American Journal of Epidemiology
Copyright © 2004 by the Johns Hopkins Bloomberg School of Public Health
All rights reserved

Vol. 159, No. 10
Printed in U.S.A.
DOI: 10.1093/aje/kwh134

Application of a New Statistical Method to Derive Dietary Patterns in Nutritional Epidemiology

Kurt Hoffmann[1], Matthias B. Schulze[2], Anja Schienkiewitz[1], Ute Nöthlings[1], and Heiner Boeing[1]

[1] Department of Epidemiology, German Institute of Human Nutrition, Bergholz-Rehbrücke, Germany.
[2] Department of Nutrition, Harvard School of Public Health, Boston, MA.

Received for publication August 12, 2003; accepted for publication December 9, 2003.

Because foods are consumed in combination, it is difficult in observational studies to separate the effects of single foods on the development of diseases. A possible way to examine the combined effect of food intakes is to derive dietary patterns by using appropriate statistical methods. The objective of this study was to apply a new statistical method, reduced rank regression (RRR), that is more flexible and powerful than the classic principal component analysis. RRR can be used efficiently in nutritional epidemiology by choosing disease-specific response variables and determining combinations of food intake that explain as much response variation as possible. The authors applied RRR to extract dietary patterns from 49 food groups, specifying four diabetes-related nutrients and nutrient ratios as responses. Data were derived from a nested German case-control study within the European Prospective Investigation into Cancer and Nutrition-Potsdam study consisting of 193 cases with incident type 2 diabetes identified until 2001 and 385 controls. The four factors extracted by RRR explained 93.1% of response variation, whereas the first four factors obtained by principal component analysis accounted for only 41.9%. In contrast to principal component analysis and other methods, the new RRR method extracted a significant risk factor for diabetes.

diabetes mellitus; diet; epidemiologic methods; nutrition; pattern analysis; statistics

Abbreviations: EPIC, European Prospective Investigation into Cancer and Nutrition; PCA, principal component analysis; PLS, partial least squares; RRR, reduced rank regression.

Epidemiologic studies on the relation between human diet and disease traditionally evaluated the effects of single nutrients or foods. However, these effects are often too small to detect. Moreover, the complexity of the human diet and especially the high correlation between intakes of various foods and nutrients makes it difficult to examine their separate effects (1, p. 22). Comprehensive dietary variables that allow for intake of many foods may show a greater effect on disease than any single component. Depending on its definition, each comprehensive variable reflects a specific dietary pattern that may represent a more accurate picture of diet than isolated foods (2).

Up to now, two different approaches have been used to derive dietary patterns. The first involves use of diet-quality scores based on recommended diets or dietary guidelines (3–10). Here, scientific evidence available prior to the current study is used to define dietary patterns. This technique, sometimes called an a priori or hypothesis-oriented approach (11–13), does not use intake data from the study to create pattern variables. The weakness of diet-quality scores is that they focus on selected aspects of diet and do not consider the correlation structure of food and nutrient intakes. Consequently, such scores do not reflect the overall effect of diet in general but only the formal sum of not-adjusted single effects.

The second approach is exploratory; thus, dietary patterns are derived from the data at hand. This approach ignores prior knowledge completely. Statistical exploratory methods that accomplish pattern derivation are principal component analysis (PCA) and factor analysis. Both are dimension-reduction techniques widely applied in nutritional epidemiology (14–32). Since they are very similar, we scrutinize the

Correspondence to Dr. Kurt Hoffmann, Department of Epidemiology, German Institute of Human Nutrition, Arthur-Scheunert-Allee 114-116, 14558 Nuthetal, Germany (e-mail: khoff@mail.dife.de).

benefit of PCA only. Applied to food intake data, PCA aims to explain the total variation in intake of many foods or food groups in terms of a few linear functions called principal components. The first principal component is the standardized linear function of foods with maximal variance, the second principal component maximizes the variance among all functions orthogonal to the first component, and so on. This procedure results in uncorrelated pattern scores that summarize and decompose the correlation structure of the original food items.

Unfortunately, PCA is sometimes not successful in deriving dietary patterns that are predictors of disease. Actually, in various applications of PCA to obtain dietary risk factors, the odds ratios for the first principal components or factors were not significantly different from 1 (12, 23, 26, 29, 31, 33). A possible reason for these disappointing results is that explaining as much variation in food intake as possible does not mean that much variation in important nutrients will be explained. Therefore, it might be wiser to focus on the variation in such nutrients that presumably affect the incidence of disease. The obvious idea of applying PCA to these nutrients is not attractive since patterns are then defined as linear functions of nutrients and therefore are not directly related to dietary habits because individual persons ultimately manipulate nutrient intake largely by their choice of foods (1, p. 21).

Apparently, a statistical method is needed that determines linear functions of predictors (foods) by maximizing the explained variation in responses (disease-related nutrients). Such a method already exists and is called reduced rank regression (RRR) or maximum redundancy analysis. However, to our knowledge, it has not yet been applied in epidemiology. RRR is neither an a priori nor a purely exploratory statistical method. Since it uses both information sources, data from the study and prior information for defining responses, it represents a so-called a posteriori method. The RRR method is implemented in SAS software (SAS Institute, Inc., Cary, North Carolina). Moreover, a third similar method called partial least squares (PLS), which is a compromise between PCA and RRR, is also available in SAS. We applied all three methods to data from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study to find out whether dietary patterns obtained by using these methods are predictive for type 2 diabetes mellitus. We also included some diet-quality scores in the comparative study to explore the efficiency of a priori methods.

MATERIALS AND METHODS

Study population

The study population was selected from participants of the EPIC cohort in Potsdam, Germany (34, 35), which contributes a general population sample of 27,548 subjects to the EPIC multicenter cohort study (36). Cases were those participants who were free of type 2 diabetes at baseline and developed type 2 diabetes during the first 2- to 3-year follow-up, depending on time of recruitment. Cases of incident diabetes were identified from self-reports, current medications, and current dietary treatment for diabetes and were verified by the primary care physician. A total of 193 cases were verified until November 1, 2001. Each case was matched by age and sex with two controls free of prevalent diabetes. One control for whom dietary variables were missing was not considered, thus leaving 385 controls for the statistical analysis.

Data collection

Baseline examination of the study population was carried out between August 1994 and September 1998. Study participants completed a self-administered food frequency questionnaire and a lifestyle questionnaire. The food frequency questionnaire assessed usual intake of 148 single food items during the 12 months before the examination. The food items were aggregated into 49 separate food groups based on culinary usage or nutrient profiles. The definition of the food groups has been published previously (24). Usual nutrient intake was estimated from the food items consumed by using the German Food Code (Federal Institute for Health Protection of Consumers and Veterinary Medicine, 1998).

Smoking status, educational attainment, and sports activity were assessed through personal computer–guided interviews in the study center. Smoking status was defined as current smoker or nonsmoker. Educational attainment was considered a dichotomized variable with the two categories "vocational training or lower degree" and "trade school, technical school, or university degree." Sports activity was calculated as number of hours of such activity per week averaged over 1 year. Anthropometric measurements of body height, body weight, waist girth, and hip circumference were performed by trained personnel (37) at baseline. Body mass index was calculated as weight in kilograms divided by height in meters squared. Waist-hip ratio was determined as the ratio of waist girth to hip circumference.

Statistical methods

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two sets of variables called predictors and responses, respectively. In the subsequent application, the predictors X_i are intakes of food groups in grams per day, whereas the responses Y_j are intakes of nutrients in grams per day or ratios of nutrient intakes. All three statistical methods (RRR, PCA, and PLS) work by extracting successive linear combinations of the predictors, called factors or components. However, the goals of these methods differ. The classic PCA method selects factors that explain as much predictor variation as possible. In contrast, RRR extracts factors that explain as much response variation as possible. The third method, PLS, balances the two goals of explaining predictor variation and explaining response variation.

The three methods are similar in terms of their mathematical foundation and their technique of deriving factors. For each method, the coefficient vectors of the extracted linear functions are eigenvectors of a covariance matrix. PCA uses the covariance matrix of predictors, whereas RRR starts from the covariance matrix of responses. PLS uses the matrix of covariances between predictors and responses. The eigenvalue belonging to an eigenvector quantifies the fraction of variation explained by the corresponding linear function of predictors. The factors obtained by PCA, PLS, and RRR usually are sorted by decreasing eigenvalues. The first factor of PCA is the linear function of predictors that maximizes the explained variation in predictors. However, in general, it is not optimal in terms of response variation. In contrast, the first factor of RRR explains more variation in response than any other linear function of predictors but possibly explains only a moderate fraction of predictor variation. Finally, the first factor of PLS maximizes the covariance between linear combinations of predictors and responses.

RRR starts from a linear function of responses called response score that will then be projected onto the space of predictors to produce a factor score, that is, a linear function of predictors. Both scores form an inseparable pair reflecting the same latent variable in different sets of original variables. Because the first aim of this method is to explain a high proportion of response variation, evaluation of factors extracted by RRR should be based on response scores rather than on factor scores. However, factor scores represent the comprehensive variables that will be used in subsequent statistical analysis.

A nice property of all three methods is that the successive extracted factors are uncorrelated. This property follows from the orthogonality of eigenvectors. Therefore, the variation in the original variables, predictors or responses, can be decomposed into fractions of variation explained by the obtained factors. Because different factors from one method are uncorrelated, they can simultaneously be chosen as independent variables in a regression model without confounding each other.

The PCA, PLS, and RRR methods are all implemented in the special procedure PLS of the SAS System for Windows, release 8.02 (SAS Institute, Inc.). This procedure includes as an option the choice of an optimal number of extracted factors by applying cross-validation and the randomization-based model comparison test proposed by van der Voet (38). On the other hand, the number of factors cannot be greater than the rank of the corresponding covariance matrix. In our application, the rank of the covariance matrix used for PCA and PLS is equal to the number *n* of food groups, whereas the rank is equal to the number *m* of selected nutrient-related responses for RRR. To ensure compatibility of the results, it is reasonable to choose the minimum of *n* and *m* as the uniform number of extracted factors for all methods. The SAS code for the applied SAS procedure PLS is given in the Appendix.

Definition of responses

The responses we used are nutrients and ratios of nutrients presumed to be important in the development of type 2 diabetes. Previous studies indicate that a higher intake of polyunsaturated fat could be beneficial (39, 40), whereas a higher intake of saturated fat could adversely affect glucose metabolism and insulin resistance (41, 42). Therefore, following the idea of Hu et al. (43), we chose the ratio of polyunsaturated fat intake to saturated fat intake as one response. As another response, we selected fiber intake because higher intake of fiber was associated with reduced diabetes risk in some large cohort studies (44–46). We also included dietary magnesium intake, which showed a strong inverse association with incidence of diabetes in the Iowa Women's Health Study (46). Finally, we chose alcohol consumption as the fourth response because moderate consumption of alcohol reduced significantly the risk of diabetes in some previous epidemiologic studies (47–52). Altogether, compared with persons free of diabetes, those with type 2 diabetes mellitus are likely to be characterized as having lower values for all four response variables.

Diet-quality scores

We defined a generalized diet-quality score on the basis of the four responses, analogous to Hu et al. (43). For this purpose, each response variable was at first categorized into quintiles, with the fifth quintile representing the lowest risk. Then, each participant was assigned a score defined as the sum of his or her quintile values for the four responses. Thus, each response was weighted equally regardless of different effect sizes and intercorrelations of responses. The score uses integer values of 4–20. Persons with high scores should have a lower risk of developing diabetes than those with low scores. To study the sensitivity of this approach, we defined four additional scores by omitting response variables one at a time. The modified score values varied between 3 and 15.

RESULTS

The mean responses for diabetes cases and controls are presented in table 1. Surprisingly, the crude means of all four responses did not differ significantly between cases and controls. Adjustment for age, sex, body mass index, waist-hip ratio, sports activity, smoking status, education, and total energy intake did not change the results materially. Only after additional adjustment for the other three response variables were the means significantly different for fiber and magnesium intakes and borderline significant for alcohol consumption and fat ratio. However, contrary to our conjecture, magnesium intake and ratio of polyunsaturated fat to saturated fat were on average higher for cases than for controls.

To study the associations between responses, we calculated Pearson's correlation coefficients separately for cases and controls (table 2). The highest correlation was determined for fiber and dietary magnesium intakes. High magnesium intake was also strongly associated with high consumption of alcohol. The correlation structure was somewhat different for cases and controls, indicating that some linear functions of all four responses can have more discriminating power for diabetes incidence than any single response.

The statistical methods PCA, PLS, and RRR were applied to explain variation in 49 predefined food groups as well as variation in the four response variables in the pooled data set of 578 cases and controls. The variation accounted for by each of the four factors is presented in table 3 for all three methods. The first four components of PCA explained 22.0 percent of food intake variation and 41.9 percent of response

TABLE 1. Mean responses for diabetes cases and controls in the European Prospective Investigation into Cancer and Nutrition-Potsdam diabetes study (193 cases, 385 controls), 2001

Response	Mean		Difference of means	95% CI*	p value
	Cases	Controls			
Polyunsaturated fat intake/saturated fat intake					
Model 1†	0.459	0.438	0.021	–0.003, 0.046	0.09
Model 2‡	0.458	0.438	0.020	–0.007, 0.047	0.15
Model 3§	0.462	0.437	0.025	–0.001, 0.052	0.07
Fiber intake (g/day)					
Model 1	22.1	22.6	–0.5	–1.6, 0.6	0.36
Model 2	22.1	22.6	–0.5	–1.4, 0.6	0.34
Model 3	21.8	22.8	–1.0	–1.7, –0.2	0.009
Magnesium intake (g/day)					
Model 1	0.353	0.341	0.012	–0.004, 0.029	0.15
Model 2	0.350	0.343	0.007	–0.002, 0.017	0.14
Model 3	0.352	0.342	0.010	0.003, 0.017	0.009
Alcohol consumption (g/day)					
Model 1	18.3	15.9	2.4	–1.3, 6.0	0.20
Model 2	16.7	16.8	–0.1	–3.7, 3.6	0.97
Model 3	14.9	17.7	–2.8	–5.8, 0.1	0.06

* CI, confidence interval.

† Not adjusted.

‡ Adjusted for age, sex, body mass index, waist-hip ratio, sports activity, smoking status, education, and total energy intake.

§ Adjusted for all variables included in model 2 and the three other response variables.

variation. In contrast, the four RRR factors explained only 13.1 percent of variation in food intake but accounted for most of the variation in the four selected responses (93.1 percent). The fractions of variation explained by the four PLS factors were between those for PCA and RRR.

Table 4 gives a more detailed picture of how much of the variation in the single-response variables was explained by the factors of the three statistical methods. Obviously, PCA failed to explain the major amount of variation in fat ratio, fiber intake, and magnesium intake. In contrast, PLS accounted for more than 67 percent of the variation in all single responses. However, RRR clearly outperformed the other two methods. RRR factors explained more than 77 percent of the variation in each response, with especially high percentages for alcohol consumption, fiber intake, and magnesium intake. The first RRR factor already accounted

TABLE 2. Pearson's correlation coefficients between responses, European Prospective Investigation into Cancer and Nutrition-Potsdam diabetes study (193 cases, 385 controls),* 2001

Response	Polyunsaturated fat intake/ saturated fat intake	Fiber intake (g/day)	Magnesium intake (g/day)	Alcohol consumption (g/day)
Polyunsaturated fat intake/ saturated fat intake		–0.09	–0.17	0.09
Fiber intake (g/day)	0.08		0.66†	–0.02
Magnesium intake (g/day)	–0.06	0.72†		0.51†
Alcohol consumption (g/day)	–0.05	–0.09	0.38†	

* Correlation coefficients for cases are presented above the diagonal and those for controls are located below the diagonal.

† Correlation significantly different from zero ($p < 0.001$).

TABLE 3. Explained variation in all food groups and responses by using different statistical methods, European Prospective Investigation into Cancer and Nutrition-Potsdam diabetes study (*n* = 578), 2001

	Explained variation in food groups*			Explained variation in responses†		
	Principal component analysis	Partial least squares	Reduced rank regression	Principal component analysis	Partial least squares	Reduced rank regression
Factor 1	7.3	6.6	4.2	16.8	32.1	44.2
Factor 2	5.9	4.6	3.7	4.8	22.1	26.1
Factor 3	4.8	4.2	3.3	15.1	17.5	19.6
Factor 4	4.0	3.1	1.9	5.2	9.4	3.2
Total	22.0	18.5	13.1	41.9	81.1	93.1

* Food items from the food frequency questionnaire were aggregated into 49 separate food groups (refer to table 1 of Schulze et al. (24)).

† Selected responses are ratio of polyunsaturated to saturated fat intake, fiber intake, magnesium intake, and alcohol consumption.

for 91.1 percent of the variation in magnesium intake and 60.7 percent of the variation in fiber intake. The second and third factors reflected considerable variation in alcohol consumption and fat ratio, respectively.

TABLE 4. Explained variation in single responses by using different statistical methods, European Prospective Investigation into Cancer and Nutrition-Potsdam diabetes study (*n* = 578), 2001

	Explained variation in			
	Poly-unsaturated fat intake/ saturated fat intake	Fiber intake	Magnesium intake	Alcohol consumption
Principal component analysis				
Factor 1	3.8	16.0	36.9	10.5
Factor 2	2.6	12.9	3.8	0.0
Factor 3	9.4	11.5	0.1	39.3
Factor 4	8.4	0.1	4.6	7.8
Total	24.2	40.5	45.4	57.6
Partial least squares				
Factor 1	2.4	37.5	69.3	18.9
Factor 2	3.4	23.1	0.0	62.2
Factor 3	59.5	9.4	1.2	0.1
Factor 4	2.0	10.5	18.8	6.0
Total	67.3	80.5	89.3	87.2
Reduced rank regression				
Factor 1	0.3	60.7	91.1	24.7
Factor 2	0.0	31.4	0.0	72.9
Factor 3	76.8	1.1	0.4	0.3
Factor 4	0.2	4.4	6.1	2.0
Total	77.3	97.6	97.6	99.9

To examine more thoroughly the relations between responses and factors, we calculated the coefficients of the response scores for all RRR factors (table 5). As already expected from the correlation structure of response variables (table 2), no factor existed with only positive or only negative response score coefficients. A high response score for factor 1 reflected a diet rich in magnesium and fiber as well as high alcohol consumption. The second response score was elevated for persons consuming alcohol but not much fiber, whereas a high score for the third factor reflected a diet with much polyunsaturated fat and not much saturated fat. Finally, a high response score for the fourth factor occurred if the person's intake of alcohol and fiber was high but of magnesium was low.

We used the four factor scores from each method as independent variables in a logistic regression model for type 2 diabetes mellitus. After adjustment for age, sex, body mass index, waist-hip ratio, sports activity, smoking status, education, and total energy intake, no factor score obtained by using the PCA and PLS methods was significantly associated with diabetes risk (table 6). Of the factors obtained by RRR, only the fourth one was found to be a risk factor for type 2 diabetes. The risk associated with an increase of one

TABLE 5. Response scores for the factors obtained by using reduced rank regression, European Prospective Investigation into Cancer and Nutrition-Potsdam diabetes study (*n* = 578), 2001

	Response score coefficient for			
	Polyunsaturated fat intake/ saturated fat intake	Fiber intake	Magnesium intake	Alcohol consumption
Factor 1	−0.04	0.59	0.72	0.37
Factor 2	0.01	−0.55	0.01	0.84
Factor 3	0.99	0.12	−0.07	0.06
Factor 4	−0.14	0.58	−0.69	0.40

TABLE 6. Relative risks and 95% confidence intervals for type 2 diabetes according to a standardized* increase in factor scores obtained by using different statistical methods, European Prospective Investigation into Cancer and Nutrition-Potsdam diabetes study (*n* = 578),† 2001

	Principal component analysis		Partial least squares		Reduced rank regression	
	RR‡	95% CI‡	RR	95% CI	RR	95% CI
Factor 1	0.92	0.66, 1.29	1.05	0.66, 1.67	1.16	0.76, 1.77
Factor 2	1.03	0.84, 1.28	1.09	0.86, 1.39	1.02	0.82, 1.26
Factor 3	0.82	0.64, 1.04	1.16	0.94, 1.43	1.10	0.89, 1.38
Factor 4	0.85	0.68, 1.07	1.05	0.84, 1.31	0.68	0.54, 0.85

* The factor score was increased by one standard deviation.

† Relative risks were calculated by using a logistic regression model containing simultaneously all four factors of a method and adjusted for age, sex, body mass index, waist-hip ratio, sports activity, smoking status, education, and total energy intake.

‡ RR, relative risk; CI, confidence interval.

standard deviation in this score was 0.68 (95 percent confidence interval: 0.54, 0.85).

We next compared RRR factor scores with diet-quality scores. Table 7 gives the relative risks of diabetes according to quintiles of scores. Again, the fourth RRR factor was the only dietary variable predictive of type 2 diabetes. The relative risks across increasing quintiles of this factor score, after adjustment for age, sex, body mass index, waist-hip ratio, sports activity, smoking status, education, and total energy intake, were 1.0, 0.92, 0.88, 0.65, and 0.49 (95 percent confidence interval: 0.25, 0.94; *p* for trend = 0.01). Surprisingly, no trend was perceptible for increasing diet-quality scores regardless of whether three or four responses were used.

In table 8, the food groups strongly associated with the diabetes-related RRR factor are presented. Together, the 10 food groups shown explained 85.8 percent of the factor score variation. The highest contributions to explained variation were from low-fat and high-fat dairy products, whole grain bread, and coffee. Whereas low-fat and high-fat dairy products, coffee, fruit juice, margarine, and processed meat had negative

TABLE 7. Relative risks and 95% confidence intervals for type 2 diabetes according to quintiles of reduced rank regression factors and diet-quality scores, European Prospective Investigation into Cancer and Nutrition-Potsdam diabetes study (*n* = 578),* 2001

	First quintile (RR†)	Second quintile		Third quintile		Fourth quintile		Fifth quintile		*p* for trend
		RR	95% CI†	RR	95% CI	RR	95% CI	RR	95% CI	
Reduced rank regression										
Factor 1	1.0	0.79	0.41, 1.52	0.64	0.32, 1.32	0.82	0.37, 1.77	0.76	0.29, 1.96	0.62
Factor 2	1.0	1.46	0.75, 2.86	1.62	0.82, 3.21	1.17	0.58, 2.36	1.04	0.51, 2.12	0.77
Factor 3	1.0	0.60	0.31, 1.16	0.82	0.43, 1.59	0.92	0.48, 1.74	1.32	0.70, 2.48	0.17
Factor 4	1.0	0.92	0.49, 1.71	0.88	0.47, 1.65	0.65	0.34, 1.23	0.49	0.25, 0.94	0.01
Diet-quality scores‡										
Score 1	1.0	0.72	0.38, 1.39	0.85	0.46, 1.59	0.96	0.49, 1.90	1.01	0.48, 2.11	0.89
Score 2	1.0	0.74	0.40, 1.38	0.51	0.26, 1.01	0.98	0.47, 2.06	0.45	0.18, 1.10	0.20
Score 3	1.0	1.04	0.56, 1.91	0.83	0.44, 1.56	1.29	0.59, 2.81	0.97	0.46, 2.06	0.98
Score 4	1.0	0.85	0.44, 1.64	1.27	0.74, 2.19	0.67	0.33, 1.37	1.07	0.54, 2.11	0.99
Score 5	1.0	1.96	1.06, 3.61	1.56	0.80, 3.07	1.64	0.82, 3.27	1.47	0.64, 3.37	0.43

* Relative risks were adjusted for age, sex, body mass index, waist-hip ratio, sports activity, smoking status, education, and total energy intake.

† RR, relative risk; CI, confidence interval.

‡ The four response variables were categorized into quintiles separately. Then, each study participant was assigned four numbers corresponding to the quintiles that contained his or her response values. Score 1 was defined as the sum of the four numbers. Scores 2, 3, 4, and 5 were defined as sums of only three numbers by omitting the quintile number of fat ratio, fiber intake, magnesium intake, and alcohol consumption, respectively.

TABLE 8. Food groups strongly associated with the diabetes-related factor score identified by using reduced rank regression, European Prospective Investigation into Cancer and Nutrition-Potsdam diabetes study (*n* = 578), 2001

Food group	Standardized score parameter	Correlation with score	Explained proportion of score variation (%)*
Low-fat dairy products	–0.50	–0.40	20.0
Whole grain bread	0.46	0.28	12.9
Coffee	–0.37	–0.31	11.5
High-fat dairy products	–0.36	–0.27	9.7
Wine	0.28	0.28	7.8
Fresh fruit	0.27	0.23	6.2
Fruit juice	–0.30	–0.18	5.4
Margarine	–0.22	–0.23	5.1
Processed meat	–0.19	–0.20	3.8
Spirits	0.19	0.18	3.4
All 10 food groups			85.8

* The score variation explained by a food group was calculated as the product of the corresponding standardized score parameter, the correlation coefficient, and 100%.

score coefficients and were negatively correlated to the factor score, whole grain bread, fresh fruit, wine, and spirits were characterized by positive score coefficients and correlation. Because an increase in the fourth RRR factor score decreased diabetes risk (tables 6 and 7), food groups positively associated with this factor score had a beneficial effect, whereas food groups with negative score coefficients had a detrimental effect regarding the incidence of diabetes. Moreover, the effect assigned to food groups can be interpreted by the corresponding response score for the fourth factor (table 5). For example, increased intake of whole grain bread, because it contains much fiber, reduced the risk of diabetes.

DISCUSSION

The statistical method RRR is a powerful tool for deriving dietary patterns important in nutritional epidemiology. In contrast to the classic PCA method, RRR can be applied to explain variation in nutrients or nutrient-related responses by linear functions of food intakes. Since biologic knowledge concerning development of a disease is based on the role of nutrients rather than of foods, dietary patterns obtained by using RRR should better clarify the importance of diet in the etiology of diseases.

The most appealing feature of RRR is the possibility of choosing disease-specific responses. Therefore, prior knowledge gained from biologic evidence, dietary intervention studies, epidemiologic studies with biomarkers, and large prospective cohort studies can be incorporated. The RRR method combines two information sources, prior information and the data from the study. It also combines the strength of PCA to consider correlation of dietary components and the advantage of diet-quality scores to account for current scientific evidence. Diet-quality scores will be significant risk factors for disease if the selected score components do not strongly interact and if their presumed effects do not contradict study findings. In the more realistic case of correlated responses and unknown mixture of effects, RRR will outperform diet-quality scores. RRR is more flexible and less sensitive to violations of assumptions concerning directional relations because it determines only response variables without fixing equal weights and expected directions of effects. Clearly, in the extreme case of missing prior knowledge, no response variables can be justified; therefore, the explorative PCA method should be preferred to RRR.

The only diabetes-related dietary pattern we could determine was the fourth factor score of RRR. This factor explained only 3.2 percent of response variation. This finding raises the question of whether such a tiny proportion reflects a direction of infrequent response variation possibly predictive for diabetes. When all response scores from using the three methods were explored, the response score for the fourth RRR factor was found to be the only one for which score coefficients for fiber and magnesium intakes had high absolute values of the opposite sign. This observation suggests that both nutrients have opposite effects on diabetes that generally offset, and the exception is reflected by, the fourth RRR pattern. This explanation can be confirmed by the high positive correlation between both nutrients (table 2) and the significance of intake differences between cases and controls after adjustment for the correlated nutrient (table 1). Obviously, the role of magnesium intake in the development of diabetes can be evaluated only by simultaneous consideration of other correlated nutrients, as performed by the RRR method. Diet-quality scores defined by the responses failed to be predictive for diabetes because they ignored the high correlation between magnesium and fiber intakes and assumed that high magnesium intake always decreased the risk of type 2 diabetes.

The objective of this study was a methodological one and consisted of presenting a statistical alternative to PCA and exploratory factor analysis to derive dietary factors. Besides the inability to use prior knowledge, main reservations to PCA and factor analysis refer to the arbitrariness in determining the number of factors extracted and the interpretation of factors. Rationally, only factors with high corresponding eigenvalues should be chosen. However, any lower bound for eigenvalues, whether predetermined or determined by a scree plot, is arbitrary. Generally, in previous epidemiologic studies, no more than four PCA factors have been extracted. Consistent with the epidemiologic literature, we also extracted the first four factors but also examined whether the subsequent factors would create diabetes-related dietary patterns. As a result, the first 10 PCA factors were found to have no association with diabetes incidence at all. In contrast to PCA, choosing the number of extracted factors in RRR is generally no serious problem. Because no more factors than responses exist and the number of responses will be small in most applications, it is obvious to extract the maximal number of factors. Thus, RRR reduces the dimension of predictor variables to the dimension of response variables. In the present study, we chose the maximum of four factors.

Extracting less than the maximal number of factors can lead to overlooking disease-related dietary patterns, as can be seen in the diabetes case-control study.

The second objection to PCA concerns interpretation of factors that include its role in the etiology of diseases. If a dietary pattern obtained by using PCA turns out to be a risk factor for a specific disease, a plausible explanation is often difficult to find. Although we know which food groups substantially contribute to the factor by looking for high factor loadings, it remains unclear why these food groups are important in the incidence of disease. The RRR method may help overcome this weakness. Factor scores extracted by RRR can always be evaluated by their corresponding response scores and by the explained variation in response variables that should be related to disease. On the other hand, associations between food groups and responses can be used to interpret beneficial or detrimental effects of food groups as components of dietary patterns.

However, application of RRR in nutritional epidemiology has several limitations. First, factor scores are not well-measured characteristics of diet but linear combinations of food intakes usually obtained by a food frequency questionnaire. Assessing dietary intake by food frequency questionnaire is subject to considerable measurement error, as demonstrated by the OPEN Study (53). Second, the coefficients of a factor score are estimated by using the data at hand and cannot be reproduced with data from another study population. For example, using the full cohort of the EPIC-Potsdam study instead of the nested case-control study results in somewhat different dietary patterns, although the fourth RRR factor remains a significant predictor of diabetes. A possible approach to reduce the data dependency of the pattern variables is to construct simplified dietary patterns by omitting food groups with low score coefficients and ignoring the weights of the remaining food groups (25). Moreover, simplified patterns are easier to interpret than patterns including all food groups. Third, although the RRR method aims to explain much response variation, the predictive value of all response variables together can be higher than the one for all RRR factors. For example, the four nutrients we considered predict diabetes better than the four RRR factor scores. However, a diabetes model with nutrients as independent variables does not tell us what foods will reduce diabetes risk.

Altogether, RRR seems to be an appropriate and promising statistical method to determine which dietary patterns are associated with development of diseases by combining prior information and dietary information from the study. Nevertheless, the usefulness of RRR needs to be confirmed in future studies for other diseases and other disease-related variables chosen as responses.

REFERENCES

1. Willett W. Nutritional epidemiology. Oxford, United Kingdom: Oxford University Press, 1998.
2. Hu F. Dietary pattern analysis: a new direction in nutritional epidemiology. Curr Opin Lipidol 2002;13:3–9.
3. Kennedy ET, Ohls J, Carlson S, et al. The healthy eating index: design and applications. J Am Diet Assoc 1995;95: 1103–8.
4. Kant AK. Indexes of overall diet quality: a review. J Am Diet Assoc 1996;96:785–91.
5. Huijbregts P, Feskens E, Rasanen L, et al. Dietary patterns and 20 year mortality in elderly men in Finland, Italy, and the Netherlands: longitudinal cohort study. BMJ 1997;315:13–17.
6. Haines PS, Siega-Riz AM, Popkin BM. The diet quality index revised: a measurement instrument for populations. J Am Diet Assoc 1999;99:697–704.
7. Kant AK, Schatzkin A, Graubard BI, et al. A prospective study of diet quality and mortality in women. JAMA 2000; 283:2109–15.
8. McCullough ML, Feskanich D, Stampfer MJ, et al. Adherence to the dietary guidelines for Americans and risk of major chronic disease in women. Am J Clin Nutr 2000;72:1214–22.
9. 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;72:1223–31.
10. 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: 1261–71.
11. Trichopoulos D, Lagiou P. Dietary patterns and mortality. Br J Nutr 2001;85:133–4.
12. 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;85:219–25.
13. Schulze MB, Hu FB. Dietary patterns and risk of hypertension, type 2 diabetes mellitus, and coronary heart disease. Curr Atheroscler Rep 2002;4:462–7.
14. Schwerin HS, Stanton JL, Riley AM Jr, et al. Food eating patterns and health: a reexamination of the Ten-State and HANES I surveys. Am J Clin Nutr 1981;34:568–80.
15. Gex-Fabry M, Raymond L, Jeanneret O. Multivariate analysis of dietary patterns in 939 Swiss adults: sociodemographic parameters and alcohol consumption profiles. Int J Epidemiol 1988;17:548–55.
16. Randall E, Marshall JR, Graham S, et al. Patterns in food use and their associations with nutrient intakes. Am J Clin Nutr 1990;52:739–45.
17. Whichelow MJ, Prevost AT. Dietary patterns and their association with demographic, lifestyle and health variables in a random sample of British adults. Br J Nutr 1996;76:17–30.
18. Gittelsohn J, Wolever TM, Harris SB, et al. Specific patterns of food consumption and preparation are associated with diabetes and obesity in a native Canadian community. J Nutr 1998;128:541–7.
19. Slattery ML, Boucher KM, Caan BJ, et al. Eating patterns and risk of colon cancer. Am J Epidemiol 1998;148:4–16.
20. Slattery ML, Edwards SL, Boucher KM, et al. Lifestyle and colon cancer: an assessment of factors associated with risk. Am J Epidemiol 1999;150:869–77.
21. Hu FB, Rimm E, Smith-Warner SA, et al. Reproducibility and validity of dietary patterns assessed with a food-frequency questionnaire. Am J Clin Nutr 1999;69:243–9.
22. Williams DEM, Prevost TP, Whichelow MJ, et al. A cross-sectional study of dietary patterns with glucose intolerance and other features of the metabolic syndrome. Br J Nutr 2000; 83:257–66.
23. Osler M, Helms AA, Heitmann B, et al. Food intake patterns and risk of coronary heart disease: a prospective cohort study examining the use of traditional score techniques. Eur J Clin Nutr 2002;568–74.
24. Schulze MB, Hoffmann K, Kroke A, et al. Dietary patterns and their association with food and nutrient intake in the European Prospective Investigation into Cancer and Nutrition

(EPIC)-Potsdam study. Br J Nutr 2001;85:363–73.

25. Schulze MB, Hoffmann K, Kroke A, et al. An approach to construct simplified measures of dietary patterns from exploratory factor analysis. Br J Nutr 2003;89:409–18.
26. 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;158:365–73.
27. Fung TT, Willett WC, Stampfer MJ, et al. Dietary patterns and the risk of coronary heart disease in women. Arch Intern Med 2001;161:1857–62.
28. 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;73:61–7.
29. Fung T, Hu FB, Fuchs C, et al. Major dietary patterns and the risk of colorectal cancer in women. Arch Intern Med 2003; 163:309–14.
30. Terry P, Hu FB, Hansen H, et al. Prospective study of major dietary patterns and colorectal cancer risk in women. Am J Epidemiol 2001;154:1143–9.
31. Terry P, Suzuki R, Hu FB, et al. A prospective study of major dietary patterns and the risk of breast cancer. Cancer Epidemiol Biomarkers Prev 2001;10:1281–5.
32. Van Dam RM, Rimm EB, Willett WC, et al. Dietary patterns and risk for type 2 diabetes mellitus in U.S. Ann Intern Med 2002;136:201–9.
33. Masaki M, Sugimori H, Nakamura K, et al. Dietary patterns and stomach cancer among middle-aged male workers in Tokyo. Asian Pac J Cancer Prev 2003;4:61–6.
34. Boeing H, Wahrendorf J, Becker N. EPIC-Germany—a source for studies into diet and risk of chronic diseases. Ann Nutr Metab 1999;43:195–204.
35. Boeing H, Korfmann A, Bergmann MM. Recruitment procedures of EPIC-Germany. Ann Nutr Metab 1999;43:205–15.
36. Riboli E, Kaaks R. The EPIC Project: rationale and study design. Int J Epidemiol 1997;26(suppl 1):S6–14.
37. Kroke A, Bergmann MM, Lotze G, et al. Measures of quality control in the German component of the EPIC study. Ann Nutr Metab 1999;43:216–24.
38. van der Voet H. Comparing the predictive accuracy of models using a simple randomization test. Chemometrics Intell Lab Syst 1994;25:313–23.
39. Storlien LH, Kriketos AD, Jenkins AB, et al. Does dietary fat influence insulin action? Ann N Y Acad Sci 1997;827:287–301.
40. Salmeron J, Hu FB, Manson JE, et al. Dietary fat intake and risk of type II diabetes in women. Am J Clin Nutr 2001;13: 1019–27.
41. Feskens EJ, Virtanen SM, Rasanen L, et al. Dietary factors determining diabetes and impaired glucose tolerance. A 20-year follow-up of the Finnish and Dutch cohorts of the Seven Countries Study. Diabetes Care 1995;18:1104–12.
42. Hu FB, van Dam RM, Liu S. Diet and risk of type II diabetes: the role of types of fat and carbohydrate. Diabetologia 2001; 44:805–17.
43. Hu FB, Manson JE, Stampfer MJ, et al. Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. N Engl J Med 2001; 345:790–7.
44. Salmeron J, Ascherio A, Rimm EB, et al. Dietary fiber, glycemic load, and risk of NIDDM in men. Diabetes Care 1997;20: 545–50.
45. Salmeron J, Manson JE, Stampfer MJ, et al. Dietary fiber, glycemic load, and risk of non-insulin-dependent diabetes mellitus in women. JAMA 1997;277:472–7.
46. Meyer KA, Kushi LH, Jacobs DR Jr, et al. Carbohydrates, dietary fiber, and incident type II diabetes in older women. Am J Clin Nutr 2000;71:921–30.
47. Ajani UA, Hennekens CH, Spelsberg A, et al. Alcohol consumption and risk of type 2 diabetes mellitus among US male physicians. Arch Intern Med 2000;160:1025–30.
48. Wei M, Gibbons LW, Mitchell TL, et al. Alcohol intake and incidence of type 2 diabetes in men. Diabetes Care 2000;23: 18–22.
49. Conigrave KM, Hu BF, Camargo CA, et al. A prospective study of drinking patterns in relation to risk of type 2 diabetes among men. Diabetes 2001;50:2390–5.
50. De Vegt F, Dekker JM, Groeneveld WJ, et al. Moderate alcohol consumption is associated with lower risk for incident diabetes and mortality: the Hoorn Study. Diabetes Res Clin Pract 2002;57:53–60.
51. Wannamethee SG, Shaper AG, Perry IJ, et al. Alcohol consumption and the incidence of type II diabetes. J Epidemiol Community Health 2002;56:542–8.
52. Wannamethee SG, Camargo CA, Manson JE, et al. Alcohol drinking patterns and risk of type 2 diabetes mellitus among younger women. Arch Intern Med 2003;163:1329–36.
53. Kipnis V, Subar AF, Midthune D, et al. Structure of dietary measurement error: results of the OPEN biomarker study. Am J Epidemiol 2003;158:14–21.
54. SAS Institute, Inc. SAS/STAT user's guide, version 8. Cary, NC: SAS Institute, Inc, 1999:2693–734.

APPENDIX

SAS Code for Deriving Dietary Patterns

This Appendix provides the SAS code to construct dietary patterns by using the RRR, PCA, and PLS methods. We used the same SAS procedure *pls* with the same options for all three methods. For more details concerning other possible options and their SAS syntax, refer to the *SAS User's Guide* (54).

Our data file named *diet* contained 49 food groups denoted by *group1*, …, *group49* and the four nutrients *fatratio*, *fiber*, *mg*, and *alcohol*. The following SAS code was used to construct dietary patterns by using the RRR method:

```
proc pls data=diet method=RRR
        nfac=4 varss details;
        model fatratio fiber mg alcohol
              = group1-group49;
        output out=pattern xscore=scorex yscore=scorey;
run;
```

The first row invokes the *pls* procedure and indicates that the data from file *diet* will be analyzed by RRR. In the second row, the option nfac=4 specifies the number of factors to extract. The further options *varss* and *details* concern the displayed output. By default, the procedure *pls* displays just the amount of predictor and response variation accounted for by each factor. The option *varss* additionally lists the amount of variation accounted for in each response and predictor; the option *details* lists the details of the fitted model for each successive factor. The model statement defines the four nutrients *fatratio*, *fiber*, *mg*, and *alcohol* as response variables and the 49 food groups *group1*, …, *group49* as predictors. In the output statement, a new data set named *pattern* is created to receive output variables of the procedure, such as extracted factors and predicted values. The options xscore=scorex and yscore=scorey produce four factor and four response scores denoted by scorex1, …, scorex4 and scorey1, …, scorey4, respectively. The factor scores can be used in subsequent statistical analysis, whereas the response scores can be helpful in interpreting factors.

PCA and PLS dietary patterns can also be derived by applying the SAS procedure *pls*. In our example, we did so by changing the option method=RRR to method=PCR and method=PLS, respectively, and letting all other statements and options from the SAS code given above remain unchanged.