Towards Measuring the Food Quality of Grocery Purchases: an Estimation Model of the Healthy Eating Index-2010 Using only Food Item Counts

Le-Thuy T. Tran, Philip J. Brewster, Valliammai Chidambaram, John F. Hurdle

Abstract

Measuring the quality of food consumed by individuals or groups in the U.S. is essential to informed public health surveillance efforts and sound nutrition policymaking. For example, the Healthy Eating Index-2010 (HEI) is an ideal metric to assess the food quality of households, but the traditional methods of collecting the data required to calculate the HEI are expensive and burdensome. We evaluated an alternative source: rather than measuring the quality of the foods consumers eat, we want to estimate the quality of the foods consumers buy. To accomplish that we need a way of estimating the HEI based solely on the count of food items. We developed an estimation model of the HEI, using an augmented set of the What We Eat In America (WWEIA) food categories. Then we mapped ~92,000 grocery food items to it. The model uses an inverse Cumulative Distribution Function sampling technique. Here we describe the model and report reliability metrics based on NHANES data from 2003-2010.

Keywords: Healthy Eating Index; WWEIA Food Categories; Statistical Modeling; NHANES Data; Grocery Food Purchases; Nutritional Sciences

1. Background and objective

There is sound literature-based evidence for the correlation between dietary habits and disease. This has led to a heightened interest in the monitoring and assessment of the nutritional status of patients with diet-related disease, as well as an interest in public health nutrition assessment. The common methods of nutritional monitoring

* Corresponding author: John F. Hurdle, MD, PhD. Tel.: +1-801-213-3232; fax: +1-801-581-4297.
E-mail address: john.hurdle@utah.edu
We want to evaluate an alternative approach: rather than measuring the quality of the foods consumers eat, we want to estimate the quality of the foods consumers buy. We plan to start with household grocery purchases, and our working hypothesis is that grocery data will be a strong predictor of overall dietary quality. Because grocery retailers collect sales data for internal purposes, such data promise to be inexpensive for research use. It is also free from self-report errors/bias and collecting it poses no burden on consumers. To accomplish this, we need a validated food quality metric paired with a way to apply that metric based solely on the counts of foods purchased.

1.1. The Healthy Eating Index 2010 (HEI-2010)

The Healthy Eating Index (HEI) as mentioned in Guenther et al. is a measure of overall diet quality that encapsulates the United States Dietary Guidelines for Americans (DGA). The HEI-2010 is the latest HEI iteration, and we use “HEI” here to refer to this version. One of the great strengths of the HEI is that it can measure quality at various levels of the food stream, e.g., the national food supply, the community food environment, restaurants, and individual food intakes as discussed in. This makes the HEI an excellent candidate as a scoring metric for our work.

The HEI assigns scores to 12 food components in a way that increases the overall quality score based on 8 food components that consumers should eat to maintain an adequate diet (e.g., total fruit, whole grains, seafood and plant proteins), while decreasing the overall quality score based on 4 components that consumers should choose in moderation (refined grains, saturated fats, sodium, and empty calories from solid fats, alcohol, and added sugars). The USDA encoded nutrition information from the ~8,000 foods in the Food and Nutrient Database for Dietary Studies (FNDDS) into the MyPyramid Equivalents Database (MPED), which summarizes the contribution of these foods to the components used by the HEI, based on standardized portions (100-gram ounce or cup equivalents). Using the MPED, which is now called the Food Patterns Equivalents Database (FPED), the HEI normalizes the amount of the foods eaten, per 100 grams, to each 1,000 kcal of total energy, and then scores the MPED-based food components. The HEI components and their corresponding scoring ranges are shown in Table 1.

1.2. The MyPyramid Equivalents Database (MPED) and the Food Patterns Equivalents Database (FPED)

MPED 2.0 provides the number of MyPyramid equivalents of the food groups and subgroups that are present in 100 grams of the 6940 food codes and 811 food modification codes reported in the WWEIA survey from 2003-2004. The Center for Nutrition Policy and Promotion (CNPP) addendum to the MPED 2.0 supplies data for 942 additional food items found in the WWEIA survey from 2005-2006. In addition, the CNPP provided supplemental data for whole fruit and fruit juices. Later replacing the MPED 2.0, the FPED provides the conversions for food codes and food modification codes reported in the WWEIA 2009-2010 surveys. There are a total of 8190 food codes.
Table 1. Scoring range and MyPyramid subgroups used in the calculation of the HEI component scores.

<table>
<thead>
<tr>
<th>HEI-2010 Components</th>
<th>Scoring range</th>
<th>MyPyramid subgroups and our constructed groups*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Vegetables</td>
<td>0-5</td>
<td>V_TOTAL, ALLMEAT*, LEGUMES</td>
</tr>
<tr>
<td>Greens and Beans</td>
<td>0-5</td>
<td>V_DRKGR, ALLMEAT*, LEGUMES</td>
</tr>
<tr>
<td>Total Fruit</td>
<td>0-5</td>
<td>F_TOTAL</td>
</tr>
<tr>
<td>Whole Fruit</td>
<td>0-5</td>
<td>WHOLEFRT</td>
</tr>
<tr>
<td>Whole Grains</td>
<td>0-10</td>
<td>G_WHL</td>
</tr>
<tr>
<td>Dairy</td>
<td>0-10</td>
<td>D_TOTAL</td>
</tr>
<tr>
<td>Total Protein</td>
<td>0-5</td>
<td>ALLMEAT*, LEGUMES</td>
</tr>
<tr>
<td>Seafood and Plant Proteins</td>
<td>0-5</td>
<td>SEAPLANT*, ALLMEAT*, LEGUMES</td>
</tr>
<tr>
<td>Fatty Acid Ratio</td>
<td>0-10</td>
<td>MFAT*, PFAT*, SFAT*</td>
</tr>
<tr>
<td>Sodium</td>
<td>0-10</td>
<td>SOD*</td>
</tr>
<tr>
<td>Refined Grains</td>
<td>0-10</td>
<td>G_NWHL</td>
</tr>
<tr>
<td>Empty Calories (SoFASS)</td>
<td>0-20</td>
<td>ADDSUG, DISCFAT_SOL</td>
</tr>
</tbody>
</table>

*The variables without an asterisk (*) are MyPyramid subgroups. The variables with an asterisk are our constructed variables, specified as follows: ALLMEAT* is the sum of 4 MyPyramid subgroups: M_MPFE, M_NUTSD, M_SOY, M_EGG; SEAPLANT* is the sum of 4 MyPyramid subgroups: M_FISH_HI, M_FISH_LO, M_SOY, M_NUTSD; MFAT*, PFAT*, SFAT*, and SOD* are not specified within the MyPyramid but are computed in the model from FNDDS food codes mapped to WWEIA food categories.

1.3. The HEI-2010 scores for grocery purchases

In future work we are interested in using the HEI as a scoring metric to measure household food quality via grocery purchases. To use the HEI directly we would need knowledge of both the quantity of food and the overall caloric value of that food to calculate its component scores, but we lack both in grocery retail transaction data. For example, we received from our grocery retail partner a data set of sales records from 144,000 households in three geographical regions spanning a period of 15 months from January 2012 to March 2013. We identified 92,062 distinct food items in the data set that we attempted to map to foods coded in the FNDDS. We were only partially successful in building that map, and we discovered that weight/volume data were incomplete for most of the grocery items. We needed an alternative food categorization scheme that could map to our grocery sales items and that could also map directly to the MPED, so we could use the HEI. We chose the 150 food categories of WWEIA, augmented by a set of pseudo-categories for whole-grain foods (which WWEIA lacked). The earlier work of Brinkerhoff et al. discovered that grocery retailers organize foods in a way similar to the way nutritionists categorize foods. Figure 1 is an example of the mapping between the organization of foods from a grocery retailer and the categorization of foods by USDA Nutrient Database. As discussed in Brinkerhoff et al., grocery retailers use hierarchical categories to organize the supermarket food items. The most granular category in this hierarchy was entitled “sub-commodity”. In the grocery database’s organization of foods, similar food items were assigned into sub-commodities. We exploited the natural correspondence between the grocery database’s organization of foods and the categorization of foods into the food categories of WWEIA to map the grocery food items to the less-granular WWEIA food categories. Examples of sub-commodities in our grocery retail data include soy milk, chicken thighs, chicken wings, and ground chicken. In the other hand, examples of WWEIA food categories include milk substitutes (grouped food codes of soy milk, coconut milk, rice beverage, etc.) and chicken (grouped of chicken wings, thighs, drumsticks, etc.). Our dietitians mapped and verified the mappings of 1887 grocery sub-commodities containing a substantial
majority of 92,062 distinct food items to the WWEIA food categories. If a corresponding WWEIA food category was found for a given sub-commodity, then all the food items within the sub-commodity were automatically assigned to the discovered WWEIA food category. The food items in the sub-commodities mapped to multiple WWEIA food categories were hand-mapped by dietitians. Since the Whole Grain category was not clearly identified in the WWEIA groups, dietitians hand-mapped food items from the grocery data to pseudo-WWEIA numbers for this component in the scoring. The linkage between the FNDDS food codes and grocery food items via the WWEIA food categories is shown in Figure 2.

1.4. The NHANES dataset

The National Health and Nutrition Examination Survey (NHANES) assesses the health and nutrition of adults and children in the United States. The survey combines interviews and physical examinations. To assess nutrition, NHANES uses experienced interviewers to solicit 1-2 days of dietary recall. NHANES data are obtained through a complex, multistage, probabilistic sampling strategy that selects participants who are representative of the civilian, non-institutionalized U.S. population. Over-sampling of certain population subgroups increases the reliability and precision of health status indicator estimates for those subgroups. In the work presented here we opted to combine the four NHANES nutrition data sets from 2003-2010. This provides a large and well-balanced food quality profile. The nutrition data in NHANES are coded so as to support the direct calculation of the HEI for individual respondents; i.e., the foods in the survey have been mapped to the FNDDS. For analysis of the NHANES dietary intakes in the WWEIA surveys, the USDA’s Food Surveys Research Group (FSRG) provides mappings between the FNDDS food codes used for calculating the HEI and the WWEIA food categories.

1.5. Objective

We report below the results of constructing a computer model that estimates the Healthy Eating Index 2010 from count data alone, and we provide reliability measures based on National Health and Nutrition Examination Survey data sets spanning 2003 to 2010.
2. Methods

Table 1 shows, in addition to the standard HEI components and their scores, the MyPyramid subgroups we used in this work; it also shows the several constructed variables that are necessary for the calculation of the corresponding HEI component scores. The nutrient density of the MyPyramid subgroup \( M \) for a set of \( n \) consumed food items \((F_1, F_2, F_3, ..., F_n)\) is defined as follows:

\[
density_M(F_1, F_2, F_3, ..., F_n) = \frac{\sum_{j=1}^{n} M(F_j) W(F_j)}{100} \frac{1}{C(F_1, F_2, F_3, ..., F_n)}
\]  

where \( M(F_j) \) and \( W(F_j) \) are, respectively, the MyPyramid equivalents of the subgroup \( M \) in 100 grams, and the weight of the food item \( F_j \) and \( C(F_1, F_2, F_3, ..., F_n) \) is the total calories of all consumed food items. The nutrient density equation (1) may also be written as follows:

\[
density_M(F_1, F_2, F_3, ..., F_n) = \sum_{j=1}^{n} subden_M(F_j)
\]  

where

\[
subden_M(F_j) = \frac{10 M(F_j) W(F_j)}{C(F_1, F_2, F_3, ..., F_n)}
\]  

Figure 3 illustrates the steps for calculating the HEI-2010 component scores for a set of consumed FNDDS food codes. The conversion from a weighted composition of the FNDDS food codes reported in the What We Eat in America (WWEIA) surveys to the MyPyramid subgroup nutrient densities is enabled by the MPED. The HEI-2010 standards and the application of these standards to derive the HEI-2010 component scores as shown in Figure 2 are discussed and implemented in Guenther et al.\textsuperscript{11} and the USDA SAS code.\textsuperscript{14}
Since a complete mapping between the FNDDS food codes and the grocery food items is not readily available, the steps illustrated in Figure 3 cannot be directly applied to obtain the HEI-2010 for a set of purchased grocery items. Without a complete mapping between the FNDDS food codes and the grocery food items, the nutrient densities of MyPyramid subgroups for a set of grocery food items can not be obtained using Equation (1). To measure the dietary quality of grocery purchases, we aimed to derive an estimation model of the HEI-2010 based on a crosswalk between the FNDDS food codes and the grocery food items, using WWEIA categories as the shared intermediary.

The USDA provides a mapping between the WWEIA food categories and the FNDDS food codes, as well as a mapping between the FNDDS food codes and the MPED. Thus WWEIA (augmented with whole-grain foods) provides a convenient crosswalk to the nutrient densities of MyPyramid subgroups starting from either grocery codes or FNDDS food codes. The WWEIA category schema groups similar foods and beverages together based on nutrient content. WWEIA food categories are labeled with a unique four-digit code and contain discrete food items - each FNDDS food code is placed in one of the mutually exclusive food categories.

Let $WWEIA\_Cat\_Map$ denote the mapping from the FNDDS food codes into the WWEIA food categories. Let $Grocery\_WWEIA\_Map$ denote the mapping between the food UPCs and WWEIA food categories. With these two maps we have established a crosswalk between the food UPCs and the FNDDS food codes using WWEIA as the intermediary. In order to obtain the distribution of nutrient densities of MyPyramid subgroups across the WWEIA food categories, we computed the value of $\sum_{m} F_m$ in Equation (3) for all FNDDS food codes and MyPyramid subgroups in Table 1 for all the NHANES respondents from 2003-2010. Using these computed values and the $WWEIA\_Cat\_Map$, we generated the cumulative distribution functions (CDFs) of these computed values organized by the MyPyramid subgroups for each of the WWEIA food categories, excluding WWEIA’s “Other” food category. For example, the food codes in WWEIA category 6002 (Apples) make significant contributions only to the nutrient density of WHOLEFRT (whole fruit), while the food codes in WWEIA category 2502 (Eggs and omelets) make significant contributions to the nutrient densities of ALLMEAT (proteins), D TOTAL (dairy), and DISCFAT SOL (solid fats). For illustration, the MPED CDFs for the WWEIA categories “Apples” and “Eggs and omelets” are shown in Figure 4. We also include in Figure 5 a diagram that illustrates the crosswalk to the nutrient densities of MyPyramid subgroups starting from either grocery codes or the FNDDS food codes.

![Fig. 3. The HEI-2010 for a set of consumed FNDDS food codes.](image-url)
Fig. 4. The cumulative distribution functions (CDFs) of the nutrient densities of the MyPyramid subgroups for WWEIA food categories 2502 and 6002.

Fig. 5. Crosswalk to the nutrient densities of MyPyramid subgroups starting from either grocery codes or the FNDDS food codes.
To construct a count-based HEI model, we start by replacing Equation (2) above with the values computed from the following equation as approximations:

\[
    \text{density}_M(F_1, F_2, F_3, \ldots, F_n) = \sum_{j=1}^{n} \text{random\_samp}_M(WWEIA\_Cat\_Map(F_j))
\]  

(4)

where \( \text{random\_samp}_M \) is a function that generates a random sample based on the CDF of the MyPyramid subgroup \( M \) for the category \( WWEIA\_Cat\_Map(F_j) \). We count the number of each food item of interest (e.g., an NHANES respondent, the entire 15-month store sample, a single store, a single household, etc.), and map those items to the WWEIA categories. Then we use uniform random sampling of the inverse-CDF for that category, repeated once for each entry, and compute the mean. For example, if a food sample had 25,000 items in the Dairy_Total WWEIA category, we sample the inverse-CDF 25,000 times and compute its mean over the 25,000 samples. In this case, the inverse-CDF provides the estimated dairy food density, and thus its contribution to the corresponding MPED category D_Total. Those MPED densities, in turn, can be used to calculate the HEI component score for Dairy using the standard SAS-based calculator provided by the USDA.\(^{14}\) In all cases, this is repeated 100 times to smooth variation resulting from the probabilistic sampling strategy.

Figure 6 shows how the model could be used to estimate the HEI component scores for a grocery market basket. The market basket data provides the food item counts, and \( Grocery\_WWEIA\_Map \) provides the crosswalk to the MPED. From those, the HEI can be calculated as noted above.

The component scores Fatty Acid Ratio and Sodium are calculated using the values for saturated fat, polyunsaturated fat, monounsaturated fat, and sodium, but these are not included in the MPED. In order to estimate these two component scores, we used a similar approach to estimating the other component scores, but here we compute the CDFs for these variables across all WWEIA food categories, using the FNDDS food codes that each one contains to fetch nutrient information from FNDDS.
3. Results

In future work we plan to verify the model’s performance assessing the quality of household market baskets. Here we assess the reliability of the model estimator using the NHANES datasets (2003-2010). For each of the NHANES respondents, we computed the “estimate” scores using our estimation model and then compared them with the “true” scores using the USDA SAS-based HEI calculator. With our estimation model, the “estimate” scores for NHANES respondents are obtained via the steps given in Figure 7. On the other hand, the “true” scores are obtained via the HEI-2010 SAS-code. We eliminated data for children who are under two years of age. For each of the remaining NHANES respondents, we separated the foods eaten at home and foods away from homes, and treated them as two individual respondents. We dropped respondents who reported eating no food components in these two categories. This left 32,085 total respondents reporting eating foods at home and 21,633 respondents reporting eating foods away from home.

We used the Cronbach’s alpha reliability coefficient to measure the internal consistency between the “true” and “estimate” component scores of all selected NHANES respondents. The Pearson’s correlation coefficients were also used to measure the relationship between the “true” and “estimate” component scores. The Cronbach’s alpha reliability coefficients and the Pearson’s correlation coefficients that measure the pairwise internal consistency and correlation between the “estimate” and “true component scores for foods at home and foods away from home are given in Table 2.

The range of the Cronbach’s alpha reliability coefficients is 0.0-1.0. The closer the coefficient is to 1, the greater the internal consistency of the items. The range of the Pearson’s correlation coefficients is -1.0 - 1.0. A value of 1 or -1 indicates a perfect positive or perfect negative linear relationship, respectively, between the two variables.

![Fig 7. The estimation model of the HEI-2010 applied to NHANES (2003-2010) respondents](image-url)
Table 2. The standardized Cronbach’s alpha coefficients and the Pearson correlation coefficients show the internal consistency and the pairwise correlations between “estimated” and “true” component scores for NHANES respondents reporting foods at home (i = 21,633) and foods away from home (n = 32,085). *denotes Cronbach’s alpha scores or Pearson’s correlation coefficients that failed to meet our pre-defined adequacy thresholds.

<table>
<thead>
<tr>
<th>HEI-2010 Components</th>
<th>Foods from home</th>
<th>Foods away from home</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standardized</td>
<td>Pearson’s Correlation</td>
</tr>
<tr>
<td></td>
<td>Cronbach’s alpha</td>
<td>Coefficients</td>
</tr>
<tr>
<td></td>
<td>Coefficients</td>
<td>Coefficients</td>
</tr>
<tr>
<td>Total Vegetables</td>
<td>0.84</td>
<td>0.72</td>
</tr>
<tr>
<td>Greens and Beans</td>
<td>0.84</td>
<td>0.72</td>
</tr>
<tr>
<td>Total Fruit</td>
<td>0.94</td>
<td>0.89</td>
</tr>
<tr>
<td>Whole Fruit</td>
<td>0.96</td>
<td>0.93</td>
</tr>
<tr>
<td>Whole Grains</td>
<td>0.78</td>
<td>0.64</td>
</tr>
<tr>
<td>Dairy</td>
<td>0.86</td>
<td>0.76</td>
</tr>
<tr>
<td>Total Protein</td>
<td>0.89</td>
<td>0.79</td>
</tr>
<tr>
<td>Seafood and Plant Proteins</td>
<td>0.90</td>
<td>0.82</td>
</tr>
<tr>
<td>Fatty Acid Ratio</td>
<td>0.82</td>
<td>0.70</td>
</tr>
<tr>
<td>Sodium</td>
<td>0.33*</td>
<td>0.20*</td>
</tr>
<tr>
<td>Refined Grains</td>
<td>0.56*</td>
<td>0.40*</td>
</tr>
<tr>
<td>Empty Calories (SoFAAS)</td>
<td>0.46*</td>
<td>0.30*</td>
</tr>
</tbody>
</table>

The HEI’s mean component scores derived from NHANES data is considered a good measure of the U.S. population’s long-term and usual intake, so we evaluated the performance of our estimation model in obtaining the mean component scores across the NHANES respondents spanning 2003-2010. The resulting “true” and “estimated” mean component scores for NHANES respondents reporting food eaten at home and food eaten away from home are shown in Figures 8 and 9, respectively.

Fig 8. The means of the “true” and “estimate” HEI component scores reported respondents by NHANES respondents (2003-2010) for foods at home.
4. Discussion

Table 2 show a strong internal consistency and a near linear relationship for several important food-based HEI components (vegetables, fruits, dairy, whole grains, and proteins) and the fatty-acid ratio, but weak correlations for refined grains, sodium, and SoFAAS. This holds for both foods eaten at home and for foods away from home. We consider a Cronbach’s alpha level of 0.70 to be the threshold for adequate internal reliability. For the Pearson correlations, coefficients in the range of 0.60-0.79 are considered strong and coefficients in the range of 0.80-1.00 are considered very strong. With this metric the model did poorly with the same three component scores for refined grains, sodium, and SoFAAS. However the model showed a strong performance for total vegetables; greens and beans; whole grains; dairy; total protein; and the fatty acid ratio, and proved very strong for whole fruit; total fruit; and seafood and plant proteins. As shown in Figures 8 and 9, the “estimated” means and “true” means matched very well for all component scores except, again, for the components refined grains, sodium, and SoFAAS.

To improve the model, we clearly need to improve its performance on refined grains, sodium, and the SoFAAS. We anticipated that the model would perform poorly with sodium and the SoFAAS. The CDFs for these two components were computed in an indirect way based on their distribution in the FNDDS foods, and this proved to be an inadequate approximation. In future work we plan to model these two components separately, stepping away from the CDF approach to exploit association rules and “sentinel foods” as features in a machine learning model. The model’s performance on refined grains is puzzling. We suspect that there may be systematic errors in our Grocery_WWEIA_Map, so we will work with our dietitians to reassess that map.

4. Conclusion

We present here a robust model that does a very good job estimating HEI component scores for nine important dimensions of purchased-food quality, and it does so using only the counts of food items. Like the HEI itself, the
model can be applied to any source of food data. To apply the model, one only needs to be able to count food items of interest and be able to map them to the fairly granular WWEIA food categories.

Acknowledgements

This work was supported by NIH grant T15-LM007124, as well as by two University of Utah institutional innovation seed grants. We are indebted to Patricia Guenther for her insights into the HEI, and to Kristine Jordan and Jennifer North for their help with the mapping described in this manuscript.

References