN")
```

Calculate accuracy score

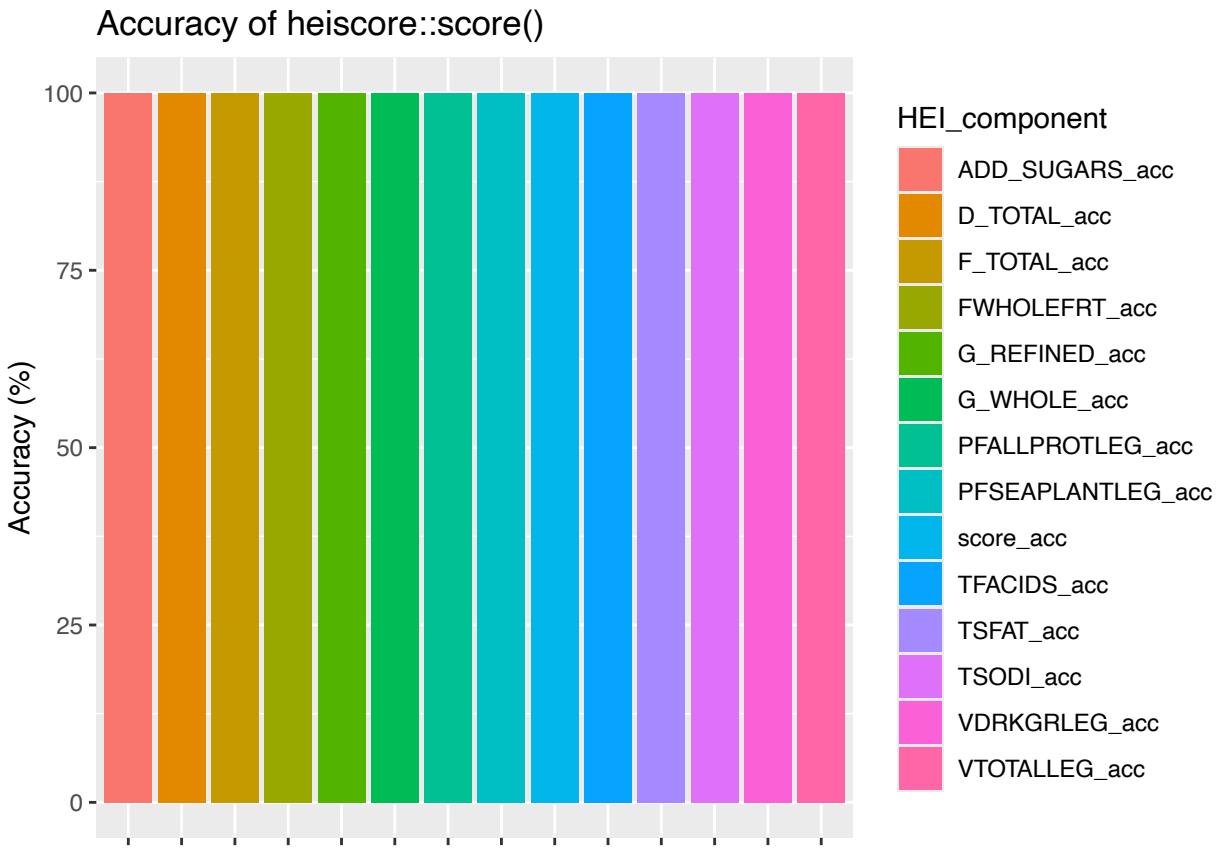
To calculate the accuracy score, both the validation and `heiscore::score()` results are rounded to 2 decimal places. Then, for each HEI component, the proportion of subjects with exact matches between the two results is calculated.

```
# Initialize a dataframe to store validation results
validation_results <- data.frame(SEQN = joined_data$SEQN)

# Find the difference between the NCI and heiscore results for each observation
validation_results <- joined_data %>%
  transmute(
    F_TOTAL_acc = abs(round(F_TOTAL, 2) - round(HEI2015C3_TOTALFRUIT, 2)),
    FWHOLEFRT_acc = abs(round(FWHOLEFRT, 2) - round(HEI2015C4_WHOLEFRUIT, 2)),
    VTOTALLEG_acc = abs(round(VTOTALLEG, 2) - round(HEI2015C1_TOTALVEG, 2)),
    VDRKGRLEG_acc = abs(round(VDRKGRLEG, 2) - round(HEI2015C2_GREEN_AND_BEAN,
    ↪  2)),
    G_WHOLE_acc = abs(round(G_WHOLE, 2) - round(HEI2015C5_WHOLEGRAIN, 2)),
    D_TOTAL_acc = abs(round(D_TOTAL, 2) - round(HEI2015C6_TOTALDAIRY, 2)),
    PFALLPROTLEG_acc = abs(round(PFALLPROTLEG, 2) - round(HEI2015C7_TOTPROT,
    ↪  2)),
    PFSEAPLANTLEG_acc = abs(round(PFSEAPLANTLEG, 2) -
    ↪  round(HEI2015C8_SEAPLANT_PROT, 2)),
    TFACIDS_acc = abs(round(TFACIDS, 2) - round(HEI2015C9_FATTYACID, 2)),
    G_REFINED_acc = abs(round(G_REFINED, 2) - round(HEI2015C11_REFINEDGRAIN,
    ↪  2)),
    TSODI_acc = abs(round(TSODI, 2) - round(HEI2015C10_SODIUM, 2)),
    ADD_SUGARS_acc = abs(round(ADD_SUGARS, 2) - round(HEI2015C13_ADDSUG, 2)),
    TSFAT_acc = abs(round(TSFAT, 2) - round(HEI2015C12_SFAT, 2)),
    score_acc = abs(round(score, 2) - round(HEI2015_TOTAL_SCORE, 2)))

# Define a function to calculate the proportion of 0s in a column
proportion_of_zeros <- function(column) {
  sum(column == 0) / length(column) * 100
```

```
}

# Apply the function to each column (component) in the dataframe
proportions <- apply(validation_results, 2, proportion_of_zeros)

# Convert the vector of accuracy scores to a dataframe
proportions_df <- data.frame(HEI_component = names(proportions), accuracy_score
↪  = unname(proportions))

# Print the accuracy scores
print(proportions_df)
```

```
##        HEI_component accuracy_score
## 1        F_TOTAL_acc      100.00000
## 2      FWHOLEFRT_acc       99.98592
## 3      VTOTALLEG_acc      100.00000
## 4      VDRKGRLEG_acc      100.00000
## 5        G_WHOLE_acc      100.00000
## 6        D_TOTAL_acc      100.00000
## 7   PFALLPROTLEG_acc       99.97184
## 8  PFSEAPLANTLEG_acc       99.97184
## 9        TFACIDS_acc       99.98592
## 10     G_REFINED_acc       99.98592
## 11         TSODI_acc      100.00000
## 12    ADD_SUGARS_acc      100.00000
## 13         TSFAT_acc       99.97184
## 14         score_acc      100.00000
```

```
# Plot the accuracy scores
ggplot(proportions_df, aes(x = HEI_component, y = accuracy_score, fill =
↪  HEI_component)) +
  geom_bar(stat = "identity") +
  ylab("Accuracy (%)") +
  xlab(NULL) +
  ggtitle("Accuracy of heiscore::score()") +
  theme(axis.text.x = element_blank())
```


Accuracy of heiscore::score()
Accuracy (%)
100
75
50
25
0
HEI_component
ADD_SUGARS_acc
D_TOTAL_acc
F_TOTAL_acc
FWHOLEFRT_acc
G_REFINED_acc
G_WHOLE_acc
PFALLPROTLEG_acc
PFSEAPLANTLEG_acc
score_acc
TFACIDS_acc
TSFAT_acc
TSODI_acc
VDRKGRLEG_acc
VTOTALLEG_acc