Data reduction techniques to classify plant based dietary patterns: Pros and cons

David R. Jacobs, Jr, PhD
Mayo Professor of Public Health
University of Minnesota
6th International Conference on Vegetarian Nutrition
Loma Linda, Feb 26, 2013
Disclosure

Unpaid consultant California Walnut Commission
Hierarchical structure of dietary patterns, foods, and nutrients for study of food synergy

<table>
<thead>
<tr>
<th>Food synergy level</th>
<th>Examples of dietary component at each synergy level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 5: Dietary pattern</td>
<td>‘Prudent diet’, ‘Western diet’, other combinations of food groups</td>
</tr>
<tr>
<td>Level 4: Food groups</td>
<td>Whole grain, dairy, fruit, vegetables, meat</td>
</tr>
<tr>
<td>Level 3: Whole grain</td>
<td>Whole wheat, brown rice, rolled oats</td>
</tr>
<tr>
<td>Level 2: Whole wheat</td>
<td>Bran, germ, endosperm; extract of fat-soluble portion</td>
</tr>
<tr>
<td>Level 1: Bran or a single phytochemical</td>
<td>Specific nutrients or phytochemicals</td>
</tr>
</tbody>
</table>

Dietary Patterns: Statistical Construct or Statistical Discovery?

- To a large extent, individuals do not consider dietary pattern when deciding what to eat
  - Vegetarianism is an example of a dietary pattern decision, but it only determines 1 aspect of the diet (not eating meat); individuals may not balance their intake to fit a pattern over months or years
- A consistent finding in nutritional epidemiology is that dietary patterns predict future chronic disease and death
- Different dietary patterns tend to resemble each other with common elements
- The A Priori Diet Quality Score has correlations over 18-20 years of about 0.6 in CARDIA and IWHS, 0.74 over 1.3 years in AHS-2
- These correlations and ability to predict suggest that dietary pattern is a personal characteristic over long periods of time, even if individuals are not aware of their intake as a pattern.
Different Statistical Methods

• In CARDIA, we formed the A Priori Diet Quality Score and 2 principal components (PCA). The latter 2 scores are data-driven.
  • $r^2$ for estimating the a priori score simultaneously from the “meat and refined grain” and “fruit/vegetable” PCAs ~ 0.6 at each exam
  • Something links the a priori or guideline ideas with consistent dietary choice or consumption, such that some people follow the a priori ideas, while others oppose them
Different Statistical Methods

- In CARDIA, we also formed 2 cluster scores, dichotomies that we called “prudent” and “western”. These 2 scores are data-driven.
- \( r^2 \) for estimating the a priori score simultaneously from the dichotomies are \( \sim 0.38 \) at each exam, lower than for PCA in part because the cluster scores have only 2 levels.
- Again, something links the a priori or guideline ideas with consistent dietary choice or consumption, such that some people follow the a priori ideas, while others oppose them.
Different Statistical Methods

• In IWHS, we formed the A Priori Diet Quality Score and two version of the Alternative Healthy Eating Index (2002, McCullough and 2010, Chiuve). These scores are all a priori.
  • $r^2$ for estimating the a priori score from the AHEI is about 0.6, reflecting similarities in the a priori thinking of the respective score authors
  • The correlation of the two AHEIs is about 0.7, while the tracking correlation over 18 years was about 0.55 for A Priori Diet Quality Score and 0.45 for AHEI
Different Statistical Methods

• Correlations and tracking with the Trichopolou Mediterranean Diet score are lower
  • This score has only 9 points, based on within population distributional medians of the different dietary variables
  • Tracking is probably lower because it misses nuances that the more complex scores capture
Complexity in Nutrition and Biology

• Nutrition is biology. Both are extraordinarily complex. Some simple pathways are discernible, especially in medicine, where the goal is to interrupt a problematic or “misfiring” pathway to restore better health. Fundamentally, however, complexity is the rule. Solutions that incorporate complexity are likely to answer questions that reductionist approaches cannot address.

• In statistical theory, data reduction is a method for reducing dimensionality of a multivariable problem to isolate one or a few dimensions that capture the essence of the full, multivariable data. This reduction in dimensionality enables the analyst to make simple, yet still comprehensive statements.

• In the field of dietary patterns, researchers initially took this perspective to reduce the complex nutritional problem encompassing many foods and nutrients consumed to a single score. They employed factor/principal components analysis, which adds several dietary elements together.
Summing correlated elements spreads them out

- Many dietary patterns combine elements by using a weighted sum.
- If the elements are uncorrelated, when one element goes up the other may go up or down, and the sum increases proportional to the number of elements added.
- If the elements are positively correlated, however, they tend to go up together or down together. If one of the variables goes up, the sum goes up disproportionately to the number of elements added.
- Thus the distribution of the sum of elements tends to spread out.
- Now suppose there is risk associated with increasing level of each element. That risk goes up with the increasing levels of the elements.
- If there are risk synergies, the combinations stay intact and the risk gradient over the summed elements reflects the synergistic risk increase.
- Thus a pattern may reflect risk better the sum of the expected risk for each element.
Example of the effect of summing risk over elements

1. 8 people with 3 dichotomous risk elements and arbitrary risk scale
2. Element risk: 1 for no, 5 for yes
3. Interaction: elements 1 and 2 have reduced risk (6 instead of 10, interaction coefficient -4) and elements 2 and 3 have increased risk (17 instead of 10, interaction offset +7)

any single element risk
(averaged over the other 2 elements)
Details of all classes vs sum

Risk all categories e1y e2y: -4, e2y e3y: +7

Risk over sum variable

- Interaction
- No interaction
Pieces of a pattern score

- What data elements (foods, food groups, nutrients, behaviors, preparation practices)
- What metric
  - What is 1 unit in each element?
- What grouping or combination strategy
  - Grouping similar people (cluster analysis)
    - Not nuanced
  - Combining by taking a function of the data elements
    - Addition most common
    - Metrics usually standardized
    - Weighted somehow
- Data driven
  - PCA standardizes all variables, then weight according to their correlatedness among data elements (risk not included!)
  - Cluster analysis (maximizes differences from centroids of groups)
- A priori
  - Usually specifies data elements in clinically/nutritionally sensible way and adds with “sensible” weighting
Deconstruction of a pattern score

• A Priori
  • This makes sense in terms of characterizing what is known

• Data Driven
  • No intrinsic reason that a pattern should be related to actual risk
  • Patterns that explain more variability in diet in the sample analyzed are not necessarily better at risk prediction
  • Patterns that contain good combinations of food, properly weighted are predictive of risk

• Complex to less complex
  • Which elements or groups of elements relate to risk, independent of the rest of the elements?
  • How to best express each element
  • How to find missing elements (particularly level of processing)?
  • One element at a time is likely not very predictive
  • Sums allow many small contributors and hidden synergies to express themselves
    • Cancellation of effects across elements can happen
  • Remove one at a time
  • Remove and isolate groups of elements
  • Change signs or weighting
  • Very little systematic work has been done along these lines
<table>
<thead>
<tr>
<th>Food Category</th>
<th>A Priori Diet Quality Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vegetables</td>
<td>(+) Beans and legumes, green vegetables, other vegetables, tomato</td>
</tr>
<tr>
<td></td>
<td>(-) Fried potatoes</td>
</tr>
<tr>
<td></td>
<td>(0) Potatoes</td>
</tr>
<tr>
<td>Fruit</td>
<td>(+) Fruit</td>
</tr>
<tr>
<td></td>
<td>(0) Fruit juice</td>
</tr>
<tr>
<td>Nuts and soy protein</td>
<td>(+) Seeds and nuts, soy products</td>
</tr>
<tr>
<td></td>
<td>(-) Red meat, liver, processed meat</td>
</tr>
<tr>
<td></td>
<td>(0) Eggs</td>
</tr>
<tr>
<td>Meat</td>
<td>(+) Fish, poultry</td>
</tr>
<tr>
<td></td>
<td>(-) Salty snacks, sweets</td>
</tr>
<tr>
<td></td>
<td>(0) Margarine</td>
</tr>
<tr>
<td>Grains</td>
<td>(+) Whole grain</td>
</tr>
<tr>
<td></td>
<td>(-) Fried foods, butter, whole fat dairy</td>
</tr>
<tr>
<td></td>
<td>(0) Refined grain</td>
</tr>
<tr>
<td>Trans fat/salt</td>
<td>(+) Low fat dairy, oil</td>
</tr>
<tr>
<td></td>
<td>(0) Chocolate, eggs</td>
</tr>
<tr>
<td>Fats</td>
<td>(+) Beer, liquor, wine</td>
</tr>
<tr>
<td>Alcohol</td>
<td>(0) Diet soft drink</td>
</tr>
<tr>
<td>Other beverages</td>
<td>(+) Coffee, tea</td>
</tr>
<tr>
<td></td>
<td>(-) Soft drink</td>
</tr>
</tbody>
</table>