



Practice of Epidemiology

Comparison of 3 Methods for Identifying Dietary Patterns Associated With Risk of Disease

Julia R. DiBello, Peter Kraft, Stephen T. McGarvey, Robert Goldberg, Hannia Campos, and Ana Baylin

Initially submitted April 3, 2008; accepted for publication July 31, 2008.

Reduced rank regression and partial least-squares regression (PLS) are proposed alternatives to principal component analysis (PCA). Using all 3 methods, the authors derived dietary patterns in Costa Rican data collected on 3,574 cases and controls in 1994–2004 and related the resulting patterns to risk of first incident myocardial infarction. Four dietary patterns associated with myocardial infarction were identified. Factor 1, characterized by high intakes of lean chicken, vegetables, fruit, and polyunsaturated oil, was generated by all 3 dietary pattern methods and was associated with a significantly decreased adjusted risk of myocardial infarction (28%–46%, depending on the method used). PCA and PLS also each yielded a pattern associated with a significantly decreased risk of myocardial infarction (31% and 23%, respectively); this pattern was characterized by moderate intake of alcohol and polyunsaturated oil and low intake of high-fat dairy products. The fourth factor derived from PCA was significantly associated with a 38% increased risk of myocardial infarction and was characterized by high intakes of coffee and palm oil. Contrary to previous studies, the authors found PCA and PLS to produce more patterns associated with cardiovascular disease than reduced rank regression. The most effective method for deriving dietary patterns related to disease may vary depending on the study goals.

cardiovascular diseases; diet; myocardial infarction; principal component analysis

Abbreviations: PCA, principal component analysis; PLS, partial least-squares regression; PRESS, predicted residual sum of squares; RRR, reduced rank regression.

Principal component analysis (PCA), one of the major methods for deriving dietary patterns, combines correlated variables into weighted summations that “account for” much of the information contained in the original variables. PCA is usually based on the correlation matrix of foods included in a food frequency questionnaire, resulting in a set of uncorrelated factors (dietary patterns) characterized by the different foods (1, 2).

This method has been used in previous studies relating dietary patterns to the risk of cardiovascular disease in a wide variety of populations and cultures (3–17). One limitation of this approach, when attempting to relate dietary patterns to the occurrence of disease, is that patterns explaining large percentages of variation in food intake do not necessarily explain an equivalent amount of variation

in nutrient intake. The effects of diet on the risk of disease are probably mediated through specific macro- and micro-nutrients, and thus patterns derived using PCA may not be associated with disease risk (18).

Reduced rank regression (RRR) and partial least-squares regression (PLS) are proposed alternatives to PCA for deriving dietary patterns (19). RRR is similar to PCA in that it is also a dimension reduction technique resulting in uncorrelated summary variables representing a larger set of initial factors. However, the goal of RRR is different from that of PCA, as combinations of predictor variables are derived that account for much of the information contained in a set of response variables. Key nutrients or biomarkers function as the response variables, and linear combinations of foods are derived which maximize the explained variance in these

Correspondence to Dr. Ana Baylin, Department of Community Health, Brown University, Box G-S121, 121 South Main Street, 2nd Floor, Providence, RI 02903 (e-mail: ana_baylin@brown.edu).

responses. The PLS method is a compromise between PCA and RRR; the goal of this approach is to explain variability in select nutrients or biomarkers, as well as foods (18).

The results from a limited number of investigations using RRR have shown stronger relations between derived dietary patterns and cardiovascular disease than results from studies using PCA (6, 18). However, RRR has not been used across diverse study populations with varying dietary exposures, and this method needs to be tested further to establish its effectiveness in deriving dietary patterns related to cardiovascular disease. Costa Ricans represent a novel population in which to apply RRR. This population is of particular interest because intake of the predominant dietary items established as risk factors for myocardial infarction in Costa Ricans are changing and the relations between these exposures have not been explored (20, 21).

Our purpose in the present study was to examine the relations among dietary exposures related to myocardial infarction risk in Costa Rica by deriving dietary patterns using all 3 methods: PCA, PLS, and RRR. Additionally, we tested the ability of these methods to derive patterns related to myocardial infarction and attempted to characterize the most appropriate context in which to apply each method.

MATERIALS AND METHODS

Study population

Details on the study population have been published elsewhere (22, 23). In brief, data were collected in Costa Rica (1994–2004) for a population-based case-control study of first, incident nonfatal myocardial infarction ($n = 3,939$). Participation was 98% in cases and 88% among controls. Controls were selected randomly and were matched with cases by age (± 5 years), sex, and current area of residence (county) using information from the Costa Rican Census. All participants gave written informed consent. This study was approved by the Human Subjects Committee of the Harvard School of Public Health and the Ethics Committee of the University of Costa Rica.

Data collection

The data collection procedures have been described elsewhere in detail (23–25). In brief, trained field-workers visited study participants at their homes in the morning and collected information about sociodemographic characteristics, medical history, lifestyle habits, and dietary intake. They also made anthropometric measurements and took subcutaneous adipose tissue samples. Self-reported history of diabetes and hypertension were validated as previously described (23).

Dietary information was obtained using a 135-item food frequency questionnaire that was a modified version of the Willett questionnaire (26) and was developed and validated specifically for use in Costa Rica (27). Subcutaneous adipose tissue biopsies were collected from the upper buttock using a modification of a method described previously (28). Levels of α -linolenic acid and *trans*-fatty acids in adipose tissue were quantified by gas-liquid chromatography as previously described (25).

Response variables

We chose response variables for RRR and PLS that are known to be related to atherosclerosis and myocardial infarction and are of particular interest in this population. These variables included adipose tissue levels of α -linolenic and *trans*-fatty acids and dietary intakes of saturated fat, fiber, and folate.

Adipose tissue levels of α -linolenic and *trans*-fatty acids are likely to more accurately reflect long-term dietary intake than dietary information obtained from a food frequency questionnaire in case-control studies of chronic disease, since adipose tissue has a turnover rate of 1–2 years and is not subject to recall bias (25). Currently there are no accurate biomarkers for long-term intakes of saturated fat, fiber, and folate.

Adipose tissue α -linolenic and *trans*-fatty acids and dietary saturated fat were associated with risk of myocardial infarction in previous investigations in this population (29–32). However, exposure to these fats has changed over time in Costa Rica as intake of palm oil has decreased and has been replaced predominately by partially hydrogenated soybean oil. It is unclear whether the replacement of palm oil with soybean oil results in improved cardiac health due to the combination of variable amounts of atherogenic *trans*-fatty acid and heart-beneficial α -linolenic acid found in partially hydrogenated soybean oil (20, 21). The inclusion of saturated fat, α -linolenic acid, and *trans*-fatty acids as response variables allowed for evaluation of the relations among these variables and the current risk of myocardial infarction associated with their intake.

The mechanisms by which folate and fiber decrease the risk of atherosclerosis have not been fully elucidated, but they have been associated with decreased levels of markers of inflammation and decreased risk of cardiovascular disease in numerous cohort studies (33, 34). Folate and fiber represent important plant-based dietary exposures in this population, and to our knowledge their association with myocardial infarction has not been previously explored.

Statistical methods

To test for differences in mean values or distributions of demographic and lifestyle variables between cases and controls, we used McNemar's test for categorical variables and paired *t* tests for continuous variables. Mean values for the response variables were calculated using regression analysis to adjust for confounding. Dietary responses (folate, fiber, saturated fat) were adjusted for energy intake using the residual method (35). The mean values for all response variables were adjusted for age, sex, and area (matching factors), smoking status, waist:hip ratio, physical activity, and other response variables. Pearson correlation coefficients among response variables for cases and controls were calculated. Data on adipose tissue *trans*-fatty acid were log-transformed to improve normality.

Foods from the food frequency questionnaire were classified into 43 groups based upon nutrient profile or culinary use or were left alone to minimize within-person variation in intakes of individual foods (3). The food groups used in this study are shown in Table 1. Daily consumption frequencies

Table 1. Food Groupings Used in Dietary Analyses for a Case-Control Study of First Incident Myocardial Infarction ($n = 3,574$), Costa Rica, 1994–2004

Food or Food Group	Food Item(s)
Liquor	Whiskey, rum, gin, vodka, “guaro”
Wine and beer	Red and white wine, beer
Coffee	Caffeinated coffee
Tea	Tea (regular, not herbal), water (tap and bottled)
High-energy drinks	Regular drinks with and without caffeine (such as carbonated beverages), instant powdered drinks, fruit drinks (packaged and fresh)
No-energy drinks	Diet-type beverages with and without caffeine (such as carbonated beverages and ginger ale)
Fruit juices	Orange juice, other fruit juices
Cereal, cold breakfast	Refined grain-based cold breakfast cereals (such as corn flakes)
Eggs	Hen eggs
Low-fat dairy products	Skim milk, yogurt, cottage cheese
High-fat dairy products	Whole and 2% dairy-fat milk, sour cream, ice cream, fresh white cheese, cream cheese, white and yellow processed cheese
Grains, refined	White bread, white rice, macaroni, spaghetti, pancakes
Grains, whole	Oatmeal, corn-based products, whole-wheat bread
Grains, whole, other	Bran (oats and wheat), wheat germ
Legumes	Black beans, string beans, peas
Meat, chicken	Chicken with skin and fat
Meat, chicken, lean	Chicken with skin and/or fat removed
Meat, fish	Canned tuna, sardines, white Pacific fish, shrimp, lobster
Meat, organ	Liver (beef, pork, chicken)
Meat, processed	Ham, salami, bologna, mortadela, sausages, hot dogs
Meat, red	Beef, pork, bacon, fried pork rind
Nuts	Peanuts, cashews
Fruits	Pineapples, papayas, bananas, cantaloupes, avocados, apples, pears, mangoes, oranges, watermelons, cashew pulp, tangerines, sweet lemons, grapefruit, lychees, medlars, “jocotes” (<i>Spondias purpurea</i>), apricots, peaches, plums
Tomatoes	Raw tomatoes, tomato-sauce dishes
Vegetables, cruciferous	Broccoli, cauliflower, cabbage
Vegetables, dark yellow	Carrots (raw, cooked), peach palm (<i>Bactris gasipaes</i>), yellow squash (<i>Cucurbita moschata</i>)
Vegetables, green leafy	Cooked spinach, lettuce, mustard greens, other leafy vegetables
Vegetables, other	Cucumbers, plantains, chayotes (<i>Sechium edule</i>), tender squash, corn on the cob, beets, eggplant
Vegetables, root, starchy	Potatoes (baked, boiled, mashed, French fries), yams (<i>Dioscorea alata</i>), sweet potatoes (<i>Ipomoea batatas</i>), cassava (<i>Manihot esculenta</i>), “tiquisque” (a root vegetable; <i>Xanthosoma sagittifolium</i>)
Pizza	Pizza
Soup, thick	Thick and creamy soups
Snacks	Saltine-type crackers, popcorn, chips (potato, cassava, plantain)
Sweets and desserts	Chocolate bars, chocolate confectionery, candy, cookies, brownies, doughnuts, chocolate drinks (hot or cold), homemade and commercial cookies, cakes, and pastry
Oil, other unsaturated	Corn, sunflower, olive, and other oils used for cooking
Oil, soybean	Soybean oil
Oil, palm	Hardened palm oil (manteca) used for cooking
Spreads	Butter, margarine, hydrogenated vegetable oil-butter blend (“lactocrema”)
Dressings	Olive oil (as salad dressing, added to bread or food at the table), lemon and vinegar, and other dressings
Sugar	Sugar added to hot and cold beverages and cereal
Saccharin	Saccharin (pills and drops)
Condiments	Ketchup, salt, mayonnaise, hot chili sauce (added to foods at the table)
Condiments, fresh	Celery, parsley, sweet peppers, onions, cilantro, garlic
Condiments, sweet	Jam, jelly, honey

of food items on the food frequency questionnaire were aggregated into food groups, transformed where appropriate to improve normality, and adjusted for total energy intake by the residual method (35).

The PROC PLS statement in SAS (SAS Institute Inc., Cary, North Carolina) was used to conduct PCA, PLS, and RRR analyses in the combined sample of cases and controls, changing the "method =" option to reflect each desired method. A more detailed description of these methods is given by Hoffman et al. (19). Additionally, we used random-sample cross-validation and van der Voet's test (36) as a guide in selecting the number of factors to retain. The final number of factors selected represents the model with residuals that were insignificantly larger than the model with the minimum predicted residual sum of squares (PRESS) (37). PRESS is equivalent to the residual sum of squares from ordinary regression analyses, except that PRESS is calculated as the difference between predicted values in the training set used to fit the model and the actual values in the portion of the data excluded from model fitting using cross-validation. Using van der Voet's criteria (36), a large number of factors were retained for PCA ($n = 42$) and PLS ($n = 14$), with many of the factors explaining a small percentage of predictor (PCA, PLS) and response (PLS) variation. Thus, we also considered interpretability of the factors, as those explaining little predictor variation were unlikely to represent major dietary patterns in our population. We chose to retain 5 factors for all 3 methods based on interpretability and to increase comparability across methods, since there were 5 response variables used in analyses and the number of factors retained in RRR is constrained by the rank of the covariance matrix of responses (19).

The continuous factor scores produced by PCA, PLS, and RRR analyses for all retained factors were entered into conditional logistic regression models (1 model per method) with myocardial infarction status as the outcome. Factors that were significant in these models were then divided into quintiles and regressed on myocardial infarction status to produce adjusted estimates of odds ratios and 95% confidence intervals. To test for trends across quintiles of dietary patterns, we set the median intake of each quintile to each subject in the same quintile and treated the resulting variable as a continuous variable in regression analyses (8).

In secondary analyses carried out to confirm the derived patterns, the sample was randomly split in 2 and patterns were derived in 1 subsample. Foods were also ungrouped and entered into the analyses to evaluate the effects of food grouping decisions on the resulting dietary patterns. Additionally, foods were entered into analyses unadjusted for energy intake. To examine potential variation in dietary patterns and in the relations of these patterns with risk of myocardial infarction, we stratified our sample by age and sex, excluded participants with previously diagnosed comorbidity (hypertension, hypercholesterolemia, and diabetes), and derived dietary patterns in these subgroups again. There were no major differences in the derived patterns in any of the secondary analyses. Data were analyzed using SAS software, version 9.1 (SAS Institute Inc.).

Table 2. Characteristics of Cases and Controls in a Case-Control Study of First Incident Myocardial Infarction ($n = 3,574$), Costa Rica, 1994–2004^a

Characteristic	Cases ($n = 1,787$)	Controls ($n = 1,787$)
Age, years ^b	58 (11)	58 (11)
Sex, % female ^b	26.5	26.5
Waist:hip ratio	0.97 (0.07)	0.95 (0.07)***
Physical activity, metabolic equivalents/day	1.50 (0.67)	1.55 (0.67)**
Income, \$US/month	506 (394)	581 (427)***
Hypertension, % ^c	38.7	29.5***
Hypercholesterolemia, % ^c	30.6	27.0*
Angina, % ^c	12.8	4.4***
Diabetes mellitus, % ^c	24.3	14.4***
Current smoker, %	39.7	21.3***
Total energy intake, kcal/day	2,700 (944)	2,440 (757)***
Total fat intake, %	32.4 (5.9)	31.7 (5.7)**
Total carbohydrate intake, %	54.3 (7.5)	55.4 (7.3)***
Total protein intake, %	13.2 (2.2)	13.0 (2.1)**

* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$ (all P values are 2-sided).

^a Data presented are mean values (with standard deviations in parentheses) or percentages.

^b Matching factor.

^c Based upon self-reported history of this medical condition.

RESULTS

Study sample characteristics

We excluded participants who were missing the majority of the items on the food frequency questionnaire or had missing data for confounding factors, leaving a total of 3,574 participants. On average, cases had significantly higher waist:hip ratios and lower physical activity levels than controls. Cases also had significantly lower incomes, had a greater frequency of comorbidity, consumed more calories, had a larger percentage of calories from fat and protein, and were more likely to be smokers than controls (Table 2).

Response variables

Mean values for the response variables for cases and controls are presented in Table 3. Adjusted mean levels of folate, fiber, and adipose tissue α -linolenic acid were higher among controls, whereas mean levels of saturated fat and adipose tissue *trans*-fatty acids were higher among cases. Only levels of saturated fat were significantly different across cases and controls, with α -linolenic acid having borderline significance.

In examining the correlations among the response variables, the overall correlation structure was similar for cases and controls (Table 4). Fiber and folate had the highest correlation. Fiber was negatively correlated with saturated fat and positively correlated with α -linolenic acid, with these associations being slightly stronger in controls. Levels

Table 3. Mean Responses for Cases and Controls in a Case-Control Study of First Incident Myocardial Infarction ($n = 3,574$), Costa Rica, 1994–2004^a

Response Variable	Mean		Difference Between Mean Values	95% Confidence Interval
	Controls	Cases		
Dietary intake				
Fiber, g/day				
Model 1 ^b	24.54	23.68	0.87	0.48, 1.26
Model 2 ^c	23.92	23.53	0.39	0.001, 0.79
Model 3 ^d	23.95	23.88	0.07	-0.25, 0.38
Folate, $\mu\text{g/day}$				
Model 1 ^b	434.08	425.85	8.23	0.61, 15.84
Model 2 ^c	423.52	422.19	1.33	-6.41, 9.06
Model 3 ^d	425.72	429.56	-3.84	-10.13, 2.44
Saturated fat, % of energy				
Model 1 ^b	10.37	11.11	-0.74	-0.92, -0.56
Model 2 ^c	10.46	11.13	-0.67	-0.86, -0.49
Model 3 ^d	10.46	10.98	-0.53	-0.70, -0.36
Adipose tissue, % of total fatty acids				
<i>trans</i> -Fatty acids				
Model 1 ^e	3.57	3.50	0.07	-0.005, 0.138
Model 2 ^c	3.50	3.48	0.02	-0.06, 0.09
Model 3 ^d	3.48	3.51	-0.03	-0.10, 0.04
α -Linolenic acid				
Model 1 ^e	0.65	0.62	0.03	0.02, 0.04
Model 2 ^c	0.64	0.62	0.02	0.01, 0.04
Model 3 ^d	0.64	0.63	0.01	-0.002, 0.024

^a Cases and controls were matched by age (± 5 years), sex, and area of residence (county).

^b Results were adjusted for matching factors and total energy intake.

^c Results were adjusted for variables included in model 1 and waist:hip ratio, smoking status, and total metabolic equivalent-hours per day.

^d Results were adjusted for all variables in model 2 and the other response variables.

^e Results were adjusted for matching factors.

of *trans*-fatty acids were negatively correlated with saturated fat and positively associated with α -linolenic acid among both cases and controls.

Comparison of dietary pattern methods

The percentage of variation in response variables explained for each method is displayed in Table 5. As expected, PCA explained the least amount of variation in response variables, followed by PLS and RRR. However, the difference in explained response variability for PLS and RRR was relatively small.

Table 6 shows the model fit statistics for regression of myocardial infarction status on the five retained factors as continuous variables for each method. Interestingly, all 3 methods explained a similar amount of variation in myocardial infarction status in this population in comparison with the null model, having only marginally different -2 log likelihood and Akaike's Information Criterion values.

Dietary patterns

The main factor loadings (correlations of the food groups with the dietary patterns) for patterns significantly associated with the occurrence of myocardial infarction are shown in Figure 1. The first factor derived (vegetable pattern) was similar across all 3 methods and was primarily characterized by high intakes of fruit, dark yellow vegetables, green leafy vegetables, other vegetables, and polyunsaturated oil and a low intake of palm oil. Factors 2 and 3 were not significantly associated with risk of myocardial infarction for any of the methods examined; they consisted of a mix of "healthy" and "unhealthy" foods such as vegetables and high-fat dairy products (including whole milk, ice cream, and cheese). Factor 4 was significantly associated with myocardial infarction using PCA and PLS but was not associated with the occurrence of this endpoint using RRR. Factor 4, based on PCA analyses (palm oil and coffee pattern), was characterized by high intakes of coffee, sugar, and palm oil

Table 4. Pearson Correlation Coefficients for Correlations Between Response Variables in a Case-Control Study of First Incident Myocardial Infarction ($n = 3,574$), Costa Rica, 1994–2004^a

Response Variable	Fiber	Folate	Saturated Fat	<i>trans</i> -Fatty Acids	α -Linolenic Acid
Fiber ^b , g/day	1.0	<i>0.56**</i>	<i>-0.29**</i>	<i>-0.01</i>	<i>0.10**</i>
Folate ^b , μ g/day	<i>0.54**</i>	1.0	<i>-0.15**</i>	<i>0.06*</i>	<i>0.30**</i>
Saturated fat, % of energy	<i>-0.33**</i>	<i>-0.18**</i>	1.0	<i>-0.15**</i>	<i>-0.27**</i>
<i>trans</i> -Fatty acids ^c	<i>-0.03</i>	<i>0.09**</i>	<i>-0.16**</i>	1.0	<i>0.30**</i>
α -Linolenic acid ^c	<i>0.14**</i>	<i>0.28**</i>	<i>-0.31**</i>	<i>0.32**</i>	1.0

* $P < 0.05$; ** $P < 0.001$.^a Correlation coefficients for cases are above the diagonal (italic type), and those for controls are below the diagonal.^b Dietary variables were adjusted for total energy intake.^c Adipose tissue variables (% of total fatty acids).

and, based on PLS (alcohol and polyunsaturated oil pattern), by high intakes of high-energy drinks (including soda and fruit drinks), liquor, wine, beer, and other unsaturated oil and a low intake of high-fat dairy products. Finally, factor 5 (alcohol, legumes, and polyunsaturated oil pattern) was associated with myocardial infarction using patterns derived from PCA only and was characterized by high intakes of liquor, wine, beer, legumes, and other unsaturated oil.

Associations between dietary patterns and myocardial infarction

In examining the adjusted odds ratios across quintiles of the significant dietary patterns resulting from each method, factor 1 (vegetable pattern) was associated with decreased risk of myocardial infarction across increasing quintiles of the dietary pattern regardless of the method used.

Table 5. Explained Variation in Response Variables Using Principal Component Analysis, Partial Least-Squares Regression, and Reduced Rank Regression in a Case-Control Study of First Incident Myocardial Infarction ($n = 3,574$), Costa Rica, 1994–2004

	Explained Variation in Response Variable, %				
	Fiber	Folate	Saturated Fat	<i>trans</i> -Fatty Acids	α -Linolenic Acid
Principal component analysis					
Factor 1	8.1	7.4	1.0	3.4	0.06
Factor 2	30.7	7.8	6.7	0.03	6.3
Factor 3	2.2	0.8	0.02	0.25	1.4
Factor 4	1.6	0.2	12.3	6.6	14.2
Factor 5	1.8	1.3	4.5	2.3	0.5
Total	44.4	17.5	24.5	12.6	22.5
Partial least-squares regression					
Factor 1	53.5	24.6	23.4	2.1	7.0
Factor 2	1.9	0.7	39.9	1.8	14.0
Factor 3	21.6	2.5	0.1	8.3	0.4
Factor 4	0.1	0.4	12.5	2.0	6.5
Factor 5	4.1	7.7	0.1	1.7	0.04
Total	81.2	35.8	76.0	15.9	27.9
Reduced rank regression					
Factor 1	66.8	27.1	41.0	0.8	8.5
Factor 2	17.6	5.2	31.2	4.0	8.7
Factor 3	0.02	2.3	5.7	9.6	8.3
Factor 4	1.6	4.5	0.1	0.4	2.5
Factor 5	0.7	1.1	0.1	3.7	1.6
Total	86.9	40.2	78.1	18.5	29.6

Table 6. Model Fit Statistics for Regression of All Retained Factors Using Principal Component Analysis, Partial Least-Squares Regression, and Reduced Rank Regression in a Case-Control Study of First Incident Myocardial Infarction ($n = 3,574$), Costa Rica, 1994–2004

Model ^a	No. of Parameters	–2 Log Likelihood	Akaike's Information Criterion
Null (with matching factors)	1	2,477.31	2,477.31
Principal component analysis	29	2,009.24	2,065.24
Partial least-squares regression	29	2,010.56	2,066.56
Reduced rank regression	29	2,016.10	2,072.10

^a Results were adjusted for age, sex, county, income (in quintiles, plus an additional category for missing variables), current smoking status, physical activity (in quintiles), waist:hip ratio (in quintiles), total energy intake (in quintiles), history of hypertension, high cholesterol, and diabetes, and other dietary patterns. Additional adjustment for multivitamin use did not alter the results.

Additionally, when factor 1 was derived in the subsample of participants without various comorbid conditions (hypertension, high cholesterol, and diabetes), a slightly stronger inverse association was seen between this pattern and risk of myocardial infarction using PCA- and RRR-derived patterns (fifth quintile of dietary pattern vs. the first: PCA—odds ratio = 0.60, 95% confidence interval: 0.40, 0.88; PLS—odds ratio = 0.57, 95% confidence interval: 0.40, 0.82; RRR—odds ratio = 0.51, 95% confidence interval: 0.36, 0.73). There was an increased risk of myocardial infarction across increasing quintiles of factor 4 derived using PCA (palm oil and coffee pattern) and a decreased risk across increasing quintiles of factor 4 derived using PLS (alcohol and polyunsaturated oil pattern). Finally, there was a decreased risk of myocardial infarction across increasing quintiles of factor 5 derived using PCA (alcohol, legumes, and polyunsaturated oil pattern) (Table 7).

DISCUSSION

Costa Ricans served as a novel population in which to test the relative ability of 3 different methods to derive dietary patterns related to the risk of myocardial infarction.

We identified 4 dietary patterns associated with varying risks of myocardial infarction in Costa Rica. Factor 1 (vegetable pattern), characterized by high intake of lean chicken, vegetables, fruit, and polyunsaturated oil, was generated by all 3 dietary pattern methods and was robust across the many sensitivity analyses performed. This dietary pattern was associated with a markedly decreased risk of myocardial infarction and was inversely associated with saturated fat intake across all 3 dietary pattern methods. Interestingly, partially hydrogenated soybean oil was positively correlated with the vegetable pattern derived using PLS and RRR, suggesting that the potentially atherogenic effects of *trans*-fatty acids present in soybean oil in Costa Rica may be offset by the protective effects of α -linolenic acid.

PCA and PLS produced additional factors with strong negative correlations with saturated fat. Factor 4 from PLS analyses (alcohol and polyunsaturated oil pattern) and factor 5 from PCA (alcohol, legumes, and polyunsaturated oil pattern) were uniquely characterized across their quintiles by a decreasing intake of saturated fat and increasing intakes of wine, beer, and liquor. Intake of moderate amounts of alcohol has been linked to decreased risk of coronary heart disease in numerous studies (38) and may account for some of the cardioprotective effect associated with these patterns. The RRR method focuses solely on explaining variation in response variables, which may explain why this method did not produce a pattern characterized by alcohol intake and related to decreased risk of myocardial infarction, since alcohol intake was not included as a response variable in the analyses. Factor 4 from PCA analyses (palm oil and coffee pattern) was the only factor associated with increased risk of myocardial infarction, suggesting that palm oil still plays a role in the pathogenesis of myocardial infarction in this population despite its declining use (20). This factor was characterized by a high intake of saturated fat and a low adipose tissue level of α -linolenic acid. While PLS and RRR also produced patterns with this profile of increased saturated fat and low α -linolenic acid, these patterns were not positively associated with intake of coffee as was factor 4 from PCA analyses. While the literature on coffee and risk of myocardial infarction is mixed, there are compounds in coffee, namely diterpenes (present in unfiltered coffee) and caffeine, that have been linked to an increased risk of coronary heart disease (39). PCA may have produced a pattern with the unique combination of high palm oil and coffee intake, as this method is based on explaining variability in food intake without respect to response variability. Coffee intake is unlikely to account for much variability in the chosen response variables, potentially explaining the weak association of this beverage with PLS- and RRR-derived patterns.

The RRR-derived dietary pattern that was associated with decreased risk of myocardial infarction in this study is comparable to patterns seen in previous investigations. An RRR-derived pattern associated with decreased risk of coronary heart disease among Germans was characterized by high intake of whole grains, fruits and vegetables, olive oil, wine, and nuts (6), much like the vegetable pattern derived in this study. Previous studies among US and European populations derived patterns positively associated with coronary heart disease or subclinical markers of coronary heart disease that were characterized mainly by high intake of meat, fats, and oils high in saturated fat and low intake of vegetables (14, 18). The palm-oil-and-coffee pattern derived in this study using PCA, associated with increased risk of myocardial infarction, was similarly characterized by high intake of saturated fat but was only weakly associated with intake of meats and vegetables.

Contrary to previous studies, we found PCA and PLS to derive more patterns associated with disease risk than RRR. Sample size may be an important factor explaining these divergent results, since the previous study comparing these 3 methods included 193 cases and 385 controls (19), whereas our sample included 1,787 case-control pairs. However, our

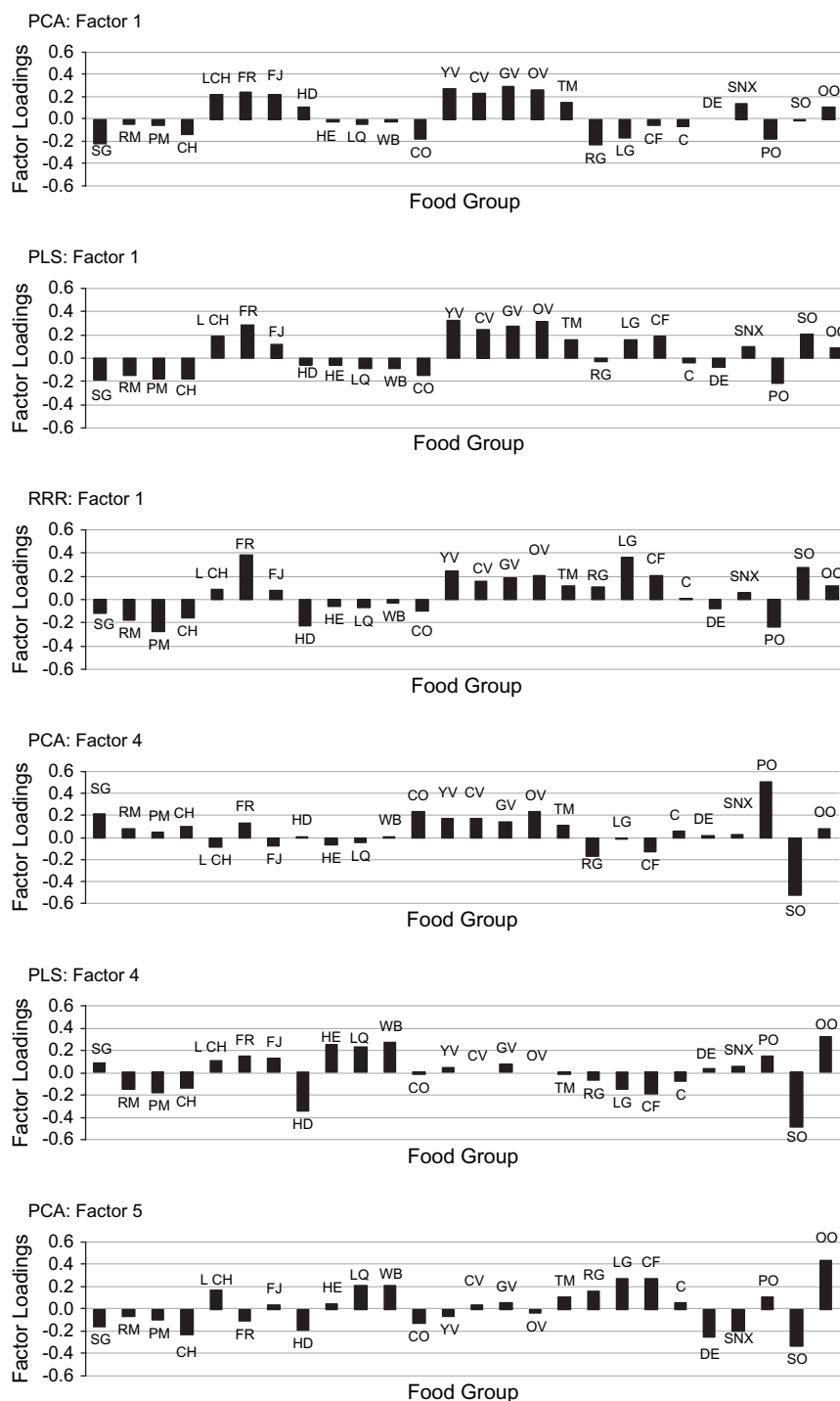


Figure 1. Main factor loadings for patterns derived using principal component analysis (PCA), partial least-squares regression (PLS), and reduced rank regression (RRR) that were significantly associated with myocardial infarction in a case-control study of first incident myocardial infarction ($n = 3,574$), Costa Rica, 1994–2004. Abbreviations for food groups used in dietary pattern analyses: SG, sugar; RM, red meat; PM, processed meat; CH, chicken with skin and fat; L CH, lean chicken; FR, fruit; FJ, fruit juice; HD, high-fat dairy foods; HE, high-energy drinks; LQ, liquor; WB, wine and beer; CO, coffee; YV, yellow vegetables; CV, cruciferous vegetables; GV, green leafy vegetables; OV, other vegetables; TM, tomatoes; RG, refined grains; LG, legumes; CF, condiments, fresh; C, condiments; DE, desserts; SNX, snacks; PO, palm oil; SO, soybean oil; OO, other oil.

results may also help to clarify the context in which each of these methods is most appropriate. In this study, RRR was the most limiting of the 3 methods, because it reduced

the range of possible myocardial infarction-related exposures to the subset of nutrients included as responses. When the goal of an investigation is to test hypotheses limited to

Table 7. Odds Ratios for Myocardial Infarction According to Quintiles of Significant Dietary Patterns in a Case-Control Study of First Incident Myocardial Infarction ($n = 3,574$), Costa Rica, 1994–2004^a

	Quintile of Factor Scores										P for Trend
	1 ^b	2		3		4		5			
		OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI		
Principal component analysis											
Factor 1	1.0	0.90	0.71, 1.15	0.68	0.52, 0.87	0.75	0.57, 0.97	0.72	0.55, 0.94		0.02
Factor 4	1.0	1.22	0.96, 1.56	1.09	0.86, 1.39	1.44	1.13, 1.85	1.38	1.07, 1.78		0.006
Factor 5	1.0	0.95	0.74, 1.20	0.85	0.66, 1.10	0.92	0.72, 1.17	0.69	0.53, 0.88		0.004
Partial least-squares regression											
Factor 1	1.0	0.62	0.49, 0.79	0.64	0.50, 0.82	0.66	0.52, 0.85	0.54	0.42, 0.70		<0.001
Factor 4	1.0	0.99	0.78, 1.26	0.82	0.65, 1.05	0.75	0.58, 0.97	0.77	0.60, 1.0		0.01
Reduced rank regression											
Factor 1	1.0	0.79	0.62, 1.00	0.85	0.67, 1.08	0.76	0.60, 0.97	0.66	0.52, 0.83		0.001

Abbreviations: CI, confidence interval; OR, odds ratio.

^a Results were adjusted for age, sex, county, income (in quintiles, plus an additional category for missing variables), current smoking status, physical activity (in quintiles), waist:hip ratio (in quintiles), total energy intake (in quintiles), history of hypertension, high cholesterol, and diabetes, and other significant dietary patterns. Additional adjustment for multivitamin use did not alter the results.

^b Reference category (OR = 1.0).

a preset group of response variables, RRR may be the most appropriate method. However, our goal in this study was to explore the relations between relevant dietary exposures in Costa Rica and myocardial infarction risk, and by applying all 3 methods we gained information about important myocardial infarction-related dietary exposures that were not explicitly included as response variables. PLS includes knowledge about intermediary variables on the pathway to disease in the derivation of dietary patterns and offers more flexibility than RRR, which may allow for the discovery of important disease-related dietary exposures in a population. PCA was the most flexible of the 3 methods evaluated in this investigation, allowing for the derivation of the largest number of patterns related to myocardial infarction, including the only pattern associated with increased risk of myocardial infarction. PCA has been criticized for being a purely data-driven method, but it may have advantages over PLS and RRR when a more exploratory analysis is warranted.

Strengths of this study include a large sample size ($n = 3,574$), the use of population-based controls, and the availability of biomarkers of dietary intake for α -linolenic and *trans*-fatty acids. This study also had potential limitations. First, dietary information was collected after cases had experienced a myocardial infarction. However, because associations between specific foods and health are not part of the general knowledge of the study population, systematic recall bias in reporting of foods is unlikely. Second, there is the possibility of residual confounding. Third, some of the effects of the dietary patterns in this study may be due to nutrients correlated with the chosen response variables, not to the responses themselves. Clinical trials are needed to separate the effects of particular nutrients from the overall dietary patterns derived in this study. Finally, PLS and RRR can produce chance associations with response variables

that do not generalize beyond the data set used to derive dietary patterns. We minimized this risk by using cross-validation to choose the patterns to investigate.

In conclusion, dietary pattern analysis allows for examination of the combined and potentially synergistic effects of many correlated dietary exposures. In this study, the PCA, PLS, and RRR methods identified similar dietary patterns associated with significantly decreased risk of myocardial infarction. However, PLS and PCA identified additional patterns that were significantly associated with myocardial infarction and were characterized by nutrients not previously considered to play a major role in the pathogenesis of myocardial infarction in this population. The flexibility allowed by PLS and PCA to derive patterns partially (PLS) or completely (PCA) unconstrained by the choice of response variables is important to consider when deciding which method to use in a particular population.

ACKNOWLEDGMENTS

Author affiliations: Department of Community Health, International Health Institute, Brown University, Providence, Rhode Island (Julia R. DiBello, Stephen T. McGarvey, Robert Goldberg, Ana Baylin); Center for Health Policy and Research, University of Massachusetts Medical School, Worcester, Massachusetts (Robert Goldberg); Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, Massachusetts (Peter Kraft); and Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts (Hannia Campos).

This study was supported by grants HL081549, HL071888, and HL60692 from the National Institutes of Health.

The authors are grateful to Xinia Siles for data collection and study management in Costa Rica.

Conflict of interest: none declared.

REFERENCES

- Hatcher L, ed. *A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling*. Cary, NC: SAS Institute Inc; 1994.
- Harris RJ, ed. *A Primer of Multivariate Statistics*. New York, NY: Academic Press, Inc; 1975.
- Hu FB, Rimm EB, Stampfer MJ, et al. Prospective study of major dietary patterns and risk of coronary heart disease in men. *Am J Clin Nutr*. 2000;72(4):912–921.
- 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.
- Osler M, Helms Andraesen A, Heitmann B, et al. Food intake patterns and risk of coronary heart disease: a prospective cohort study examining the use of traditional scoring techniques. *Eur J Clin Nutr*. 2002;56(7):568–574.
- Weikert C, Hoffmann K, Dierkes J, et al. A homocysteine metabolism-related dietary pattern and the risk of coronary heart disease in two independent German study populations. *J Nutr*. 2005;135(8):1981–1988.
- Kerver JM, Yang EJ, Bianchi L, et al. Dietary patterns associated with risk factors for cardiovascular disease in healthy US adults. *Am J Clin Nutr*. 2003;78(6):1103–1110.
- Martínez-Ortiz JA, Fung TT, Baylin A, et al. Dietary patterns and risk of nonfatal acute myocardial infarction in Costa Rican adults. *Eur J Clin Nutr*. 2006;60(6):770–777.
- Lopez-Garcia E, Schulze MB, Fung TT, et al. Major dietary patterns are related to plasma concentrations of markers of inflammation and endothelial dysfunction. *Am J Clin Nutr*. 2004;80(4):1029–1035.
- van Dam RM, Grievink L, Ocké MC, et al. Patterns of food consumption and risk factors for cardiovascular disease in the general Dutch population. *Am J Clin Nutr*. 2003;77(5):1156–1163.
- 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.
- Neumann AI, Martins IS, Marcopito LF, et al. Dietary patterns associated with risk factors for cardiovascular disease in a Brazilian city [in Portuguese]. *Rev Panam Salud Publica*. 2007;22(5):329–339.
- Harriss LR, English DR, Powles J, et al. Dietary patterns and cardiovascular mortality in the Melbourne Collaborative Cohort Study. *Am J Clin Nutr*. 2007;86(1):221–229.
- 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.
- Mikkilä V, Räsänen L, Raitakari OT, et al. Major dietary patterns and cardiovascular risk factors from childhood to adulthood. The Cardiovascular Risk in Young Finns Study. *Br J Nutr*. 2007;98(1):218–225.
- Shimazu T, Kuriyama S, Hozawa A, et al. Dietary patterns and cardiovascular disease mortality in Japan: a prospective cohort study. *Int J Epidemiol*. 2007;36(3):600–609.
- Fung TT, Rimm EB, Spiegelman D, et al. Association between dietary patterns and plasma biomarkers of obesity and cardiovascular disease risk. *Am J Clin Nutr*. 2001;73(1):61–67.
- Hoffmann K, Zyriax BC, Boeing H, et al. A dietary pattern derived to explain biomarker variation is strongly associated with the risk of coronary artery disease. *Am J Clin Nutr*. 2004;80(3):633–640.
- 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.
- Colón-Ramos U, Kabagambe EK, Baylin A, et al. Socio-economic status and health awareness are associated with choice of cooking oil in Costa Rica. *Public Health Nutr*. 2007;10(11):1214–1222.
- Kabagambe EK, Baylin A, Ascherio A, et al. The type of oil used for cooking is associated with the risk of nonfatal acute myocardial infarction in Costa Rica. *J Nutr*. 2005;135(11):2674–2679.
- Campos H, Bailey SM, Gussak LS, et al. Relations of body habitus, fitness level, and cardiovascular risk factors including lipoproteins and apolipoproteins in a rural and urban Costa Rican population. *Arterioscler Thromb*. 1991;11(4):1077–1088.
- Campos H, Siles X. Siesta and the risk of coronary heart disease: results from a population-based, case-control study in Costa Rica. *Int J Epidemiol*. 2000;29(3):429–437.
- Sen-Banerjee S, Siles X, Campos H. Tobacco smoking modifies association between Gln-Arg192 polymorphism of human paraoxonase gene and risk of myocardial infarction. *Arterioscler Thromb Vasc Biol*. 2000;20(9):2120–2126.
- Baylin A, Kabagambe EK, Siles X, et al. Adipose tissue biomarkers of fatty acid intake. *Am J Clin Nutr*. 2002;76(4):750–757.
- Willett WC, Sampson L, Stampfer MJ, et al. Reproducibility and validity of a semiquantitative food frequency questionnaire. *Am J Epidemiol*. 1985;122(1):51–65.
- Kabagambe EK, Baylin A, Allan DA, et al. Application of the method of triads to evaluate the performance of food frequency questionnaires and biomarkers as indicators of long-term dietary intake. *Am J Epidemiol*. 2001;154(12):1126–1135.
- Beynen AC, Katan MB. Rapid sampling and long-term storage of subcutaneous adipose-tissue biopsies for determination of fatty acid composition. *Am J Clin Nutr*. 1985;42(2):317–322.
- Baylin A, Kabagambe EK, Ascherio A, et al. Adipose tissue α -linolenic acid and nonfatal acute myocardial infarction in Costa Rica. *Circulation*. 2003;107(12):1586–1591.
- Baylin A, Kabagambe EK, Ascherio A, et al. High 18:2 trans-fatty acids in adipose tissue are associated with increased risk of nonfatal acute myocardial infarction in Costa Rican adults. *J Nutr*. 2003;133(4):1186–1191.
- Kabagambe EK, Furtado J, Baylin A, et al. Some dietary and adipose tissue carotenoids are associated with the risk of nonfatal acute myocardial infarction in Costa Rica. *J Nutr*. 2005;135(7):1763–1769.
- Kabagambe EK, Baylin A, Siles X, et al. Individual saturated fatty acids and nonfatal acute myocardial infarction in Costa Rica. *Eur J Clin Nutr*. 2003;57(11):1447–1457.
- Ignarro LJ, Balestrieri ML, Napoli C. Nutrition, physical activity, and cardiovascular disease: an update. *Cardiovasc Res*. 2007;73(2):326–340.
- Bazzano LA, He J, Ogden LG, et al. Dietary fiber intake and reduced risk of coronary heart disease in US men and women: The National Health and Nutrition Examination Survey I Epidemiologic Follow-up Study. *Arch Intern Med*. 2003;163(16):1897–1904.

35. Willett W, Stampfer MJ. Total energy intake: implications for epidemiologic analyses. *Am J Epidemiol*. 1986;124(1):17–27.
36. van der Voet H. Comparing the predictive accuracy of models using a simple randomization test. *Chemometr Intell Lab Syst*. 1994;25:313–323.
37. Tobias R. An introduction to partial least squares regression. In: *Proceedings of the Twentieth Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc; 1995:1250–1257.
38. Rimm EB, Klatsky A, Grobbee D, et al. Review of moderate alcohol consumption and reduced risk of coronary heart disease: is the effect due to beer, wine, or spirits? *BMJ*. 1996;312(7033):731–736.
39. Cornelis MC, El-Sohemy A. Coffee, caffeine, and coronary heart disease. *Curr Opin Clin Nutr Metab Care*. 2007;10(6):745–751.