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",
```

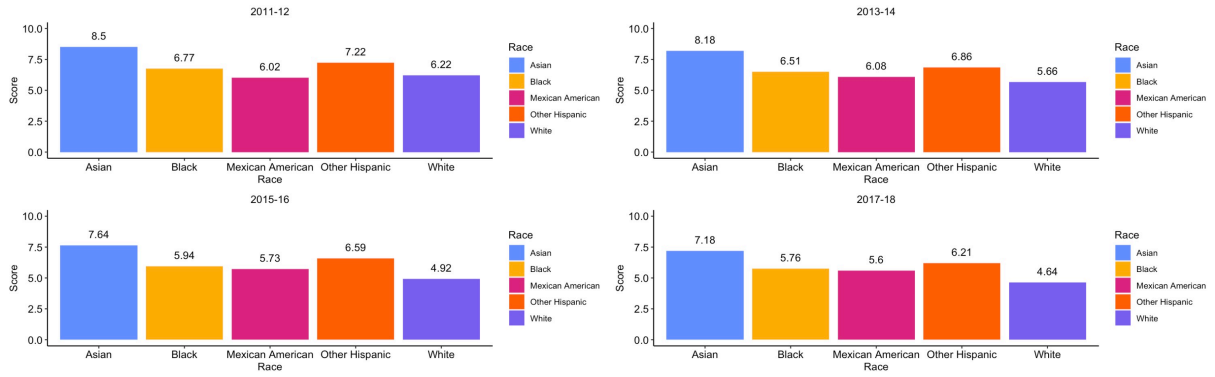


Figure S.1: Bar plots showing the Saturated Fat Mean Ratio scores across 5 primary race/ethnicity categories for each of the four NHANES cycles between 2011 and 2018. Each of the four charts displays the data for one NHANES cycle. Each bar represents the score for the respective race/ethnicity group,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

```
                          "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)) +
  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.

```r
# 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

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

Calculate accuracy score

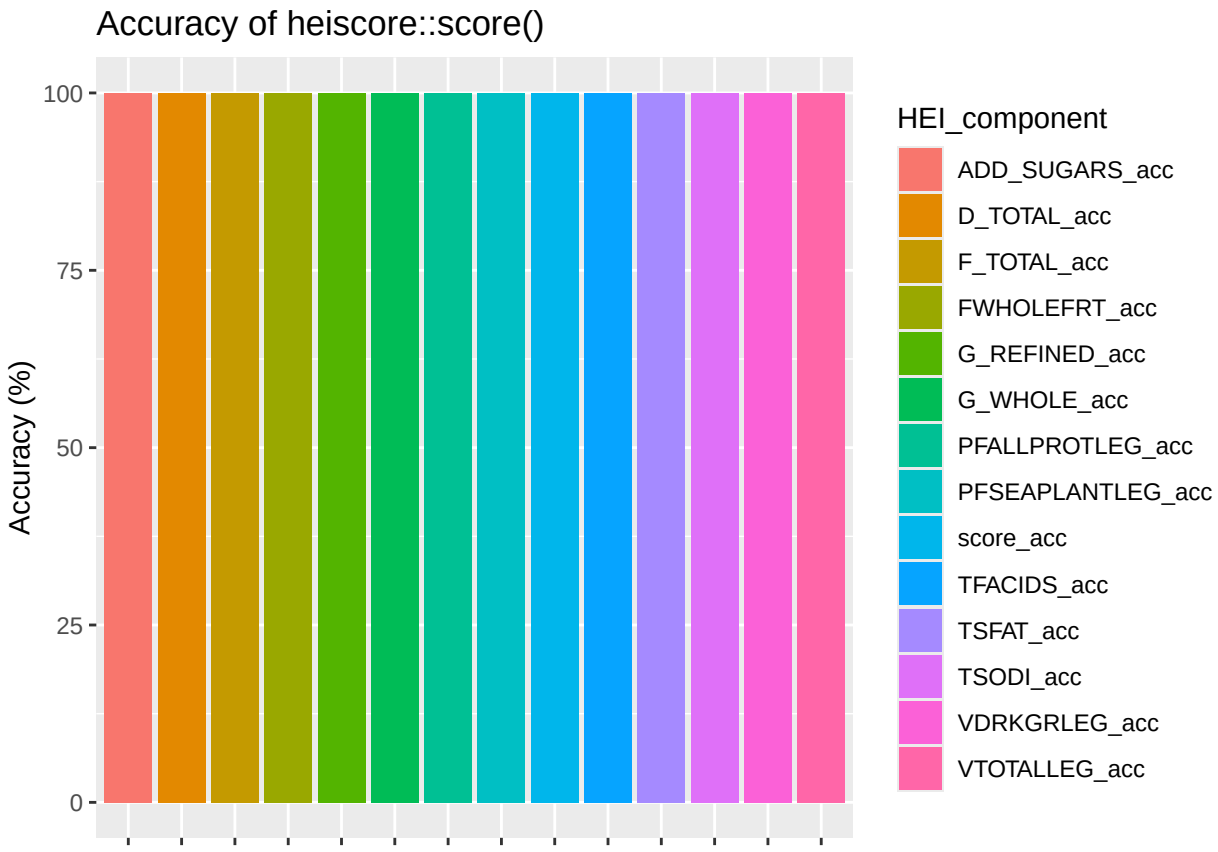
To calculate the accuracy score, both the validation and `heiscore::score()` results are rounded to 2 decimal places. Then, for each HEI component, the proportion of subjects with exact matches between the two results is calculated.

```r
# 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)))
```

```r
# 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
```

```r
# 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