American Journal of Epidemiology
Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health 2009.

Vol. 171, No. 4
DOI: 10.1093/aje/kwp393
Advance Access publication:
December 21, 2009

Practice of Epidemiology

Comparing 3 Dietary Pattern Methods—Cluster Analysis, Factor Analysis, and Index Analysis—With Colorectal Cancer Risk

The NIH–AARP Diet and Health Study

Jill Reedy*, Elisabet Wirfält, Andrew Flood, Panagiota N. Mitrou, Susan M. Krebs-Smith, Victor Kipnis, Douglas Midthune, Michael Leitzmann, Albert Hollenbeck, Arthur Schatzkin, and Amy F. Subar

* Correspondence to Dr. Jill Reedy, Division of Cancer Control and Population Sciences, Applied Research Program, Risk Factor Monitoring and Methods Branch, National Cancer Institute, 6130 Executive Boulevard, EPN 4005, MSC 7344, Bethesda, MD 20892-7344 (e-mail: reedyj@mail.nih.gov).

Initially submitted June 15, 2009; accepted for publication November 4, 2009.

The authors compared dietary pattern methods—cluster analysis, factor analysis, and index analysis—with colorectal cancer risk in the National Institutes of Health (NIH)–AARP Diet and Health Study (n = 492,306). Data from a 124-item food frequency questionnaire (1995–1996) were used to identify 4 clusters for men (3 clusters for women), 3 factors, and 4 indexes. Comparisons were made with adjusted relative risks and 95% confidence intervals, distributions of individuals in clusters by quintile of factor and index scores, and health behavior characteristics. During 5 years of follow-up through 2000, 3,110 colorectal cancer cases were ascertained. In men, the vegetables and fruits cluster, the fruits and vegetables factor, the fat-reduced/diet foods factor, and all indexes were associated with reduced risk; the meat and potatoes factor was associated with increased risk. In women, reduced risk was found with the Healthy Eating Index-2005 and increased risk with the meat and potatoes factor. For men, beneficial health characteristics were seen with all fruit/vegetable patterns, diet foods patterns, and indexes, while poorer health characteristics were found with meat patterns. For women, findings were similar except that poorer health characteristics were seen with diet foods patterns. Similarities were found across methods, suggesting basic qualities of healthy diets. Nonetheless, findings vary because each method answers a different question.

colorectal neoplasms; food habits; risk

Abbreviations: AHEI, Alternate Healthy Eating Index; HEI-2005, Healthy Eating Index-2005; MDS, Mediterranean Diet Score; NIH, National Institutes of Health.

Cluster analysis, factor analysis, and index analysis use distinct statistical approaches to approximate dietary patterns. Experts have recommended comparing these methods in relation to a disease outcome to better understand the different patterns, but such investigation has been limited (1–4).

To address this gap, we designed a comparison of the 3 most common dietary pattern methods—cluster analysis, factor analysis, and index analysis—with colorectal cancer within the same cohort, the National Institutes of Health (NIH)–AARP Diet and Health Study (n = 492,306). To our knowledge, such a comparison has not been done before. We planned 4 analyses: 1) cluster analysis (5), 2) factor analysis (6), and 3) index analysis (7) to separately investigate colorectal cancer risk and 4) a comparative analysis of the 3 approaches. Our goal here is to compare the findings from the earlier work side-by-side, illustrate if/how individuals are categorized into different patterns, and examine the health behavior characteristics within the pattern groups.

Table 1 highlights the different key questions, distinguishing features, and strategies used to assess the risk associated with each method. Cluster analysis and factor

Table 1. Three Methods of Determining Dietary Patterns (Cluster Analysis, Factor Analysis, and Index Analysis): Key Questions, Distinguishing Features, and Comparison Strategy Associated With Each Method, NIH–AARP Diet and Health Study, 1995–2000

Dietary Pattern Approach	Key Questions	Distinguishing Features	Comparison Strategy Used to Assess Colorectal Cancer Risk
Cluster analysis	Which people cluster together with regard to dietary intake within the defined population?	Large clusters represent behaviors shared by many, and small clusters represent very specific behaviors shared by a few individuals (outliers). Food choices common among most individuals contribute little to cluster formation.	Largest cluster used as the reference category
	What typifies each cluster's diet?	Clusters are categories where the variation of individual foods is not considered after classification.	
	How does each cluster relate to cancer outcome?	No gradients are formed.	
Factor analysis	What foods are correlated, suggesting underlying factors within diets of the population?	Factors are scales based on the correlations among many foods for which individuals have low, medium, or high scores.	Highest quintile compared with the lowest quintile of factor scores for each of the factors
	How do individuals score on those factors?	A gradient is formed.	
	How does each factor relate to cancer outcome?		
Index analysis	How do individuals score on each index?	Assigns scores for the total diet based on food guidance recommendations	Highest quintile compared with the lowest quintile of index scores for each of the indexes
	How do indexes relate to cancer outcome?	Ranks individuals with low scores (diets that are less favorable in many respects) versus those with high scores (diets that are more favorable in many respects). Individuals with moderate scores (diets that are favorable in some respect(s) and not in others) constitute a mix of many different exposures.	
		A gradient is formed.	

Abbreviation: NIH, National Institutes of Health.

analysis are broadly categorized as "data-driven" approaches that derive a posteriori patterns, while index analysis is an "investigator driven" approach that creates patterns based on a priori decisions. Patterns identified via cluster analysis and factor analysis are influenced by the given population and an investigator-driven food-grouping strategy, while index analysis patterns are influenced by an investigator-driven schema and food-grouping strategy. Although cluster analysis and factor analysis are both empirical methods, they are distinct in their approaches. Cluster analysis finds *people* who share similar frequency patterns for consumption of foods, whereas factor analysis finds *foods* that are correlated and then scores people based on the degree to which their diets show the same pattern of variation. Index-based analysis, however, imposes an *external structure* and assesses the degree to which individuals fit within it. Additional details about each method are included below.

In *cluster analysis*, the *k*-means cluster analysis methodology identifies aggregates of individuals in multidimensional space, where each food variable constitutes an axis (using the squared Euclidian distances between observations to determine cluster position). Each individual is positioned in space on the basis of intake of numerous foods. Food choices common to all contribute less to cluster formation than those choices made by some and not by others. The interindividual variation of food variables within a defined cluster is no longer considered once the cluster position is established, despite the fact that the variation of intake within some clusters is greater than others. Thus, no differentiation is made among individuals within the same cluster who have somewhat dissimilar diets; that is, there is no gradient.

Factor analysis (or principal component analysis) examines the correlation matrix of food variables and searches for underlying traits (or factors) that explain most of the variation in the data. Thus, a large number of food variables are reduced to a smaller set of variables that capture the major dietary traits in the population. Commonly, the emerging factors are adjusted by using an "orthogonal rotation" so that the final factors are uncorrelated. For each factor, scores are obtained that define the position of each individual along a gradient.

Index-based analysis uses a numerical scoring system defined on the basis of a priori knowledge. Indexes generate scores for different sets of dietary components based on the researcher's scoring approach and interpretation of dietary guidance. The individual components of the index are summed to a total score so that all participants are ranked from the minimum to maximum score. Index-based analysis allows for comparability across cohorts as the scoring is not driven by the specific population. Indexes may differ in design, structure, and interpretation of dietary guidance.

MATERIALS AND METHODS

Study participants

We used data from the NIH–AARP Diet and Health Study, a prospective cohort study designed to investigate diet and cancer. AARP members who were aged between 50 and 71 years and residents of 6 states (California, Florida, Louisiana, New Jersey, North Carolina, Pennsylvania) or 2 metropolitan areas (Atlanta, Georgia; Detroit, Michigan) were contacted in 1995–1996 to participate in the NIH–AARP Diet and Health Study; 18% ($n = 617,119$) returned the questionnaire. After reviewing surveys with satisfactory dietary data ($n = 566,407$), we excluded questionnaires completed by proxy ($n = 15,760$), respondents who reported previous cancer ($n = 52,867$) or end-stage renal disease ($n = 997$), and individuals with energy outliers, as defined by a Box-Cox transformation ($n = 4,401$) (8). Finally, we excluded cluster outliers ($n = 76$) as determined through cluster analysis, and therefore these analyses included 492,306 people (293,576 men and 198,730 women).

Cohort follow-up and identification of cancer cases

Study participants were followed from enrollment in 1995–1996 through December 31, 2000. Vital status was determined by annual linkage of the cohort to the Social Security Administration Death Master File on deaths in the United States, follow-up searches of the National Death Index for matched subjects, cancer registry linkage, questionnaire responses, and responses to other mailings. Incident cases of cancer were identified by probabilistic linkage between the NIH–AARP membership and 8 state cancer registry databases. In a previous analysis to study the validity of this approach, approximately 90% of all cancers were assessed (9). Further details on study design have been described elsewhere (9). The NIH–AARP Diet and Health Study was approved by the Special Studies Institutional Review Board of the National Cancer Institute.

During follow-up, we identified 3,110 incident colorectal cancer cases (2,151 in men and 959 in women). Cases were invasive and defined on the basis of *International Classification of Diseases for Oncology*, Third Edition, codes C180–C189, C199, C209, and C260. If multiple cancers were diagnosed in the same participant, we included the colorectal cancer case only if it was the first malignancy diagnosed during the follow-up period.

Exposure assessment

At baseline, study participants completed a 124-item food frequency questionnaire, an early version of the Diet History Questionnaire, to assess dietary intake over the past year. The Diet History Questionnaire has been calibrated (10, 11), and further validation was done with the AARP food frequency questionnaire and two 24-hour recalls within the NIH–AARP Diet and Health Study (12).

Our methods to identify dietary patterns by use of cluster analysis, factor analysis, and index analysis were the same as those described previously (5–7). Excluding the 76 study participants identified as cluster outliers in the cluster analysis did not change the factor analysis and index analysis findings. To create the clusters and factors, we used 181 food groups based on the 204 food items drawn from the food frequency questionnaire (because line items contain more than one food item, the final number of food items from the 124-item food frequency questionnaire was 204). We energy adjusted the food groups (expressed as grams per day) by dividing the intake of each food group by total energy and multiplying by 1,000 and then standardized these variables to a mean of 0 and standard deviation of 1. To construct index scores, we used the food group and nutrient variables from the food frequency questionnaire.

Wirfält et al. (5) identified 4 large and stable clusters for men (many foods, vegetables and fruits, fatty meats, fat-reduced foods) and 3 large and stable clusters for women (many foods, vegetables and fruits, diet foods/lean meats). For both men and women, smaller clusters were found (<10,000 individuals), but these were characterized by very specific foods and therefore were not included. Flood et al. (6) identified 3 factors for men and 3 similar factors for women: fruits and vegetables, fat-reduced and diet foods, and meat and potatoes. Reedy et al. (7) scored 4 indexes, including the Healthy Eating Index-2005 (HEI-2005) (13), the Alternate Healthy Eating Index (AHEI) (14–16), the alternate Mediterranean Diet Score (alternate MDS, modified for an American diet) (17–19), and the Recommended Food Score (20). The scoring standards are the same for men and women for all indexes except the MDS, which is based on sex-specific median intake.

Statistical analysis

We used SAS, version 8.1, software (SAS Institute, Inc., Cary, North Carolina) for statistical analyses. We defined clusters, factors, and index scores as described previously (5–7), separately for men and women. We examined the adjusted relative risks and 95% confidence intervals for colorectal cancer risk on the basis of previous analyses for cluster analysis (using the largest cluster, many foods, as the reference category); factor analysis (comparing the highest with the lowest quintiles for factor scores on each factor, quintile 5 vs. quintile 1); and index analysis (comparing the highest with the lowest quintiles for each index score, quintile 5 vs. quintile 1). We calculated the percentage of men and women from each cluster in the highest and lowest quintiles of each factor and index score. Finally, we compared health behavior characteristics for men and women in key clusters, the highest quintile for each factor, and the highest quintile for each index. The variables that we compared were as follows: energy intake (kilocalories); protein (all nutrients based on grams or milligrams per 1,000 kcal); total fat; carbohydrate; calcium; dietary fiber; folate; body mass index (18.5–24.9, 25–29, 30–34, 35–39, $\geq$40 kg/m^2); education (less than high school, high school, some college, college graduate); smoking (never smoker, former smoker of $\leq$1 pack per day, former smoker

Table 2. Adjusted Relative Risks[a] and 95% Confidence Intervals for Colorectal Cancer ($n = 492{,}306$) According to Different Approaches to Dietary Pattern Analysis: Cluster Analysis, Factor Analysis, and Index Analysis, NIH–AARP Diet and Health Study (n = 492,306), 1995–2000

	Men (n = 293,576)			Women (n = 198,730)		
	No.	Relative Risk	95% Confidence Interval	No.	Relative Risk	95% Confidence Interval
Cluster (largest cluster in the reference category)[b]						
Many foods	176,127	1.00		87,109	1.00	
Vegetables and fruits	81,318	0.85	0.76, 0.94[c]	64,671	0.90	0.77, 1.06
Fatty meats	22,756	0.94	0.80, 1.10			
Fat-reduced foods	11,273	0.88	0.70, 1.11			
Diet foods/lean meats				32,426	1.04	0.87, 1.24
Factor (highest vs. lowest quintile scores)[d]						
Fruits and vegetables	58,696	0.81	0.71, 0.93[c]	39,735	1.06	0.87, 1.30
Fat-reduced and diet foods	58,710	0.81	0.71, 0.93[c]	39,732	0.87	0.71, 1.07
Meat and potatoes	58,694	1.18	1.02, 1.35[c]	39,733	1.48	1.20, 1.83[c]
Index (highest vs. lowest quintile scores)[e]						
Healthy Eating Index-2005	58,717	0.72	0.62, 0.83[c]	39,749	0.80	0.64, 0.98[c]
Alternate Healthy Eating Index	58,721	0.70	0.61, 0.81[c]	39,752	0.80	0.64, 1.00
Mediterranean Diet Score	75,205	0.72	0.63, 0.83[c]	44,434	0.89	0.72, 1.11
Recommended Food Score	60,007	0.75	0.65, 0.87[c]	38,539	1.01	0.80, 1.28

Abbreviation: NIH, National Institutes of Health.

[a] Adjusted for age, ethnicity, education, body mass index, smoking, physical activity, energy, and menopausal hormone therapy (women only).

[b] Refer also to cluster analysis of Wirfält et al. (5).

[c] Confidence interval does not include “1.”

[d] Refer also to factor analysis of Flood et al. (6).

[e] Refer also to index analysis of Reedy et al. (7).

of >1 pack per day, current smoker of ≤1 pack per day, current smoker of >1 pack per day); and physical activity (≥20 daily minutes reported rarely or never, 1–3 times per month, 1–2 times per week, 3–4 times per week, ≥5 times per week).

RESULTS

Table 2 presents adjusted relative risks and 95% confidence intervals for colorectal cancer for men and women based on previous cluster analysis, factor analysis, and index analysis (5–7). In men, the vegetables and fruits cluster, fruit and vegetables factor, fat-reduced/diet foods factor, and all indexes (HEI-2005, AHEI, MDS, Recommended Food Score) were associated with reduced risk for colorectal cancer; the meat and potatoes factor was associated with increased risk. In women, a significantly reduced risk was found with the HEI-2005, and an increased risk was found only with the meat and potatoes factor.

Table 3 examines the percentage of men and women from each cluster in the highest quintile of each factor and index score. Fifty-seven percent of men in the vegetables and fruits cluster were classified in the highest quintile of the fruits and vegetables factor, and 48% of men in the vegetables and fruits cluster were in the highest quintile of the HEI-2005. Although 86% of men in the fat-reduced foods cluster were in the highest quintile of the fat-reduced/diet foods factor, only 37% were in the highest quintile of the HEI-2005. Again, 86% of men in the fatty meats cluster were in the highest quintile of the meat and potatoes factor, and 5% were in the highest quintile of the HEI-2005.

For women, 48% and 41% of the vegetables and fruits cluster were also in the highest quintiles for the fruits and vegetables factor and the HEI-2005, respectively. Forty-three percent of women in the diet foods/lean meats cluster were classified in the highest quintile of the fat-reduced/diet foods factor, and just 22% of women in the diet foods/lean meats cluster were also in the highest quintile of the HEI-2005.

Table 4 presents the converse relation—specifically, the percentage of men and women from each of the clusters in the *lowest* quintile of each factor and index score. For men, the classification in the lowest quintiles appears to be clearer than those for the highest quintiles, as just 1% of the men in the vegetables and fruits cluster are also in the lowest quintiles for the fruits and vegetables factor and the HEI-2005. This percentage is similarly low (0%–1%) with the

Table 3. Percentage of Men ($n = 293{,}576$) and Women ($n = 198{,}730$) From Each Cluster in the Highest Quintile of Each Factor and Index Score, NIH–AARP Diet and Health Study, 1995–2000

	Cluster						
	Men (%)				Women (%)		
	Many Foods (n = 176,127)	Vegetables and Fruits (n = 81,318)	Fatty Meats (n = 22,756)	Fat-reduced Foods (n = 11,273)	Many Foods (n = 87,109)	Vegetables and Fruits (n = 64,671)	Diet Foods/ Lean Meats (n = 32,426)
Factor[a]							
Fruits and vegetables	3	57	17	20	3	48	13
Fat-reduced/diet foods	10	36	11	86	17	2	43
Meat and potatoes	16	10	86	16	2	30	54
Index[a]							
Healthy Eating Index-2005	8	48	5	37	6	41	22
Alternate Healthy Eating Index	14	34	12	29	12	54	17
Mediterranean Diet Score	15	52	12	38	10	44	18
Recommended Food Score	11	40	25	27	8	32	24

Abbreviation: NIH, National Institutes of Health.

[a] Highest quintile only.

fat-reduced foods cluster and lowest quintile of the fat-reduced/diet foods factor, as well as with the fatty meats cluster and the lowest quintile of the meat and potatoes factor. This pattern is consistent for women as well. A small percentage (3% and 2%) of those women in the vegetables and fruits cluster is also in the lowest quintile of the fruits and vegetable factor and the HEI-2005, respectively.

Tables 5 and 6 present the demographic and nutrient intake characteristics of men and women by key clusters, factors, and index scores (index scores are represented in the table by only one index, the HEI-2005, but the characteristics were consistent for all indexes; data not shown). The men in the vegetable and fruits cluster, the fat-reduced foods cluster, and the highest quintiles of the fruits and vegetable factor, fat-reduced/diet foods factor, and HEI-2005 have a similar—and generally favorable—nutrient and health behavior profile. In contrast, the men in the fatty meats cluster, the top quintile of the meat and potatoes factor, and the lowest quintile of the HEI-2005 are systematically different from the participants in the other groups. These men had less favorable health profiles; they were less likely to be nonsmokers and to be overweight (reflected in their greater total energy intake and less physical activity). They also reported fewer years of education and had diets that indicated greater consumption of total fat and less calcium, fiber, and folate than those in the other groups.

The women in the vegetable and fruits cluster, as well as the women in the highest quintiles of the fruits and vegetable factor and HEI-2005, also shared favorable health behavior profiles. However, the women in the 2 so-called diet food pattern groups (diet foods/lean meats cluster and the highest quintile of the fat-reduced/diet foods factor) have a generally

Table 4. Percentage of Men ($n = 293{,}576$) and Women ($n = 198{,}730$) From Each Cluster in the Lowest Quintile of Each Factor and Index Score, NIH–AARP Diet and Health Study, 1995–2000

	Cluster						
	Men (%)				Women (%)		
	Many Foods (n = 176,127)	Vegetables and Fruits (n = 81,318)	Fatty Meats (n = 22,756)	Fat-reduced Foods (n = 11,273)	Many Foods (n = 87,109)	Vegetables and Fruits (n = 64,671)	Diet Foods/ Lean Meats (n = 32,426)
Factor[a]							
Fruits and vegetables	30	1	10	18	33	3	18
Fat-reduced/diet foods	25	8	33	1	15	40	2
Meat and potatoes	15	10	0	26	34	9	0
Index[a]							
Healthy Eating Index-2005	28	1	29	5	35	2	8
Alternate Healthy Eating Index	25	8	27	10	29	7	17
Mediterranean Diet Score	24	2	25	7	26	3	13
Recommended Food Score	30	4	18	11	32	5	11

Abbreviation: NIH, National Institutes of Health.

[a] Highest quintile only.

Table 5. Baseline Characteristics Among Men in Key Clusters, Factors, and Indexes, NIH–AARP Diet and Health Study, 1995–2000

	Vegetables and Fruits (Cluster)	Fruits and Vegetables (Factor)[a]	Healthy Eating Index-2005 (Index)[a]	Fat-reduced Foods (Cluster)	Fat-reduced/Diet Foods (Factor)[a]	Fatty Meat (Cluster)	Meat and Potatoes (Factor)[a]	Healthy Eating Index-2005 (Index)[b]
No.	81,318	58,696	58,717	11,273	58,710	22,756	58,694	58,717
Energy, kcal	1,706	1,724	1,788	1,800	1,744	2,253	2,122	2,368
Body mass index, kg/m^2	26.6	26.7	26.9	27.0	27.1	27.8	27.8	27.3
Physical activity $\geq$5/week, %	29.2	28.9	28.2	27.2	25.3	17.3	17.5	15.6
Never smoker, %	35.0	35.2	36.7	28.1	31.7	28.5	29.8	20.3
College graduate, %	55.9	55.2	52.6	48.5	47.6	36.0	36.7	33.5
Protein, g/1,000 kcal	40.1	39.6	41.8	42.3	42.8	41.7	41.3	33.3
Total fat, g/1,000 kcal	28.4	30.4	30.0	29.7	30.3	40.8	39.6	33.5
Carbohydrate, g/1,000 kcal	143.3	140.0	143.1	136.5	137.1	112.9	117.8	111.9
Calcium, mg/1,000 kcal	454.0	423.9	502.3	463.0	473.6	351.5	363.6	331.4
Dietary fiber, g/1,000 kcal	13.8	14.2	13.8	12.1	12.0	9.0	9.7	7.1
Folate, μg/1,000 kcal	276.9	277.9	278.3	247.3	249.8	200.2	207.0	166.5
Alcohol, g/1,000 kcal	5.5	5.2	2.6	5.8	4.3	4.3	3.5	16.3

Abbreviation: NIH, National Institutes of Health.

[a] Highest quintile only.

[b] Lowest quintile only.

poor health behavior profile, similar to the women in the highest quintile of the meat and potatoes factor.

DISCUSSION

Rather than suggesting that one approach is superior, our results demonstrate that findings can vary depending on the methods used to elucidate dietary patterns, because each method is designed to answer a different question. Cluster analysis and factor analysis ask what accounts for the variation in intakes and how well those variances relate to risk, whereas index analysis asks whether variation from a predefined diet relates to risk. Nonetheless, similarities were seen across methods, suggesting some basic qualities of healthy diets.

Overall, we can summarize the evidence regarding dietary patterns and risk as follows: For men, cluster analysis, factor analysis, and index analysis come together to help us understand patterns that can *reduce* risk—diets rich in fruits and vegetables and diets including lower fat foods—and the evidence for patterns (based on factor analysis and index analysis) that can *increase* risk—diets defined by a meat and potatoes pattern. For women, the results were less consistent, as only one factor revealed increased risk (meat and potatoes factor) and one index pattern showed decreased risk (HEI-2005).

The different findings between men and women could be due to the greater heterogeneity in women's diets (9), biologic differences, increased measurement error among women (21), differences in how men and women completed the food frequency questionnaire, or other reasons. Additionally, though, we found differences in the health behavior characteristics of men and women in similar-looking patterns that might help to explain why these patterns produced different results. The women in the diet food pattern groups (defined as diet foods/lean meats cluster and fat-reduced/diet foods factor) look like "dieters," women who are in poorer health/overweight, trying to change their behaviors, or at least report a "good" diet. On the other hand, the men in the diet food pattern groups (defined as fat-reduced foods cluster and fat-reduced/diet foods factor) look "health-conscious." Thus, the women and men in the diet food pattern groups had dissimilar health behavior characteristics. The "women dieters" had profiles most similar with those individuals in the meat and potatoes factor, and the "health-conscious men" looked more like those in the fruits and vegetables pattern groups and all indexes.

Among men, however, we also saw differences in cancer risk. We did not see the same association with colorectal cancer for men in the fatty meats cluster and the meat and potatoes factor. This may be due to differences in group size, but it also reflects that, even when the health characteristics and nutrient profiles are similar, *the actual foods and/or people that make up these patterns differ because they are defined by using different statistical procedures.*

In cluster analysis and factor analysis, labels such as "fruits and vegetable factor" are commonly attached to factors and clusters that emerge analytically. However, similar labels can represent meaningfully different patterns. In

Table 6. Baseline Characteristics Among Women in Key Clusters, Factors, and Indexes, NIH–AARP Diet and Health Study, 1995–2000

	Vegetables and Fruits (Cluster)	Fruits and Vegetables (Factor)[a]	Healthy Eating Index-2005 (Index)[a]	Diet Foods/Lean Meats (Cluster)	Fat-reduced/Diet Foods (Factor)[a]	Meat and Potatoes (Factor)[a]	Healthy Eating Index-2005 (Index)[b]
No.	64,671	39,735	39,753	32,426	39,732	39,733	39,759
Energy, kcal	1,387	1,429	1,480	1,583	1,432	1,789	1,723
Body mass index, kg/m^2	25.9	26.1	26.5	27.7	27.0	28.1	27.1
Physical activity $\geq$5/week, %	23.6	24.2	21.2	14.1	18.7	11.6	10.5
Never smoker, %	45.7	47.7	49.8	47.4	44.7	46.0	36.1
College graduate, %	39.4	36.5	35.8	26.5	32.3	21.0	21.8
Protein, g/1,000 kcal	40.7	40.0	42.7	42.7	44.8	40.5	33.8
Total fat, g/1,000 kcal	27.4	29.7	29.3	34.6	30.3	39.9	37.8
Carbohydrate, g/1,000 kcal	149.8	146.6	146.1	130.0	138.5	120.6	122.9
Calcium, mg/1,000 kcal	519.6	472.3	561.1	442.1	499.0	372.4	374.1
Dietary fiber, g/1,000 kcal	15.0	15.9	14.3	11.4	12.6	10.0	7.8
Folate, μg/1,000 kcal	290.4	298.5	281.5	234.2	254.5	212.2	182.8
Alcohol, g/1,000 kcal	3.4	2.8	1.6	2.7	2.6	2.2	7.4

Abbreviation: NIH, National Institutes of Health.

[a] Highest quintile only.

[b] Lowest quintile only.

the analyses presented by Wirfält et al. (5) and Flood et al. (6), the clusters and factors were derived separately for men and women and, despite the similar names, they are not defined by exactly the same foods, nor are they the same as clusters and factors similarly named in other studies. These methods are data driven and dependent on the intake within the population from which they are drawn. Labels help to clarify the discussion of the findings; indeed, we have used labels here regarding "fruits and vegetables," "diet food," and "meat" pattern groups. Using labels makes for easier presentation to an audience, but it makes less clear the fact that clusters or factors with similar or identical names may be quite different.

Other comparative work with cluster analysis and factor analysis has focused on the stability and reproducibility of clusters and factors and, to a lesser extent, on the general picture provided by the methods. Research that has compared different methods with a biomarker or health outcome includes comparisons of cluster analysis and factor analysis with plasma lipid biomarkers (22), factor analysis and reduced rank regression with biomarkers of subclinical atherosclerosis (23), and factor analysis and index analysis with plasma sex hormone concentrations (24), mortality (25), and hypertension (26). In related analyses of cluster analysis and factor analysis, Newby et al. (2) also found that some associations were significant for men and not for women (white bread cluster and lower high density lipoprotein cholesterol), some were significant when using factor analysis but not cluster analysis (sweets factor and lower high density lipoprotein cholesterol), and some were similar with cluster analysis and factor analysis (healthy pattern and lower plasma triacylglycerols). Nettleton et al. (4) found that prior information about inflammation included with reduced rank regression strengthened the ability to detect an association (no association was found for factor analysis). Although differences were found in the foods in the patterns, this did not entirely account for the lack of association when using factor analysis. This reinforces the unique information provided by different pattern analysis methods (4).

There have been 3 analyses that have compared index analysis and factor analysis by using different outcomes: Fung et al. (22) found an association with index analysis (higher AHEI score and lower levels of free estradiol) but not factor analysis for plasma sex hormone concentrations; Osler et al. (23) found an association with factor analysis (prudent pattern and all-cause and cardiovascular morality) but not index analysis; and Schulze et al. (24) found no associations with either index analysis or factor analysis for hypertension (although the third of 4 quintiles measured with a Dietary Approaches to Stop Hypertension (DASH) Index was associated with a reduced risk). Although Fung et al. (22) and Schulze et al. (24) postulate that index analysis may provide a stronger ability to find more significant effects on disease risk than factor analysis, this is likely because of the inclusion of relevant, evidence-based components within a given index (22, 24). For example, Fung et al. (22) suggest that the reason they found a relation with the AHEI and not with factor analysis may be due to the emphasis on soy in the index used.

However, although an index may include a critical component, it may suffer from dilution if some dietary components are not relevant (24).

Comparisons across the methods are somewhat limited here by our decisions to define our initial food variables. Index analysis used aggregated food groups as used in food-based recommendations. However, both cluster analysis and factor analysis used single foods or minimally aggregated food groups.

Regardless of the food grouping strategy selected, we recommend using energy-adjusted variables—as we did—to account for the energy compositions of the diet rather than using variables that are derived from absolute dietary intakes. This adjustment is suggested because energy needs are determined by body size, age, physical activity, and other factors and also because diet quality is of greater interest rather than absolute intakes. Energy adjustment may also help to reduce measurement error (21), although future work is needed in this area.

The goal with dietary pattern analyses is to examine the multiple dimensions of the diet simultaneously relative to a given outcome. Thus, we consider the best way to operationalize and model the multidimensionality of the total diet. Although cluster analysis, factor analysis, and index analysis are useful and answer different questions, perhaps we should not limit ourselves to these common approaches (25). Other methods hold promise for new ways to explain the complexity of dietary data and would allow us to ask other questions: What combination of foods explains the variation in a set of intermediate health markers (reduced rank regression) (26)? What combination of foods minimizes cancer risk (neural networks) (27)? What features of the diet are most strongly associated with a reduced risk of cancer (classification and regression trees) (28)?

Dietary pattern analyses play a unique role in assessing the relations between diet and disease. Although most research with dietary patterns has been shown to be more strongly related to risk of disease than individual parts of the diet (29), the World Cancer Research Fund Panel stated that there was insufficient evidence to make judgments regarding dietary patterns and cancer risk (30). Our results are consistent with their summaries for specific foods and dietary components and reinforce the Panel's recommendation that additional research be done investigating dietary patterns.

ACKNOWLEDGMENTS

Author affiliations: National Cancer Institute, Bethesda, Maryland (Jill Reedy, Susan M. Krebs-Smith, Victor Kipnis, Douglas Midthune, Arthur Schatzkin, Amy F. Subar); University Hospital Regensburg, Regensburg, Germany (Michael Leitzmann); Lund University, Malmo, Sweden (Elisabet Wirfält); University of Minnesota, Minneapolis, Minnesota (Andrew Flood); World Cancer Research Fund, London, United Kingdom (Panagiota N. Mitrou); University of Cambridge, Cambridge, United Kingdom (Panagiota N. Mitrou); and AARP, Washington, DC (Albert Hollenbeck).

This research was supported by the Intramural Research Program of the National Cancer Institute, National Institutes of Health.

The authors gratefully acknowledge the contributions of Lisa Kahle, Leslie Carroll, and David Campbell at Information Management Services and Tawanda Roy at the Nutritional Epidemiology Branch for research assistance. Cancer incidence data from the Atlanta metropolitan area were collected by the Georgia Center for Cancer Statistics, Department of Epidemiology, Rollins School of Public Health, Emory University. Cancer incidence data from California were collected by the Cancer Surveillance Section, California Department of Health Services. Cancer incidence data from the Detroit metropolitan area were collected by the Michigan Cancer Surveillance Program, Community Health Administration, state of Michigan. The Florida cancer incidence data used in this report were collected by the Florida Cancer Data System under contract to the Department of Health. Cancer incidence data from Louisiana were collected by the Louisiana Tumor Registry, Louisiana State University Medical Center, New Orleans. Cancer incidence data from New Jersey were collected by the New Jersey State Cancer Registry, Cancer Epidemiology Services, New Jersey State Department of Health and Senior Services. Cancer incidence data from North Carolina were collected by the North Carolina Central Cancer Registry. Cancer incidence data from Pennsylvania were supplied by the Division of Health Statistics and Research, Pennsylvania Department of Health, Harrisburg, Pennsylvania.

The views expressed herein are solely those of the authors and do not necessarily reflect those of the contractor (Florida Cancer Data System) or the Florida Department of Health. The Pennsylvania Department of Health specifically disclaims responsibility for any analyses, interpretations, or conclusions.

Conflict of interest: none declared.

REFERENCES

1. Kant AK. Dietary patterns and health outcomes. *J Am Diet Assoc.* 2004;104(4):615–635.
2. Newby PK, Muller D, Tucker KL. Associations of empirically derived eating patterns with plasma lipid biomarkers: a comparison of factor and cluster analysis methods. *Am J Clin Nutr.* 2004;80(3):759–767.
3. Fung TT, Hu FB, Barbieri RL, et al. Dietary patterns, the Alternate Healthy Eating Index and plasma sex hormone concentrations in postmenopausal women. *Int J Cancer.* 2007;121(4):803–809.
4. Nettleton JA, Steffen LM, Schulze MB, et al. Associations between markers of subclinical atherosclerosis and dietary patterns derived by principal components analysis and reduced rank regression in the Multi-Ethnic Study of Atherosclerosis (MESA). *Am J Clin Nutr.* 2007;85(6):1615–1625.
5. Wirfält E, Midthune D, Reedy J, et al. Associations between food patterns defined by cluster analysis and colorectal cancer incidence in the NIH–AARP Diet and Health Study. *Eur J Clin Nutr.* 2009;63(6):707–717.
6. Flood A, Rastogi T, Wirfält E, et al. Dietary patterns as identified by factor analysis and colorectal cancer among middle-aged Americans. *Am J Clin Nutr.* 2008;88(1):176–184.

7. Reedy J, Mitrou PN, Krebs-Smith SM, et al. Index-based dietary patterns and risk of colorectal cancer: the NIH–AARP Diet and Health Study. *Am J Epidemiol*. 2008;168(1):38–48.
8. Adams KF, Schatzkin A, Harris TB, et al. Overweight, obesity, and mortality in a large prospective cohort of persons 50 to 71 years old. *N Engl J Med*. 2006;355(8):763–778.
9. Schatzkin A, Subar AF, Thompson FE, et al. Design and serendipity in establishing a large cohort with wide dietary intake distributions: the National Institutes of Health–American Association of Retired Persons Diet and Health Study. *Am J Epidemiol*. 2001;154(12):1119–1125.
10. Thompson FE, Subar AF, Brown CC, et al. Cognitive research enhances accuracy of food frequency questionnaire reports: results of an experimental validation study. *J Am Diet Assoc*. 2002;102(2):212–225.
11. Thompson FE, Midthune D, Subar AF, et al. Development and evaluation of a short instrument to estimate usual dietary intake of percentage energy from fat. *J Am Diet Assoc*. 2007; 107(5):760–767.
12. Thompson FE, Kipnis V, Midthune D, et al. Performance of a food-frequency questionnaire in the US NIH–AARP (National Institutes of Health–American Association of Retired Persons) Diet and Health Study. *Public Health Nutr*. 2008; 11(2):183–195.
13. Guenther PM, Krebs-Smith SM, Reedy J, et al. Healthy Eating Index-2005. Fact sheet no. 1. Alexandria, VA: Center for Nutrition Policy and Promotion, US Department of Agriculture, 2008. (http://www.cnpp.usda.gov/Publications/HEI/healthyeatingindex2005factsheet.pdf).
14. McCullough ML, Feskanich D, Stampfer MJ, et al. Diet quality and major chronic disease risk in men and women: moving toward improved dietary guidance. *Am J Clin Nutr*. 2002;76(6):1261–1271.
15. McCullough ML, Feskanich D, Rimm EB, et al. Adherence to the Dietary Guidelines for Americans and risk of major chronic disease in men. *Am J Clin Nutr*. 2000;72(5):1223–1231.
16. McCullough ML, Feskanich D, Stampfer M, et al. Adherence to the Dietary Guidelines for Americans and risk of major chronic disease in women. *Am J Clin Nutr*. 2000;72(5):1214–1222.
17. Trichopoulou A, Kouris-Blazos A, Wahlqvist ML, et al. Diet and overall survival in elderly people. *BMJ*. 1995;311(7018): 1457–1460.
18. Trichopoulou A, Orfanos P, Norat T, et al. Modified Mediterranean diet and survival: EPIC-elderly prospective cohort study [electronic article]. *BMJ*. 2005;330(7498):991.
19. Mitrou PN, Kipnis V, Thiébaut AC, et al. Mediterranean dietary pattern and prediction of all-cause mortality in a US population: results from the NIH–AARP Diet and Health Study. *Arch Intern Med*. 2007;167(22):2461–2468.
20. Kant AK, Schatzkin A, Graubard BI, et al. A prospective study of diet quality and mortality in women. *JAMA*. 2000;283(16): 2109–2115.
21. Kipnis V, Subar AF, Midthune D, et al. Structure of dietary measurement error: results of the OPEN Biomarker Study. *Am J Epidemiol*. 2003;158(1):14–21; discussion 22–26.
22. Fung TT, Willett WC, Stampfer MJ, et al. Dietary patterns and the risk of coronary heart disease in women. *Arch Intern Med*. 2001;161(15):1857–1862.
23. Osler M, Heitmann BL, Gerdes LU, et al. Dietary patterns and mortality in Danish men and women: a prospective observational study. *Br J Nutr*. 2001;85(2):219–225.
24. Schulze MB, Hoffmann K, Kroke A, et al. Risk of hypertension among women in the EPIC–Potsdam Study: comparison of relative risk estimates for exploratory and hypothesis-oriented dietary patterns. *Am J Epidemiol*. 2003;158(4): 365–373.
25. Moeller SM, Reedy J, Millen AE, et al. Dietary patterns: challenges and opportunities in dietary patterns research: an Experimental Biology workshop, April 1, 2006. *J Am Diet Assoc*. 2007;107(7):1233–1239.
26. Hoffmann K, Schulze MB, Schienkiewitz A, et al. Application of a new statistical method to derive dietary patterns in nutritional epidemiology. *Am J Epidemiol*. 2004;159(10):935–944.
27. Rothney MP, Neumann M, Béziat A, et al. An artificial neural network model of energy expenditure using nonintegrated acceleration signals. *J Appl Physiol*. 2007;103(4):1419–1427.
28. Camp NJ, Slattery ML. Classification tree analysis: a statistical tool to investigate risk factor interactions with an example for colon cancer (United States). *Cancer Causes Control*. 2002; 13(9):813–823.
29. Kant AK. Indexes of overall diet quality: a review. *J Am Diet Assoc*. 1996;96(8):785–791.
30. World Cancer Research Fund/American Institute for Cancer Research. *Food, Nutrition, Physical Activity, and the Prevention of Cancer: A Global Perspective*. Washington, DC: American Institute for Cancer Research; 2007.