What Percentage of Children in the U.S. are Eating a Healthy Diet?

Raymond J. Carroll
Department of Statistics
Faculty of Nutrition
Center for Statistical Bioinformatics
Institute for Applied Mathematics and Computational Science
Texas A&M University
http://stat.tamu.edu/~carroll
Surveillance Measurement Error Group

Sue Krebs-Smith, Dennis Buckman, Raymond Carroll, Kevin Dodd, Larry Freedman, Patricia Guenther, Victor Kipnis, Doug Midthune, Amy Subar, Janet Tooze
Outline

- Context
- The Healthy Eating Index-2005
- Episodically consumed foods and their properties
- Results
- Modeling
Background and goals → Major questions → HEI 2005 → Usual intake

Assumptions and model → Computational issues → Cancer and diet
The 24HR Recall

- The 24HR is a good measure of intake on a single day, but as a measure of usual intake it does not account for day-to-day variability.

- The sample mean 24HR value can be used as an estimate of the population mean usual intake.

- The sample distribution of 24HR is not a good estimate of the population distribution of usual intake.
The 24HR Recall

- Nutritionists want to understand longer term average daily intake as measured by an instrument, not intake on 1 day

- We call this usual intake

- That we cannot observe usual intake is called measurement error
Context

• The data we use are from the NHANES survey of children aged 2-8 in the U.S.

• The data set has 2,300 children with a 24hr

• There are roughly 700 with two 24hr

• This is a real survey, and survey weights are incorporated into the analysis
Background and goals

Major questions

HEI 2005

Usual intake

Assumptions and model

Computational issues

Cancer and diet
Our goal is to give realistic estimates of dietary intake distributions and dietary patterns that account for the day-to-day variability.

We also want to estimate the real relationship between nutrition and cancer, while accounting for day-to-day variability.

Today we focus most effort on the first problem. But we also do an analysis of the second.
The Healthy Eating Index - 2005

- The HEI-2005 is a composite score that is based on the USDA 2005 Dietary Guidelines

- It is a multi-component dietary quality index involving ratios of interrelated dietary components to energy (calories)

- It ranges from 0 to 100, with higher scores better
The **Healthy Eating Index (HEI)** is a measure of diet quality that assesses conformance to Federal dietary guidance. The original HEI was created by the U.S. Department of Agriculture (USDA) in 1995. Release of new Dietary Guidelines for Americans in 2005 motivated a revision of the HEI. The food group standards are based on the recommendations found in My Pyramid (see Britten et al., *Journal of Nutrition Education and Behavior* 38(6S) S78-S92). The standards were created using a density approach, that is, they are expressed as a percent of calorie or per 1,000 calories. The components of the HEI-2005 and the scoring standards are shown below.

<table>
<thead>
<tr>
<th>Component</th>
<th>Maximum points</th>
<th>Standard for maximum score</th>
<th>Standard for minimum score of zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Fruit (includes 100% juice)</td>
<td>5</td>
<td>≥0.8 cup equiv. per 1,000 kcal</td>
<td>No Fruit</td>
</tr>
<tr>
<td>Whole Fruit (not juice)</td>
<td>5</td>
<td>≥0.4 cup equiv. per 1,000 kcal</td>
<td>No Whole Fruit</td>
</tr>
<tr>
<td>Total Vegetables</td>
<td>5</td>
<td>≥1.1 cup equiv. per 1,000 kcal</td>
<td>No Vegetables</td>
</tr>
<tr>
<td>Dark Green and Orange Vegetables and Legumes</td>
<td>5</td>
<td>≥0.4 cup equiv. per 1,000 kcal</td>
<td>No Dark Green or Orange Vegetables or Legumes</td>
</tr>
<tr>
<td>Total Grains</td>
<td>5</td>
<td>≥3.0 oz equiv. per 1,000 kcal</td>
<td>No Grains</td>
</tr>
<tr>
<td>Whole Grains</td>
<td>5</td>
<td>≥1.5 oz equiv. per 1,000 kcal</td>
<td>No Whole Grains</td>
</tr>
<tr>
<td>Milk</td>
<td>10</td>
<td>≥1.3 cup equiv. per 1,000 kcal</td>
<td>No Milk</td>
</tr>
<tr>
<td>Meat and Beans</td>
<td>10</td>
<td>≥2.5 oz equiv. per 1,000 kcal</td>
<td>No Meat or Beans</td>
</tr>
<tr>
<td>Oils</td>
<td>10</td>
<td>≥12 grams per 1,000 kcal</td>
<td>No Oil</td>
</tr>
<tr>
<td>Saturated Fat</td>
<td>10</td>
<td>≤7% of energy</td>
<td>≥15% of energy</td>
</tr>
<tr>
<td>Sodium</td>
<td>10</td>
<td>≤0.7 gram per 1,000 kcal</td>
<td>≥2.0 grams per 1,000 kcal</td>
</tr>
<tr>
<td>Calories from Solid Fats, Alcoholic beverages, and Added Sugars (SoFAAS)</td>
<td>20</td>
<td>≤20% of energy</td>
<td>≥50% of energy</td>
</tr>
</tbody>
</table>
A component of HEI-2005 is dark green and orange vegetables and legumes, which we call **DOL**.

Another component is calories from solid fats, alcoholic beverages and added sugars, which we call **SoFAAS**.

- These might be thought of as “empty calories”. They are not good for you!
The Healthy Eating Index - 2005

- The 13 components of the HEI-2005 are **Total Fruits**, **Whole Fruits**, **Total Grains**, **Whole Grains**, **Total Vegetables**, dark green and orange vegetables and legumes (**DOL**), **Milk**, **Meat and Beans**, **Oil**, **Saturated Fat**, **Sodium** and calories from solid fats, alcoholic beverages and added sugars (**SoFAAS**).
The scores assigned to each component are nonlinear functions with truncations.

Total fruits for example are measured as:

\[
\text{Adjusted Total Fruits} = \frac{\text{Cups}}{\text{Energy/1000}}
\]

The score increases linearly up to 0.8 equivalents per 1,000 kilocalories with a maximum score of 5, and does not increase with intakes above 0.8 cup equivalents per 1,000 kilocalories.
The Healthy Eating Index - 2005

• The score for **SoFAAS** as a percentage of energy is

\[
\begin{align*}
  &= 0 \quad \text{if } \geq 50 \\
  &= 20 \quad \text{if } \leq 20 \\
  &= \text{linearly interpolated otherwise}
\end{align*}
\]
Goal

- What is the distribution of dietary pattern scores, such as the HEI-2005 or the Mediterranean index?
The Healthy Eating Index - 2005

• Now comes the hard part.

\[
\text{Adjusted Whole Fruits} = \frac{\text{Cups}}{\text{Energy/1000}}
\]

• In this formula, “Cups” is long-term daily average number of cups consumed.

• “Energy” is long-term daily average number of calories consumed

• These are called Usual Intakes
HEI Total Score Histogram (Density)

The 24hr overestimates the % with diet scores < 40

It also overestimates the % with diet scores > 80
One 24hr is too pessimistic: thinks 25% have total score < 40 ("alarming")

We estimate that 8% of children have an HEI-2005 total score < 40

Major policy implications?
Note that essentially no children in the U.S. have a total HEI score of greater than 80.
Recently, the **White House Task Force on Obesity** was about to announce a goal that **all children would have a HEI-2005 usual intake total score > 80**

- The 99th percentile = 79.4. 😊
A Vignette

• Confronted with the results, the Task Force changed their goal to have children move to a mean of 80.

• However, the mean is 53, so there is a long way to go.
Dietary Reality

• **Assumption**: 24HR’s are unbiased measures of usual intake as measured by multiple 24HR
Note here that a single 24hr is shifted left compared to usual intake, although the means are the same due to some unusually high days of intake.

Source: EATS
Transformations

• To deal with the skewness, it is typical to transform the data so that day-to-day variation has a nice Gaussian-like distribution.

• One analyzes in this transformed scale, and then back-transforms to the original nutrient scale


Modeling With Transformations

Original Scale

Transformed Scale

Transform

Backtransform

Source: EATS
Episodically Consumed Foods

- The HEI-2005 has 6 components that are episodically consumed. Among children aged 2-8 in the U.S., here are the percentages of reported non-consumption on a 24hr
  - Total Fruit = 17%
  - Whole Fruit = 40%
  - Whole Grains = 42%
  - Total Veggies = 3%
  - DOL = 50%
  - Milk = 12%
Episodically Consumed Foods

- Observed food intakes are often zero

![Whole Grains](chart.png)
Model for a Single Food

- For a single food, we have developed a flexible modeling framework, which we call the **NCI Method**

- For SAS programs based on NL MIXED, see [http://riskfactor.cancer.gov/diet/usualintakes/](http://riskfactor.cancer.gov/diet/usualintakes/)

- More stable MCMC methods are described in a paper in IJ B and *Annals of Applied Statistics*
Model for a Single Food

• It is possible to get estimates of the distribution of each energy-adjusted dietary component and each HEI-2005 dietary score component, **SEPARATELY**

• This approach allows estimating the mean of the HEI total score in a population

• It does not allow estimation of **percentiles** of the HEI-2005 total score

• Percentiles require a multivariate model
So, What’s the Big Deal?

• HEI is complex, because it has 6 episodically consumed foods, 6 daily-consumed foods and nutrients, and energy.

• The newest HEI is much larger in number of components
So, What’s the Big Deal?

• The bottom line is that when we turn to things like the HEI-2005, we have four problems.

• **Problem #1:** The dimensionality of the integration is too great for PROC NLMI XED
So, What’s the Big Deal?

- **Problem #2**: Figure out a model that can allow analysis of HEI-2005

- **Problem #3**: Overcome certain technical issues about patterned covariance matrices

- **Problem #4**: Compute!
Multivariate Model

- Generically, $X$ will denote covariates
  - Demographics
  - Food frequency questionnaire if available

- Generically, $U$ denotes how people with the same covariates differ from one another in their long term intake

- Finally, $\varepsilon$ will denote day-to-day variability
A Multivariate Model

• I will just do 2 foods plus energy here, and briefly mention what happens with many foods, nutrients and energy.

• We have to reformulate the consumption model to allow day-to-day energy to be correlated with day-to-day consumption.

• We use a choice-based probit model for this task.
A Multivariate Model

- Here, $i$ will denote person
- Also, $k$ will denote replicate of the 24HR
- Finally, $j$ will denote an index
- There are $M_1 = 6$ episodically consumed foods
- There are $M_2 = 6$ nutrients
- There is also energy
- I will illustrate in the case of 2 foods and energy

$$Y_{Eijk} = \text{Energy}$$
For $j=1,3$, define a latent variable

$$Q_{Fijk} = X_i^T \beta_j + U_{ij} + \epsilon_{ijk}$$

The epsilon term describes day-to-day variability for an individual: some days a food is consumed, some days it is not.

For technical reasons (identifiability), the epsilon terms are normally distributed with variance $= 1$. 
Consumption?

- For j=1,3, define a latent variable

\[ Q_{Fijk} = X_i^T \beta_j + U_{ij} + \varepsilon_{ijk} \]

- The U term is a random variable indicating that two people with exactly the same demographics and other covariates have different probabilities of consumption.
Consumption?

• For j=1,3, define a latent variable

\[ Q_{Fijk} = X_i^T \beta_j + U_{ij} + \varepsilon_{ijk} \]

• Consumption of the food for person i on day k is distributionally equivalent to a probit model defined through

\[ Q_{Fijk} > 0 \]
• For j=1,3, define a latent variable

\[ Q_{Fijk} = X_i^T \beta_j + U_{ij} + \epsilon_{ijk} \quad Q_{Fijk} > 0 \]

• The probability of consuming a food on a given day for a particular individual is

\[ \Phi(X_i^T \beta_j + U_{ij}) \]

\[ \Phi \text{ is the normal distribution function} \]
Amount

• When a food is consumed, it is positive, so we use transformations

• The Box-Cox transformation is denoted by

\[ g_{tr}(x, \lambda) = \frac{x^\lambda - 1}{\lambda} \]
Amount

- Once again, the epsilon terms will describe day-to-day variability.

- One again, the U terms will describe how people with the same covariate pattern differ from one another
Amount

- For $j = 2, 4$, we have a second latent variable, involving consumption of the food.

$$g_{tr}(Q_{Fijk}, \lambda_j) = X_i^T \beta_j + U_{ij} + \varepsilon_{ijk}$$

- We get to observe this latent variable only if there is consumption
Energy

• Energy is always positive, so for \( j = 5 \), we observed

\[
g_{tr}(Y_{Eijk}, \lambda_j) = X_i^T \beta_j + U_{ij} + \varepsilon_{ijk}
\]

• We assume that

\[
\tilde{U}_i = (U_{i1}, \ldots, U_{i5}) = \text{Normal}(0, \Sigma_u)
\]

\[
\tilde{\varepsilon}_{ik} = (\varepsilon_{i1k}, \ldots, \varepsilon_{i5k}) = \text{Normal}(0, \Sigma_{\varepsilon})
\]
Background and goals → Major questions → HEI 2005 → Usual intake → Assumptions and model

Computational issues → Cancer and diet
Technical Issue

• The biggest technical issue is that $\sum_\varepsilon$ is restricted, with structural 1’s along the diagonals related to consumption.

• The analysis of nonlinear mixed effects models with a restricted covariance matrix is extremely difficult.

• We worked out a parameterization with no constraints (important for computation).
There are many **technical issues** related to fitting the resulting model.

These are of great interest to statisticians, but may be less interesting (or boring) for everyone else.

The full details can be found in a paper in the *Annals of Applied Statistics*. See the (8 page!) appendix.
The computational issue is that the components of the day-to-day variability, the epsilons, are all correlated.

So too are the components of the individual usual intake, the U’s.

Maximum likelihood requires integration (area under the curve of a function)
**Computation**

- We have developed a fully parametric model, with transformations and lots of latent variables.

- All parameters are free to roam or on a fixed interval.

- Even so, standard software will not work (PROC NL MIXED dies, hurts your computer, for example)
The most commonly used computational method to do the integration is called **Markov Chain Monte Carlo**.

It generally uses what are called **Gibbs sampling** and **Metropolis-Hastings steps**.

It is an iterative numerical procedure; in this particular case, we had to write our own program to do the computation.
Computation

- Statisticians have made huge gains in computing integrals using Monte-Carlo techniques

- The literature is vast
Result Verification

• For each food and nutrient, it is possible to use standard nonlinear mixed effects software to get the distribution of adjusted usual intake and HEI-2005 component score, one at a time, not jointly.

• Our results are in very close agreement with these results, but we can also do the multivariate case.
Relationship with Health Effects

- We have applied the model to the analysis of the NIH-AARP Diet and Health Study.
- The outcome was colorectal cancer, separately for men and women.
- The general goal is to study association of dietary patterns, assessed using dietary quality indices adjusted for measurement error, and a health outcome.
Relationship with Health Effects

• The effect of the day-to-day variability (measurement error) in the 24 hour recalls is generally to attenuate relative risks, i.e., make them closer to 1.0 (no effect), and hence make them look not very important.

• Relative risks below 1.0 indicate that higher scores lead to less disease.

• We measure relative risk as the ratio of disease risk from the 10th to the 90th percentile.
Relationship with Health Effects

- We did a survival analysis using person years
- **The first analysis** done was using the FFQ for the HEI-2005 total score as well as energy
- **The second analysis** was a measurement error corrected analysis, based on regression calibration to understand the effect of usual intake, not the FFQ
- The same covariates were used to fit the HEI-2005 total score model in a calibration sub-study
Relationship with Health Effects

- We did a survival analysis using person years

- Variables in the model include age, ethnicity, education, BMI, smoking status, physical energy, energy and hormone replacement therapy (for women)

- The HEI total score was also in the model, in a loglinear continuous risk model
Relationship with Health Effects

• We have applied the model to the analysis of the NIH-AARP Diet and Health Study

• Using the HEI-2005 total score from the FFQ, the relative risk for going from the 10\textsuperscript{th} to the 90\textsuperscript{th} percentile for women is estimated as \textbf{0.80}

• After measurement error correction, it is \textbf{0.62}

• Note the attenuation in the FFQ that we expected
Relationship with Health Effects

• We have applied the model to the analysis of the NIH-AARP Diet and Health Study

• Using the HEI-2005 total score, the relative risk for going from the 10th to the 90th percentile for women is estimated as 0.62

• For men, it is 0.45
Summary

• 24hr recalls have great day-to-day variability

• Adjusting for this variability to estimate the distributions of usual intakes of multiple episodically consumed foods and nutrients is extremely challenging.

• We have provided the first solution to the problem
Summary

• The methods allow us to understand dietary patterns, estimate distributions, consider risk models, etc.

• The NCI has a working version of the model fitting in SAS, which is under development

• Children in the US do not have ideal diets 😊

• Eating healthy protects against colon cancer