Public Health Nutrition: 4(5), 989–997 DOI: 10.1079/PHN2001168

Analysis of patterns of food intake in nutritional epidemiology: food classification in principal components analysis and the subsequent impact on estimates for endometrial cancer

Susan E McCann[1,*], James R Marshall[2], John R Brasure[1], Saxon Graham[1] and Jo L Freudenheim[1]

[1]Department of Social and Preventive Medicine, 270 Farber Hall, University at Buffalo, Buffalo, NY 14214, USA: [2]Arizona Cancer Center, 1515 N. Campbell Avenue, PO Box 245024, Tucson, AZ 85724-5024, USA

Submitted 4 January 2001: Accepted 6 March 2001

Abstract

Objective: To assess the effect of different methods of classifying food use on principal components analysis (PCA)-derived dietary patterns, and the subsequent impact on estimation of cancer risk associated with the different patterns.

Methods: Dietary data were obtained from 232 endometrial cancer cases and 639 controls (Western New York Diet Study) using a 190-item semi-quantitative food-frequency questionnaire. Dietary patterns were generated using PCA and three methods of classifying food use: 168 single foods and beverages; 56 detailed food groups, foods and beverages; and 36 less-detailed groups and single food items.

Results: Classification method affected neither the number nor character of the patterns identified. However, total variance explained in food use increased as the detail included in the PCA decreased (~8%, 168 items to ~17%, 36 items). Conversely, reduced detail in PCA tended to attenuate the odds ratio (OR) associated with the healthy patterns (OR 0.55, 95% confidence interval (CI) 0.35–0.84 and OR 0.77, 95% CI 0.49–1.20, 168 and 36 items, respectively) but not the high-fat patterns (OR 0.95, 95% CI 0.57–1.58 and OR 0.85, 0.51–1.40, 168 and 36 items, respectively).

Conclusions: Greater detail in food-use information may be desirable in determination of dietary patterns for more precise estimates of disease risk.

Keywords
Dietary patterns
Endometrial cancer
Statistical methods

Diet is a complex behaviour often correlated with non-dietary behaviours such as cigarette smoking and physical activity. Furthermore, intake of a single nutrient is often associated with intakes of other nutrients and dietary components, all of which may have interdependent physiological effects and common food sources[1–3]. The accurate and appropriate characterisation of dietary intake has long been a source of methodological concern in studies of diet and disease. For many years, the focus of research was on individual nutrients as risk factors for an outcome of interest. As a single food may provide a source of multiple nutrients and other components, risk associated with individual foods or food groups has also been considered. However, these approaches may not fully incorporate the complicated behaviour of food consumption. Recently, attempts have been made to characterise detailed dietary intake information using statistical techniques such as principal components analysis (PCA) that reduce this detail into a smaller number of factors which describe specific patterns of dietary behaviour[4–6]. Individuals may then be arrayed on their scores for the identified factors, and risk of a disease calculated for levels of the factor scores. To date, this approach has been primarily used in aetiological studies of colon cancer[7,8] and, most recently, for coronary heart disease[9].

Previous studies involving dietary patterns and colon cancer have suggested that dietary patterns explain risk in addition to that from single foods and nutrients. Although the use of dietary patterns to describe dietary behaviour in association with risk of a disease appears intuitively practical, the method has been criticised for the subjective manner in which food use is classified and reduced before PCA is applied[10]. Furthermore, the original purpose of PCA was to reduce a larger amount of detailed information into a smaller number of interpretable factors that also have the characteristic of explaining the largest amount of variability in the targeted behaviour[11]. Although this has been a demonstrably successful technique in the social sciences, the use of this technique

*Corresponding author: Email mccann@acsu.buffalo.edu

to efficiently explain dietary behaviour may be less successful. Whereas the goal of principal components analysis should be to explain 80 to 90% of the variance in dietary intake, in the available literature, the percentage variance explained totals no more than 37% for interpretable identified dietary patterns[7,8]. The small amount of variability explained by dietary patterns is often justified as being a result of the multidimensionality of diet, the authors preferring to rely more on the interpretability of the identified factors[6,12].

In this study, we had several aims. Firstly, we wanted to determine if the manner in which detailed food frequency data was classified and reduced prior to PCA would affect dietary pattern identification. Secondly, we wished to examine how classification might affect the percentage variance explained in frequency of food use. Finally, we were interested in how classification might impact estimation of disease risk associated with the identified dietary patterns. We had previously investigated the relationship between intakes of specific foods and nutrients and endometrial cancer[13], and were interested in the use of PCA-derived dietary patterns in estimating risk of this cancer with diet. Therefore, we assessed this final aim using data from a case–control study of diet and endometrial cancer.

Methods

Data collection

The present study utilised data from a series of case–control studies of diet and cancer of the breast, endometrium, ovary and prostate in western New York, the methods for which have been described in detail elsewhere[13]. The study protocol was approved by the Institutional Review Board of the State University of New York at Buffalo, and informed consent was obtained from all subjects. Briefly, cases for each site were identified in western New York hospitals by trained nurse case-finders. Controls were randomly selected from driver's licence lists for women <65 years of age and from Health Care Finance Administration lists for women ≥65 years of age, and frequency-matched to cases on age and county of residence. Of 1685 eligible female controls contacted, 863 (51%) were interviewed. The primary reason for not being interviewed was refusal (47%). For determination of dietary patterns, dietary data from the entire female control series were utilised ($n = 863$). For analyses of risk of endometrial cancer associated with dietary patterns, 213 controls with prior hysterectomies and 11 women reporting menopause before age 37 were excluded (because there were no cases with menopause before this age) to give a total of 639 controls.

Between October 1986 and March 1991, 523 histologically confirmed cases of endometrial carcinoma were identified in women between the ages of 40 and 85 years. One hundred and twenty-four women (24%) were not interviewed because their physician refused permission or could not be reached, 106 (20%) refused to participate, 10 (2%) had died or were too ill to participate, and 18 (3%) were not interviewed for other reasons. Cases with other concomitant primary cancers ($n = 5$) or non-adenomatous carcinoma of the endometrium ($n = 27$) were excluded, leaving 232 cases for analyses.

Women were interviewed in their homes by trained nurse interviewers using a detailed interview. Interviews took approximately two hours to complete and included questions on diet, reproductive history, family history of cancer, medical history, health habits such as cigarette smoking and physical activity, and other lifestyle and occupational factors.

Diet in the year two years before the interview was investigated by a series of detailed questions regarding usual frequency and quantity of intake of 168 foods and beverages. Portion size for each food was established by reference to food models. Additional questions regarding seasonality of use and food preparation methods resulted in 190 total items concerning food use. Nutrient intake was calculated using food composition data from the United States Department of Agriculture (USDA) and published food composition tables[14,15]. Composition data for individual carotenoids were obtained from the National Cancer Institute Carotenoid Food Composition database, version 1, 1993[16,17].

Classification of food use

Principal components analysis was performed using dietary data from the entire Western New York Diet Study female control series ($n = 863$). This form of analysis may be affected by the number of items included in the determination of factors, especially if the number of subjects from which data are obtained is limited. Although we had sufficient numbers of subjects ($n = 863$) to include all the food items queried ($n = 190$), many of the items on our diet questionnaire could be over-represented in a principal components analysis because of questions concerning in-season use and use during other months of the year. As one of the aims of the study was to assess the effect of classification of food use on dietary patterns identified by principal components analysis, three separate schemes were used to recode weekly frequency of food use into smaller numbers of items for analysis (Table 1).

In Method 1, for those foods for which seasonal use was queried (tomatoes, spinach, greens, asparagus, green beans, yellow beans, corn, cantaloupe, other melon, grapes, berries, apples, pears, peaches, apricots, prunes or plums, pineapple, grapefruit, cherries, and oranges or tangerines), average weekly frequency of use was calculated as the average of the sum of in-season weekly frequency of use and other months' weekly frequency of use. Total weekly frequency of use of each fruit item was calculated as the sum of weekly frequency of use of the

https://doi.org/10.1079/PHN2001168 Published online by Cambridge University Press

Table 1 Classification of foods utilised in the three principal components methods of identifying dietary patterns

Method 1 (168 items)	Method 2 (56 items)	Method 3 (36 items)
Green and red peppers, mushrooms, onions, cucumbers, radishes, celery, lettuce	Salad	Vegetables
Cauliflower, cabbage, broccoli, Brussels sprouts, sauerkraut	Cruciferous	
Tomatoes, spinach, carrots, broccoli, greens, winter squash, sweet potatoes	High-carotenoid vegetables	
Asparagus, green and yellow beans, corn, peas, beets, lima beans, summer squash	Other vegetables	
Coleslaw	Coleslaw	
Pickles and olives	Pickles and olives	
Mashed potatoes, baked potatoes, fried potatoes, and other white potatoes	Mashed potatoes, baked potatoes, fried potatoes, and other white potatoes	
Baked beans	Baked beans	
Tomato juice	Tomato juice	
Cantaloupe and other melon, grapes, fresh berries, apples, pears, peaches, apricots, prunes or plums, pineapple, grapefruit, cherries, oranges or tangerines, bananas, lemons or limes	Fresh fruit	Fruit
Canned or frozen berries, apple sauce, pears, peaches, apricots, prunes or plums, grapefruit, cherries, oranges or tangerines, fruit cocktail	Canned or frozen fruit	
Raisins, other dried fruit	Dried fruit	
Olives		
Apple juice, orange or citrus juice	Fruit juice	
Margarine	Margarine	Fats, regular
Butter	Butter	
Mayonnaise and miracle whip	Mayonnaise and miracle whip	
Gravy		
Salad dressing	Salad dressing	
Reduced-calorie mayonnaise, salad dressing	Reduced-calorie mayonnaise, salad dressing	Fats, reduced-calorie
Hard cheese, processed cheese, sour cream and dips, half and half, whipped cream	Dairy, high-fat	Dairy, high-fat
Cottage cheese, 2% milk, skimmed milk	Dairy, low-fat	Dairy, low-fat
Yoghurt	Yoghurt	
Eggs	Eggs	Eggs
Cold cereal, other cooked cereal, white bread, white rolls, English muffins, bagels, or biscuits, other muffins, pancakes or waffles, French toast, doughnuts or pastries, rice, noodles	Grains, refined	Grains, refined
Bran cereals, oatmeal, dark bread, dark rolls	Grains, whole	Grains, whole
Popcorn, salty snacks, wheat crackers, saltines	Snacks	Snacks
Ice milk or sherbet, ice cream, cookies, pound or sponge cake, other cake, eclairs or cream puffs, cheese cake, custard or cream pies, fruit pies, pumpkin pies, pudding, brownies	Desserts	Desserts
Jams or jellies	Jams or jellies	Jams or jellies
Candy bars, chocolate candy, other candy	Candy	Candy
Peanut butter, peanuts, other nuts	Nuts	Nuts
Steak, round steak, hamburger patties, other hamburger, other beef, veal, lamb	Red meat	Meat
Beef or calves liver, chicken liver	Liver	
Pork roast, pork chops, spare ribs, ham, breakfast sausage, sausage, bacon	Pork	

https://doi.org/10.1079/PHN2001168 Published online by Cambridge University Press

Table 1. *continued*

Method 1 (168 items)	Method 2 (56 items)	Method 3 (36 items)
Hot dogs, bologna, liverwurst, other cold cuts, pepperoni	Processed meat	Processed meat
Fresh or frozen fish, canned fish, shrimp, other shellfish	Fish	Fish
Chicken wings, fried chicken, other chicken, other poultry	Poultry	Poultry
Macaroni and cheese	Macaroni and cheese	Macaroni and cheese
Spaghetti, lasagne or other pasta with tomato sauce	Spaghetti, lasagne or other pasta with tomato sauce	Spaghetti, lasagne or other pasta with tomato sauce
Pizza	Pizza	Pizza
Tacos	Tacos	Tacos
Chilli	Chilli	Chilli
Pot pies	Pot pies	Pot pies
Soup	Soup	Soup
Fast food cheeseburger, hamburger, French fries, fried chicken, fried fish	Fast foods	Fast foods
Decaffeinated coffee	Decaffeinated coffee	Decaffeinated coffee
Regular coffee	Regular coffee	Regular coffee
Hot tea	Hot tea	Hot tea
Iced tea	Iced tea	Iced tea
Hot cocoa	Hot cocoa	Hot cocoa
Regular soft drinks	Regular soft drinks	Regular soft drinks
Diet soft drinks	Diet soft drinks	Diet soft drinks
Beer	Beer	Beer
Wine	Wine	Wine
Liquor	Liquor	Liquor

fresh fruit and weekly frequency of use of the canned or frozen fruit items. Total weekly frequency of use of each vegetable was calculated as the sum of the weekly frequencies of use of the cooked and raw forms of each item. Foods were not combined further, resulting in 168 individual foods and beverages for principal components analysis.

For Method 2, the 168 individual food items from Method 1 were classified into categories based on nutrient content and usage (Table 1). Based on this categorisation, weekly frequency of use of the following groups was considered in principal components analysis: salad vegetables, cruciferous vegetables, high-carotenoid vegetables, other vegetables, fresh fruit, canned or frozen fruit, dried fruit, fruit juices, high-fat dairy products, low-fat dairy products, refined grains, whole grains, snacks, desserts, candy, nuts, red meat, pork, processed meats, fish, poultry, soup, and fast foods, calculated as the sum of the weekly frequency of use of the individual food items included in each group. Several foods and beverages were not categorised as they were thought to reflect specific dietary behaviours (coleslaw, pickles, mashed potatoes, baked potatoes, fried potatoes, other potatoes, baked beans, tomato juice, margarine, butter, mayonnaise and miracle whip, salad dressing, reduced-calorie mayonnaise and salad dressing, yoghurt, eggs, jams or jellies, liver, macaroni and cheese, spaghetti or lasagne, pizza, tacos, chilli, pot pies, and non-dairy beverages). These foods were included in the principal components analysis as separate items, resulting in 56 groups and foods for Method 2.

Finally, the groups and food items were further categorised into broader groups based on USDA food classification definitions and nutrient content (Method 3). The following groups were considered for determination of dietary patterns: vegetables, fruit, regular fats, reduced-calorie fats, high-fat dairy, low-fat dairy (including yoghurt), eggs, refined grains, whole grains, snacks, desserts, jams or jellies, candy, nuts, meat, processed meats, fish, poultry, and fast foods. As in Methods 1 and 2, weekly frequency of use of each broad group was calculated by summing the weekly frequencies of the member groups or foods. Macaroni and cheese, spaghetti or lasagne, pizza, tacos, chilli, pot pies, soup, and non-dairy beverages were left as separate items for a total of 36 groups and foods.

Dietary patterns

For each method, dietary patterns were identified by principal components analysis (PCA) in SPSS for Windows, version 8, using standard statistical methods[10]. Factors were rotated with an orthogonal (varimax) rotation to improve interpretability and minimise the correlation between the factors. The number of factors retained from each food classification method was determined by factor interpretability and the amount of variance explained by each factor. Labelling of the factors

https://doi.org/10.1079/PHN2001168 Published online by Cambridge University Press

was primarily descriptive and based on our interpretation of the pattern structures.

Cases and controls were assigned pattern-specific factor scores for each of the three food classification methods. Scores for each pattern and method were calculated as the sum of the products of the factor loading coefficient and standardised weekly frequency of use of each food associated with that pattern. Only foods with factor loadings of ≥ 0.30 and ≤ -0.20 were included in calculation of pattern scores because these items represent the foods most strongly related to the identified factor.

Estimation of risk of endometrial cancer

For descriptive purposes, we computed Pearson correlation coefficients for pattern scores by each classification method with selected subject characteristics and nutrient intake. For estimation of risk of endometrial cancer associated with dietary patterns, pattern scores were categorised into tertiles based on the distribution of the controls. For each classification method, risk of endometrial cancer in each tertile relative to the lowest (referent) tertile of pattern scores was estimated by odds ratios (ORs) and 95% confidence intervals (CIs), calculated with unconditional logistic regression adjusting for age (years), education (years), diabetes (yes/no), hypertension (yes/no), cigarette smoking (pack-years), body mass index (weight in kilograms/(height in metres)^2), age at menarche (years), parity (number of pregnancies), oral contraceptive use (ever/never), menopause status (pre- or post-menopausal), and menopausal oestrogen use (ever/never). Risk estimates obtained for each of the patterns derived using the three food classification methods were compared by examination of the impact on the confidence limits associated with each pattern-specific odds ratio. As is commonly done in studies of diet and disease, to estimate the relative risk of endometrial cancer with pattern scores, we had categorised the scores for the cancer subjects into tertiles of pattern score for each pattern and method. If the three classification schemes used in the PCA to identify the patterns produced relatively comparable rankings on pattern scores, we would expect the scores for the subjects to be consistently classified into low, medium and high categories of pattern scores regardless of classification scheme. To assess the degree of concordance in assignment to category of pattern score that might result from the use of different food classification schemes, agreement between the three methods for each pattern was calculated as the percentage exact agreement of classification along the diagonal.

Results

The characteristics of the women in this study have been described in detail elswhere[13]. Briefly, cases tended to have higher body mass index, be more likely to report a history of hypertension and diabetes, be post-menopausal, and have used menopausal oestrogen. Controls were younger and somewhat better educated than were the cases. Controls were also more likely than cases to have used oral contraceptives, and have slightly higher parity.

Dietary patterns

Principal components analyses identified two major dietary patterns from each of the three food classification schemes. For each method, we identified a 'healthy' pattern and a 'high fat' pattern. The factor loadings for the foods associated with each pattern and method are shown in Table 2. The higher the factor loading for a food, the stronger the association of that food with that pattern. Negative factor loadings indicate that non-use of a food was associated with the pattern. As can be seen in Table 2, higher consumption of fruits and vegetables, poultry, fish and whole grains tended to characterise the 'healthy' pattern, whereas the 'high fat' pattern was more strongly associated with higher consumption of refined grains, fast foods, high-fat mixed dishes and meats. Although a number of minor patterns were identified for each method, the additional variance in frequency of food use explained by each was small (less than 1%) and the patterns were less stable and not interpretable within current conceptual frameworks. As can be seen in Table 2, the classification method used had little impact on subsequent identification of dietary patterns. However, as foods were more broadly classified, thus reducing the number of items included in PCA, the amount of variance explained by the resultant dietary patterns increased from around 8% (168 items) to around 17% (36 items).

Pearson's correlation coefficients for the dietary patterns with selected subject characteristics and nutrient intake are shown in Table 3. Among the subject characteristics, age was negatively associated with scores on the 'high fat' pattern, but not related to the 'healthy' pattern. Dietary pattern score was unrelated to other subject characteristics. For all three methods, the 'healthy' pattern was most strongly associated with higher intakes of dietary fibre, folate, vitamin C, phytosterols and individual carotenoids, whereas the 'high fat' pattern was most strongly associated with higher intakes of energy, protein, carbohydrates, fat and cholesterol.

Endometrial cancer risk

Risk of endometrial cancer associated with dietary patterns is shown in Table 4. For the highest tertile of scores on the 'healthy' patterns identified using 168 and 56 items, respectively, we observed reduced risks of endometrial cancer with odds ratios of 0.55 and 0.59. The reduced risks observed for the first two identified 'healthy' patterns were somewhat attenuated and the confidence limits broader for the 'healthy' pattern obtained using the 36 broader food groups (OR 0.77, 95% CI 0.49–1.20). Risk

https://doi.org/10.1079/PHN2001168 Published online by Cambridge University Press

Table 2 Factor loadings for foods associated with dietary patterns identified with principal components analysis using three food-use classification schemes, Western New York Diet Study, female controls ($n = 863$)

168 items				56 items				36 items			
'Healthy'		'High fat'		'Healthy'		'High fat'		'Healthy'		'High fat'	
Carrots	0.48	White rolls	0.46	Salad	0.60	Processed meats	0.61	Vegetables	0.62	Processed meats	0.62
Pineapple	0.45	Fast food fries	0.45	Fresh fruit	0.56	Red meat	0.53	Fruit	0.61	Meat	0.57
Broccoli	0.42	Potato chips	0.45	Other vegetables	0.56	Pork	0.51	Whole grains	0.59	Pizza	0.55
Spinach	0.41	Pizza	0.45	Whole grains	0.54	Fried potatoes	0.50	Low-fat dairy	0.53	Spaghetti or lasagne	0.55
Green peppers	0.40	Spaghetti or lasagne	0.43	High-carotenoid vegetables	0.53	Spaghetti or lasagne	0.48	Refined grains	0.45	Snacks	0.46
Cottage cheese	0.38	Fried potatoes	0.42	Cruciferous vegetables	0.48	Pizza	0.47	High-fat dairy	0.43	Refined grains	0.45
Summer squash	0.38	Hot dogs	0.41	Canned fruit	0.48	Refined grains	0.45	Desserts	0.38	Fast foods	0.43
Grapefruit	0.38	Pepperoni	0.40	High-fat dairy	0.39	Snacks	0.44	Fish or seafood	0.35	Desserts	0.41
Peaches	0.38	Gravy	0.40	Low-fat dairy	0.39	Desserts	0.44	Nuts	0.35	Macaroni and cheese	0.36
Pears	0.38	Ground beef	0.39	Fish and seafood	0.37	Fast food	0.43	Low-calorie fats	0.34	Candy	0.30
Cauliflower	0.37	Other beef	0.39	White potatoes	0.33	Macaroni and cheese	0.38	Jams and jelly	0.32	Tacos	0.29
Prunes or plums	0.37	Doughnuts, pastries	0.38	Baked potatoes	0.31	Mashed potatoes	0.36	Eggs	0.31	Low-calorie fats	−0.20
Mushrooms	0.37	Hamburger	0.37	Coleslaw	0.30	Candy	0.33				
Beets	0.34	Chicken wings	0.36	Refined grains	0.40	Low-calorie salad dressing	−0.23				
Grapes	0.34	Pork chops	0.36	Desserts	0.31						
Asparagus	0.33	Bacon	0.36								
Berries	0.33	Bologna	0.36								
Melon	0.32	Cake	0.36								
Fish, not fried	0.32	Macaroni and cheese	0.36								
Apples or apple sauce	0.31	Brownies	0.35								
Red peppers	0.30	Fast food cheeseburger	0.34								
Sweet potatoes	0.30	Ham	0.34								
Cherries	0.30	Sausage	0.33								
White potatoes	0.30	Mashed potatoes	0.33								
Raisins	−0.23	Cold cuts	0.30								
		Corn	0.30								
		Candy bars	0.30								
Percentage variance explained	3.9		3.8		7.0		6.4		8.4		8.5

A negative loading indicates non-use of a food.
Absolute values <0.30 and >-0.20 were excluded from the table for simplicity and interpretability.

was not associated with the any of the 'high fat' patterns in these data.

Percentage exact agreement for the tertiles of dietary pattern scores by the three food classification schemes is shown in Table 5. For the both the 'healthy' and 'high fat' patterns, exact agreement in tertile classification decreases as the difference in the number of items used for PCA increases. In fact, for the 'healthy' pattern, almost half the subjects were misclassified on pattern score by the broader food-use classification method (36 items) compared with the more detailed method (168 items). Although concordance in classification decreased similarly for the 'high fat' patterns, the effect was less dramatic, with percentage exact agreement decreasing from 81% (168 items vs. 56 items) to 76% (168 items vs. 36 items).

Discussion

The characterisation of dietary intake has long been a challenge for nutritional epidemiology. Earlier studies

https://doi.org/10.1079/PHN2001168 Published online by Cambridge University Press

Table 3 Pearson correlation coefficients for dietary pattern scores calculated by three food-use classification schemes with nutrient intake and non-dietary characteristics, controls ($n = 639$)

	'Healthy'			'High fat'		
	168 items	56 items	36 items	168 items	56 items	36 items
Age	0.09	0.06	−0.04	−0.30	−0.26	−0.30
Education	0.04	0.02	0.08	−0.05	−0.03	0.00
Smoking history	−0.16	−0.12	−0.13	0.11	0.07	0.05
Parity	−0.02	0.03	0.04	0.10	0.06	0.07
Body mass index*	−0.03	−0.02	0.01	0.11	0.09	0.10
Energy	0.18	0.27	0.37	0.59	0.32	0.63
Protein	0.21	0.31	0.38	0.52	0.54	0.58
Carbohydrates	0.27	0.33	0.40	0.51	0.56	0.56
Total fat	0.06	0.17	0.29	0.61	0.63	0.64
Saturated fat	0.05	0.17	0.28	0.60	0.61	0.62
Monounsaturated fat	0.06	0.17	0.27	0.57	0.57	0.58
Polyunsaturated fat	0.11	0.21	0.27	0.45	0.48	0.49
Cholesterol	0.09	0.18	0.26	0.50	0.50	0.52
Dietary fibre	0.48	0.49	0.45	0.28	0.34	0.37
Folate	0.46	0.48	0.44	0.21	0.27	0.30
Phytosterols	0.34	0.35	0.33	0.11	0.18	0.20
Vitamin E	0.27	0.33	0.38	0.34	0.39	0.42
Vitamin C	0.53	0.49	0.37	0.07	0.12	0.14
Retinol	0.16	0.21	0.23	0.24	0.28	0.29
Alpha carotene	0.34	0.31	0.25	0.03	0.06	0.08
Beta carotene	0.47	0.42	0.29	−0.02	0.02	0.06
Cryptoxanthin	0.29	0.26	0.21	−0.04	0.01	0.02
Lycopene	0.22	0.23	0.18	0.14	0.14	0.17
Lutein	0.37	0.32	0.18	−0.04	−0.02	0.02

* Body mass index: kg m^{-2}.

Table 4 Odds ratios (ORs) and 95% confidence intervals (CIs) for risk of endometrial cancer with dietary pattern scores using three food-use classification schemes

Pattern score	OR (95% CI)	*P*-value
'Healthy'		
168 items		
1	1.00	0.004
2	0.70 (0.46, 1.06)	
3	0.55 (0.35, 0.84)	
56 items		
1	1.00	0.03
2	0.89 (0.59, 1.35)	
3	0.59 (0.37, 0.92)	
36 items		
1	1.00	0.52
2	0.72 (0.47, 1.10)	
3	0.77 (0.49, 1.20)	
'High fat'		
168 items		
1	1.00	0.82
2	0.83 (0.55, 1.25)	
3	0.95 (0.57, 1.58)	
56 items		
1	1.00	0.68
2	0.99 (0.66, 1.49)	
3	0.83 (0.49, 1.38)	
36 items		
1	1.00	0.41
2	0.77 (0.51, 1.16)	
3	0.85 (0.51, 1.40)	

Adjusted for energy, age, education, body mass index, diabetes, hypertension, pack-years cigarette smoking, age at menarche, menopause status, parity, oral contraceptive use, and menopausal oestrogen use.

attempted to link intakes of single or multiple nutrients with risk of disease. However, this approach may neglect food components that have yet to be identified but contribute to risk. Examination of individual food use only partially addressed this limitation, as foods are usually consumed in specific combinations or contain multiple nutrients acting in different ways *vis-à-vis* risk. Furthermore, dietary intake may be associated with non-dietary risk factors that might not necessarily be captured with commonly used diet characterisation methods.

A previous criticism of the use of principal components analysis to derive dietary patterns has been the subjective nature of food-use classification prior to PCA. Furthermore, the goal of principal components analysis should be to reduce a large amount of detail into a smaller set of interpretable factors which explain the largest amount of variance in the targeted behaviour. We found that, although similar patterns could be obtained regardless

Table 5 Agreement for tertile of dietary pattern scores identified using three different food-use classification schemes

	Agreement (%)
'Healthy' patterns	
168 items vs. 56 items	73
56 items vs. 36 items	69
168 items vs. 36 items	59
'High fat' patterns	
168 items vs. 56 items	81
56 items vs. 36 items	85
168 items vs. 36 items	76

https://doi.org/10.1079/PHN2001168 Published online by Cambridge University Press

of how food use was categorised before PCA, the classification scheme used did affect the amount of variance in food use explained. Our results indicated that as the level of detail in the items included in the PCA decreased, the variability explained by the factors increased. A possible explanation may be that as foods are more broadly classified, foods weakly associated with a pattern may be classified in the same broad category as foods more strongly associated, thus increasing the amount of information that a specific pattern might capture. Finally, given that individuals are unlikely to limit food choices to one pattern exclusively, for studies of dietary patterns and disease we may be more prudent to rely on interpretability of the factors, rather than variance explained.

On the other hand, increasing the amount of variance explained did not appear to improve estimates of risk of endometrial cancer associated with PCA-derived dietary patterns. In fact, whereas we observed reduced risks with the 'healthy' pattern generated from the 168 and 56 separate food items, the odds ratios observed for this pattern generated from the 36 more broadly classified foods became attenuated and the confidence limits became wider. Of interest, however, is that the level of detail included in the PCA affected the risk estimates associated with the healthy patterns, but not the high fat patterns. We had previously reported reduced endometrial cancer risks associated with higher vegetable intakes and nutrients associated with fruit and vegetable intakes, foods and nutrients associated with the healthy patterns. On the other hand, we found no increased risk associated with foods or nutrients associated with the high-fat patterns[13]. Our results suggest that, in a multidimensional exposure such as diet, greater detail may be necessary to adequately capture differences in dietary exposure between diseased and non-diseased subjects, at least when there is a true diet–disease relationship.

Our study is the first, to our knowledge, to examine risk of endometrial cancer with PCA-derived dietary patterns. In previous analyses of these data, we found reduced risks of endometrial cancer among women in the highest quartile of total vegetable intake (OR 0.49, 95% CI 0.28–0.88)[13]. As can be seen in Table 2, the 'healthy' patterns were more strongly associated with vegetable and fruit intake, and therefore may be simply another way to describe the same pattern of dietary behaviour. Previous studies of dietary patterns and colon cancer suggest that the patterns capture non-dietary risk factors as well, thereby providing a more accurate description of risk[7,8]. As can be seen in Table 3, this was not the case in these data as none of the identified patterns were strongly related to any of the non-dietary subject characteristics examined. Western New York has a number of strong ethnic communities, some of whom have well-established, traditional eating patterns. It is possible that, in this sample, non-dietary characteristics have less impact on individual dietary patterns than do tradition and habit.

In conclusion, our results suggest that PCA can reliably produce descriptions of dietary behaviour in a variety of populations as we obtained patterns similar to other investigations of dietary patterns and disease[7–9]. We have also shown that similar dietary patterns can be described using PCA regardless of the manner in which food use was classified. However, for estimation of disease risk associated with dietary patterns, our results imply that greater detail in food-use information may be desirable in determination of dietary patterns for more precise estimates of disease risk.

References

1 Randall E, Marshall JR, Graham S, Brasure J. Frequency of food use data and the multidimensionality of diet. *J. Am. Diet. Assoc.* 1989; **89**: 1070–5.

2 Randall E, Marshall JR, Graham S, Brasure J. High risk health behaviors associated with various dietary patterns. *Nutr. Cancer* 1991; **16**: 135–51.

3 Ursin G, Ziegler RG, Subar AF, Graubard BI, Haile RW, Hoover R. Dietary patterns associated with a low-fat diet in the national health examination follow-up study: identification of potential confounders for epidemiologic analyses. *Am. J. Epidemiol.* 1993; **137**(8): 916–27.

4 Randall E, Marshall JR, Graham S, Brasure J. Patterns in food use and their associations with nutrient intakes. *Am. J. Clin. Nutr.* 1990; **52**: 739–45.

5 Randall E, Marshall JR, Brasure J, Graham S. Patterns in food use and compliance with NCI dietary guidelines. *Nutr. Cancer* 1991; **15**: 141–56.

6 Hu FB, Rimm E, Smith-Warner SA, Feskanich D, Stampfer MJ, Ascherio A, Sampson L, Willett WC. Reproducibility and validity of dietary patterns assessed with a food-frequency questionnaire. *Am. J. Clin. Nutr.* 1999; **69**: 243–9.

7 Randall E, Marshall JR, Brasure J, Graham S. Dietary patterns and colon cancer in western New York. *Nutr. Cancer* 1992; **18**: 265–76.

8 Slattery ML, Boucher KM, Caan BJ, Potter JD, Ma KN. Eating patterns and risk of colon cancer. *Am. J. Epidemiol.* 1998; **148**(1): 4–16.

9 Hu FB, Rimm EB, Stampfer MJ, Ascherio A, Spiegelman D, Willett WC. A prospective study of major dietary patterns and risk of coronary heart disease in men. *Am. J. Clin. Nutr.* 2000; **72**(4): 912–21.

10 Johnson DE. Principal components analysis. *Applied Multivariate Methods for Data Analysts*. Pacific Grove, CA: Duxberry Press, 1998; chapter 5.

11 Martinez ME, Marshall JR, Sechrest L. Invited commentary: factor analysis and the search for objectivity. *Am. J. Epidemiol.* 1998; **148**(1): 17–9.

12 Slattery ML, Boucher KM. The senior authors' response: factor analysis as a tool for evaluating eating patterns. *Am. J. Epidemiol.* 1998; **148**(1): 20–1.

13 McCann SE, Freudenheim JL, Marshall JR, Vena JE, Laughlin R, Brasure JR, Swanson MK, Graham S. Diet in the epidemiology of endometrial cancer in Western New York. *Cancer Causes Control* 2000; **11**(10): 965–74.

14 US Department of Agriculture. *Composition of Foods*. Agriculture Handbook 8, revisions 8.1–8.17, 8.21. Washington, DC: US Government Printing Office, 1976–1989.

15 US Department of Agriculture. *Nutritive Value of American Foods in Common Units*. Agriculture Handbook 456. Washington, DC: US Government Printing Office, 1975.

16 Mangels AR, Holden JM, Beecher GR, Forman MR, Lanza E. Carotenoid content of fruits and vegetables: an evaluation of analytic data [published erratum appears in *J. Am. Diet. Assoc.* 1993; **93**: 527]. *J. Am. Diet. Assoc.* 1993; **93**: 284–96.

17 Chug-Ahuja JK, Holden JM, Forman MR, Mangels AR, Beecher GR, Lanza E. The development and application of a carotenoid database for fruits, vegetables, and selected multicomponent foods. *J. Am. Diet. Assoc.* 1993; **93**: 318–23.

https://doi.org/10.1079/PHN2001168 Published online by Cambridge University Press