

# Food patterns defined by cluster analysis and their utility as dietary exposure variables: a report from the Malmö Diet and Cancer Study

Elisabet Wirfält<sup>1,\*</sup>, Irene Mattisson<sup>1</sup>, Bo Gullberg<sup>2</sup> and Göran Berglund<sup>1</sup>

<sup>1</sup>Department of Medicine, Surgery and Orthopaedics, Lund University, SE-20502: Malmö, Sweden

<sup>2</sup>Department of Community Medicine, Lund University, SE-20502: Malmö, Sweden

Submitted 23 September 1999; Accepted 16 November 1999

## Abstract

**Objective:** To explore the utility of cluster analysis in defining complex dietary exposures, separately with two types of variables.

**Design:** A modified diet history method, combining a 7-day menu book and a 168-item questionnaire, assessed dietary habits. A standardized questionnaire collected information on sociodemographics, lifestyle and health history. Anthropometric information was obtained through direct measurements. The dietary information was collapsed into 43 generic food groups, and converted into variables indicating the per cent contribution of specific food groups to total energy intake. Food patterns were identified by the QUICK CLUSTER procedure in SPSS, in two separate analytical steps using unstandardized and standardized (Z-scores) clustering variables.

**Setting:** The Malmö Diet and Cancer (MDC) Study, a prospective study in the third largest city of Sweden, with baseline examinations from March 1991 to October 1996.

**Subjects:** A random sample of 2206 men and 3151 women from the MDC cohort ( $n = 28\,098$ ).

**Results:** Both variable types produced conceptually well separated clusters, confirmed with discriminant analysis. 'Healthy' and 'less healthy' food patterns were also identified with both types of variables. However, nutrient intake differences across clusters were greater, and the distribution of the number of individuals more even, with the unstandardized variables. Logistic regression indicated higher risks of past food habit change, underreporting of energy and higher body mass index (BMI) for individuals falling into 'healthy' food pattern clusters.

**Conclusions:** The utility in discriminating dietary exposures appears greater for unstandardized food group variables. Future studies on diet and cancer need to recognize the confounding factors associated with 'healthy' food patterns.

**Keywords**  
Dietary patterns  
Food patterns  
Dietary exposure categories  
Nutrient density  
Standardization  
Z-scores  
Diet history  
Epidemiology

Foods are consumed in a number of combinations, providing a range of nutrients and other dietary factors, which interact in very complex ways. Credible hypotheses have been formulated linking diet to cancer, either through promotion of, or protection against, cancer development<sup>1–4</sup>. However, nutrition studies have difficulties in separating the effects of individual nutrients, because food sources often are the same for many nutrients (e.g. for energy and fat, for different types of fat and for plant food constituents) resulting in highly correlated variables<sup>3–5</sup>. Strong correlations between variables (multicollinearity) result in unstable relationships, and may cause attenuation of diet–disease relations if the variables are entered simultaneously into the same model. Reviewers have pointed out that it is difficult both on practical and theoretical grounds to isolate the specific cancer-related biological activities of single nutrients or chemicals in observational epidemiology studies, and that

integrative approaches taking into account multiple dietary factors are needed<sup>3,4</sup>. Multivariate methodologies to define patterns (like cluster or factor analysis) will turn the collinearity of dietary variables into an analytical advantage. Although these methodologies are available, only a few studies have used them to examine the combined effect of many dietary factors on chronic disease risk<sup>6–10</sup>.

Cluster analysis empirically identifies patterns in complex data by grouping individuals with similar characteristics<sup>11,12</sup>, producing homogenous and non-overlapping exposure categories. The major advantage of this methodology is its simplicity and ease of interpretation<sup>11</sup>. Since the emerging clusters are influenced by the particular variables entered into the clustering procedure<sup>12</sup>, nutrition studies using clustering variables of different units have commonly standardized variables to a mean of zero and a standard deviation of one

\* Corresponding author. Email [elisabet.wirfalt@smi.mas.lu.se](mailto:elisabet.wirfalt@smi.mas.lu.se)

(i.e. Z-scores)<sup>7,8,13,14</sup>. Several studies have adopted a nutrient density approach, and have defined dietary variables as the per cent contribution of macronutrients to total energy<sup>7</sup>, or of specific foods to the total amount of energy consumed<sup>10,15–17</sup>. It can be argued that nutrient density variables are standardized and that no further transformation is necessary. It is, however, not clear how transformation of clustering variables affect the formation of dietary patterns.

The main objective of this study was to identify food patterns of total diet in an urban population using cluster analysis, and to examine whether unstandardized per cent energy food group variables and Z-scores of these variables result in food pattern clusters of similar utility. This paper examines: (i) the cluster analysis procedure and emerging food patterns; (ii) the differences across food pattern clusters in total energy, macro- and micro-nutrients and obesity indicators separately for each type of variable, and as a measure of external validity the ratio between total energy intake and basal metabolic rate; and (iii) the stability of food patterns using discriminant analysis. Finally, the study also examines the risk of falling into each specific food pattern cluster in relation to factors that potentially could distort diet–disease relations in studies using self-reported dietary intakes. The examined factors were past food habit change, indicators of obesity and energy intake relative to basal metabolic rate.

## Methods

### Population

The MDC Study, which is a population-based prospective cohort study in the third largest city of Sweden<sup>18</sup>, conducted baseline examinations from March 1991 to October 1996. All men aged 50–73 years and all women aged 45–73 years living in Malmö were invited to join the study. A total of 28 098 individuals joined the study, either spontaneously or after receiving a posted invitation. During the period November 1991 to February 1994 a random half of those joining the study ( $n=6103$ ) were invited to participate in a substudy on cardiovascular disease and for an additional visit to donate fasting blood ( $n=5533$ ). Of these a total of 5357 individuals (2206 men and 3151 women) also had complete dietary data and constitute the study population examined in this paper. The MDC cohort is very similar to the background population, but the proportions of individuals with higher education and of white collar workers in higher positions are slightly higher. Current smokers and individuals of non-Swedish origin are slightly less common. This sub-sample (including only individuals with complete dietary data) differs slightly from the MDC cohort, with smaller proportions of males, white collar workers in higher positions, divorcees and individuals of non-Swedish origin.

### Data collection

All MDC participants visited the study centre twice. At the first visit trained project staff provided groups of participants with information on the background and aims of the project, gave detailed instructions about the dietary data collection procedure and study questionnaires, and conducted direct anthropometric measurements. At the second visit (approximately 2 weeks later) individual interviews were conducted by trained dietary interviewers to complete the diet history and to check the correctness of completed questionnaires.

### Study measures

Dietary data were collected through a modified diet history method, specifically developed for the MDC Study to assess total diet within an urban population<sup>19</sup>. A 7-day menu book collected information on cooked lunch and dinner meals and cold beverages (including alcoholic beverages), and a 168-item questionnaire was used to obtain information on the frequencies of regularly consumed foods, including hot beverages, cakes, sandwiches, edible fats, breakfast cereals, yoghurt, milk, fruits, candies and snacks during the past year. The usual portion sizes reported in the questionnaire were estimated by the participant at home using a booklet with 48 black and white photographs. A total of 13 diet assistants conducted diet history interviews among the participants examined in this paper. During the interviews the menu book and questionnaire were carefully checked to make sure that diet reports did not overlap. Also, the portion sizes recorded in the menu book were estimated from a separate, more extensive, book of photos. A continuous in-service training programme, specific coding rules and interactive computer software (Kostvar, AIVO AB, Stockholm) enabled standardized coding and entering of dietary data, and facilitated homogeneity in data collection across diet assistants. The specific food information was converted to nutrient intake data by calculating average individual daily intakes using the nutrient information available in the MDC food and nutrient database. This database, specifically developed for the MDC Study, originates from the PC KOSTversion2/93 of the Swedish National Food Administration. It contains information on approximately 1600 basic foods derived from chemical analyses at the Swedish National Food Administration, the Swedish food industry or from other, mostly American, sources.

The reproducibility and concurrent validity of the diet history method has previously been examined in the Malmö Food Study with 18 days of weighed food records collected during 1 year as the reference<sup>20–22</sup>. This validation study included 241 Malmö residents (126 men and 115 women) in the age range 50–69 years. The method overestimated total energy by 18% at group level compared to the reference method. The nutrient correlations between the diet history and reference method were

generally higher than those found in validation studies of comparable dietary methods in other populations<sup>23</sup>.

The *food group variables* of this study were created to express the per cent of total energy contributed by specific food groups ('density'), and to facilitate the distinction between energy sources that are high, or low, in fat and fibre. First, all available dietary information was converted from grams consumed per day of the specific foods to the per cent of total energy contributed by the specific foods. Second, food variables were aggregated into 48 food groups. Third, food groups contributing negligible amounts of energy with a high number of non-consumers were identified and excluded from further analysis (e.g. mineral water, spices and condiments, diet soft drinks, vegetable juice, very high-fibre and high-sugar cereals). The resulting 43 food group variables were used to define food patterns in this data set.

The *nutrient variables* examined were: total energy (MJ), per cent contribution of macronutrients to non-alcohol energy (%energy), total fat (g), the ratio of polyunsaturated and saturated fatty acids (P/S), the ratio of  $\omega$ -3 and  $\omega$ -6 fatty acids ( $\omega$ 3/ $\omega$ 6), alcohol (g), fibre (g), ascorbic acid (mg),  $\beta$ -carotene (mg), folic acid ( $\mu$ g), vitamin E (mg), iron (mg), calcium (mg), selenium ( $\mu$ g) and zinc (mg).

Some *diet assistants* conducted interviews all through the baseline examination period, while others worked for shorter time periods. Although great care was taken to allocate participants uniformly, it is possible that diet assistants were not equally distributed over different population subgroups, and that differences in interviewing experience could have influenced observed dietary intakes. Diet assistant was therefore included as a controlling variable in the final multivariate analysis.

By combining a questionnaire and a menu book the dietary history of the MDC Study incorporates 'current diet' aspects into assessment, and may therefore show some seasonal variation in observed intakes. Therefore, *season of data collection* was controlled for in the multivariate analysis. A four-category variable delineating the four seasons – winter (December–February), spring (March–May), summery (June–August) and autumn (September–November) – was defined by the date of the dietary interview.

Information on *age* and *gender* was obtained through the person-identification number. A standardized questionnaire collected information on a broad range of issues related to demography, socioeconomics and lifestyle. Information on *past food habit change* was based on the questionnaire item: 'Have you, because of ill health or other reasons, substantially changed your food habits?' with yes/no response categories.

*Underreporting of energy* is an ongoing controversy in studies using self-report instruments to collect dietary information in epidemiology studies<sup>24,25</sup>. In order to examine the prevalence of low energy reporting in this

study the ratio between total energy intake (EI) and basal metabolic rate (BMR) was computed (i.e. EI/BMR) and compared across food pattern clusters. The equations recommended by WHO was used to compute BMR, separately for men and women and the two age intervals: 30–60 years and above 60 years<sup>26</sup>. A ratio of 1.35 has been suggested as the lower cut-off point for reasonable habitual energy intakes<sup>24</sup>. The EI/BMR variable was dichotomized ( $\leq 1.35$  and  $> 1.35$ ) for multivariate analysis examining the risk of falling into specific clusters.

The *body mass index* ( $\text{kg m}^{-2}$ ), and the *waist girth–hip girth ratio* (WHR) were calculated from direct measurements of weight, height, waist and hip circumferences of participants. Tertiles of the obesity indicator variables were used for multivariate analysis examining the risk of falling into specific clusters.

### Statistical methods

All analyses used the SPSS statistical system<sup>27</sup>. Cluster analysis was performed using the QUICK CLUSTER procedure, which is a K-means method that uses the Euclidean distances between observations to estimate clusters empirically from the data set. If the clustering variables have scales of very different ranges, the variables with larger values will overwhelm those with smaller values. Therefore standardization, or Z-score transformation, is often recommended to make the contribution of all variables to the distance measure more comparable<sup>12,27</sup>. This study evaluated the outcome of cluster analysis when using either: (i) unstandardized per cent energy food variables; or (ii) Z-scores of per cent energy food variables (i.e. transformed to have a mean of zero and a standard deviation of one). The cluster analysis procedure was therefore performed in two steps separately for each type of food group variable.

When using the QUICK CLUSTER procedure the number of clusters needs to be preselected. Since no information was available on the appropriate number of clusters in the data set, a series of steps were taken separately for each variable type to select the most suitable number. Firstly, several runs of cluster analysis were conducted with a varying number of clusters (from two to 10). Secondly, the analysis of variance tables of each analysis and the F-statistics of the food group variables were inspected to identify cluster solutions with well separated clusters. Thirdly, the size of the emerging clusters and the differences in food consumptions across individual clusters from each run was examined. With both variable types the six cluster solutions produced reasonably sized and well separated clusters with varied consumptions of major food energy contributors, and was therefore selected for further examinations.

The stability of the selected cluster solutions was tested through discriminant analysis. The discriminant functions that best distinguished subjects into six groups were determined, separately for each variable type, with all 43

**Table 1** Energy contribution from food groups in clusters generated from unstandardized food group variables

Per cent energy from food groups	Many foods and drinks ( <i>n</i> = 1508)	Fibre bread ( <i>n</i> = 510)	Low fat and high fibre ( <i>n</i> = 814)	White bread ( <i>n</i> = 758)	Milk fat ( <i>n</i> = 675)	Sweets and cakes ( <i>n</i> = 1092)
18.0–15.1				White bread		Sweets
15.0–12.1		Fibre bread			Bregott spread	
12.0–9.1			Fruits, low-fat milk			
9.0–6.1	Cheese, high-fat meats	High-fat meats		High-fat meats, sweets, low-fat spread, low-fat meats	Sweets, white bread, high-fat meats	High-fat meats
6.0–3.1	Cake/pastry, fruits, white bread, sweets, low-fat meats, boiled potatoes, medium-fat spread, low-fat milk, whole milk, low-fat spread, fibre bread	Sweets, fruits, low-fat meats, low-fat spread, cheese, boiled potatoes, low-fat milk, cake/pastry, vegetables	Low-fat meats, high-fat meats, sweets, vegetables, boiled potatoes, cake/pastry, fibre bread, white bread, fibre crisp-bread, low-fat spread, cereals, cheese	Cheese, fruits, boiled potatoes, cake/pastry, low-fat milk, whole milk, medium-fat spread	Cheese, whole milk, low-fat meats, cake/pastry, fruits, boiled potatoes, fibre bread	White bread, fruits, cake/pastry, cheese, low-fat meats, whole milk, low-fat milk, low-fat spread, boiled potatoes, fibre bread
3.0–0.0	Vegetables, beer, cereals, wine, chocolates, fibre crisp-bread, rice/pasta, low-fat cake, egg, high-fat fish, low-fat fish, fried potatoes, fruit juice, soft drinks, ice-cream, dressing, crackers, liquor, Bregott spread, cottage cheese, butter, coffee, oil, low-fat dressing, snacks, sherbet, tea, nuts, cream, miscellaneous	White bread, whole milk, medium-fat spread, cereals, rice/pasta, beer, fibre crisp-bread, low-fat cake, cottage cheese, high-fat fish, egg, low-fat fish, wine, chocolates, fruit juice, fried potatoes, Bregott spread, soft drinks, ice-cream, dressing, crackers, liquor, oil, butter, coffee, low-fat dressing, snacks, sherbet, tea, nuts, cream, miscellaneous	Medium-fat spread, cottage cheese, rice/pasta, high-fat fish, low-fat cake, wine, egg, low-fat fish, whole milk, chocolates, beer, fruit juice, crackers, ice-cream, fried potatoes, dressing, soft drinks, coffee, oil, Bregott spread, liquor, butter, low-fat dressing, snacks, sherbet, tea, nuts, cream, miscellaneous	Vegetables, cereals, fibre bread, beer, rice/pasta, low-fat cake, fried potatoes, chocolates, egg, high-fat fish, soft drinks, fibre crisp-bread, cottage cheese, wine, low-fat fish, ice-cream, dressing, fruit juice, crackers, Bregott spread, liquor, butter, coffee, oil, low-fat dressing, snacks, sherbet, tea, nuts, cream, miscellaneous	Low-fat milk, medium-fat spread, beer, vegetables, cereals, chocolates, fibre crisp-bread, wine, rice/pasta, low-fat cake, egg, fried potatoes, high-fat fish, low-fat fish, fruit juice, soft drinks, ice-cream, dressing, crackers, liquor, cottage cheese, coffee, low-fat spread, butter, oil, low-fat dressing, snacks, sherbet, tea, nuts, cream, miscellaneous	Medium-fat spread, vegetables, cereals, chocolates, low-fat cake, rice/pasta, Bregott spread, beer, fibre crisp-bread, egg, high-fat fish, soft drinks, low-fat fish, fruit juice, wine, fried potatoes, ice-cream, crackers, cottage cheese, dressing, butter, liquor, coffee, oil, low-fat dressing, snacks, sherbet, tea, nuts, cream, miscellaneous

**Table 2** Z-scores and food group distribution across clusters generated from standardized food group variables

Z-scores	Drinks and fries (n = 182)	Ice-cream and cake (n = 245)	Dieters (n = 47)	Healthy (n = 1317)	Traditional (n = 2493)	Mediterranean (n = 1073)
8.00–1.01 1.00–0.61	Low-fat dressing	Ice-cream	Sherbet			
0.60–0.21	Liquor, fried potatoes, wine	Chocolates, sherbet	Cottage cheese, fruit, high-fat fish, coffee, low-fat milk, miscellaneous, vegetables, fibre crisp-bread, low-fat spread	Cottage cheese, low-fat milk, low-fat spread, crackers, fibre bread, fruit, fibre crisp-bread, miscellaneous, low-fat cake, boiled potatoes	White bread, sweets, Bregott spread, whole milk	Wine, oil, vegetables, rice/pasta, low-fat fish, fruit, low-fat meats, egg, dressing, fibre crisp-bread, high-fat fish, nuts, tea, cheese
0.20 to –0.19	Beer, dressing, low-fat fish, medium-fat spread, snacks, low-fat meats, nuts, egg, cheese, fruit juice, chocolates, cream, cottage cheese, rice/pasta, soft drinks, white bread, high-fat meats, coffee, crackers, low-fat milk, ice-cream, low-fat spread, cake/pastry, miscellaneous, whole milk, butter, cereals, vegetables, high-fat fish, oil, fibre crisp-bread, sherbet, Bregott spread, boiled potatoes	Cake/pastry, cheese, low-fat cake, whole milk, dressing, fruit, butter, soft drinks, coffee, cream, high-fat fish, fried potatoes, snacks, low-fat meats, fibre crisp-bread, nuts, fruit juice, low-fat fish, rice/pasta, oil, medium-fat spread, high-fat meats, egg, low-fat spread, wine, tea, crackers, low-fat dressing, Bregott spread, cereals, miscellaneous, vegetables, beer, liquor, fibre bread, boiled potatoes, cottage cheese, low-fat milk, white bread	Ice-cream, rice/pasta, low-fat meats, low-fat fish, low-fat dressing, fibre crisp-bread, cereals, chocolates, oil, medium-fat spread, fried potatoes, egg, cream, fibre bread, nuts, cake/pastry, beer, high-fat meats, low-fat cake, boiled potatoes, soft drinks, wine, snacks, fruit juice	Vegetables, cereals, fruit juice, tea, coffee, high-fat fish, high-fat meats, low-fat fish, cake/pastry, egg, sherbet, sweets, low-fat meats, soft drinks, chocolates, low-fat dressing, nuts, rice/pasta, ice-cream, cream, butter, medium-fat spread, oil, snacks	Cheese, soft drinks, cake/pastry, high-fat meats, chocolates, beer, butter, medium-fat spread, fried potatoes, liquor, cream, snacks, low-fat cake, nuts, boiled potatoes, dressing, fruit juice, tea, sherbet, coffee, miscellaneous, low-fat meats, cereals, crackers, high-fat fish, egg, low-fat dressing, ice-cream, wine, low-fat spread, rice/pasta, fibre bread, oil	Coffee, cereals, snacks, liquor, beer, fruit juice, low-fat milk, fibre bread, fried potatoes, cream, cottage cheese, butter, medium-fat spread, miscellaneous, sherbet, low-fat dressing, boiled potatoes, crackers, ice-cream, Bregott spread
–0.20 to –0.59	Tea, fruit, sweets, white bread, low-fat cake	Sweets	White bread, dressing, liquor, butter, tea, Bregott spread, whole milk, sweets, cheese	Dressing, white bread, fried potatoes, liquor, beer, whole milk, wine, Bregott spread, cheese	Low-fat fish, fibre crisp-bread, cottage cheese, low-fat milk, vegetables, fruit	Chocolates, soft drinks, whole milk, low-fat spread, high-fat meats, low-fat cake, cake/pastry, white bread, sweets

food group variables entered simultaneously. The 'hit' tables, describing the relationship between actual and predicted group allocation, were inspected, and the percentage of subjects correctly classified by the clustering procedure determined. Secondly, the discriminant analysis was rerun using the stepwise method to identify the food group variables that contributed significantly to the formation of clusters.

Mean differences in total energy, total fat, the per cent contribution of macronutrients to non-alcohol energy, P/S and  $\omega 3/\omega 6$  ratios, dietary fibre, vitamins and minerals, alcohol consumption, BMI, WHR and EI/BMR were compared across clusters using general factorial analysis of variance, while controlling for age and gender. Comparisons of mean intakes of total fat, P/S and  $\omega 3/\omega 6$  ratios, dietary fibre, vitamins and minerals, and alcohol consumption were also controlled for total energy, using the standard multivariate approach<sup>28</sup>. All continuous variables were log-transformed prior to analysis to normalize the distribution of data. Chi-square analysis examined the relation between cluster membership and past food habit change, diet assistants and season of data collection. Finally, logistic regression analysis was used to examine the risk of falling into each specific cluster compared to all other clusters. The models examined the relation with EI/BMR, BMI, WHR and past food habit change, while controlling for age, gender, season of data collection and diet assistant.

## Results

The two cluster solutions are described in Tables 1 and 2. The clusters generated from the unstandardized variables were labelled by the foods that contributed the highest percentage of total energy, while those from Z-scores were given interpretative labels. The unstandardized per cent energy food variables produced six food pattern clusters with a fairly even distribution of individuals across clusters. The Z-scored variables, however, resulted in a six cluster solution with three large clusters and three clusters of much smaller size. Both variable types produced clusters where 'healthier' food choices (i.e. low-fat and high-fibre foods) and 'less healthy' food choices (i.e. high-fat foods) were prominent features.

The discriminant analysis indicated good stability for the selected cluster solutions of both variable types. The agreement between actual and predicted cluster allocation ranged between 91.0% and 95.2% for the unstandardized variables, and between 91.1% and 100% for the Z-scored variables. When using the stepwise function of the discriminant analysis, 18 unstandardized variables (i.e. Bregott margarine, low-fat margarine, whole milk, rice and pasta, sweets, white bread, high-fat meat, low-fat meat, fibre bread, fruits, fibre crisp-bread, vegetables, cereals, cheese, cottage cheese, low-fat milk, low-fat cake, cream)

and 31 Z-scored variables (i.e. sherbet, low-fat dressing, tea, vegetables, fruits, low-fat fish, fat milk, fibre crisp-bread, egg, low-fat meat, high-fat fish, cereals, coffee, fruit juice, ice-cream, wine, low-fat margarine, cottage cheese, cheese, low-fat milk, oil, crackers, rice and pasta, low-fat cake, beer, miscellaneous, high-fat dressing, fibre bread, nuts, boiled potatoes and fried potatoes) contributed significantly to the predicted cluster allocations.

## Food selection and nutrient intakes

### Unstandardized variables

Individuals falling into the largest cluster ( $n=1508$ ) had an energy intake distributed over many foods, and did not differ much from the total sample average (see also Appendix 1). The other five clusters emerged with one or two food groups contributing a major part of the total energy. However, high-fat meats and sweets were important energy sources in all clusters. Vegetable consumption was highest in the 'low fat and high fibre' and in the 'fibre bread' clusters. In the 'sweets and cakes' cluster ( $n=1092$ ) 18% of total energy came from sugar, candy and marmalade. Individuals in this cluster also consumed comparatively large amounts of cake and soft drinks. In three clusters (i.e. 'milk fat', 'white bread' and 'fibre bread'), sandwich consumption appeared to be a dominating food habit. The 'milk fat' cluster had 12% of total energy coming from Bregott margarine (manufactured from a butter fat-oil mixture), and was comparatively high in cheese and whole milk. White bread and sweets were also high in this cluster. Individuals in the 'white bread' cluster took 16% of total energy from white bread. Low-fat margarine, high-fat meats, sweets and low-fat meats were other major energy sources in this cluster. Individuals of the 'fibre bread' cluster consumed 15% of their energy from fibre bread. Other major energy sources were meats, sweets and fruits; low-fat margarine and boiled potatoes also ranked comparatively high in this cluster. The 'low fat and high fibre' cluster ( $n=814$ ) appeared overall to have the healthiest food choices. Fruits (10.5%) and low-fat milk (8%) were, together with low-fat meats, high-fat meats and sweets, the major energy contributors in this cluster. Fibre-rich and low-fat foods like vegetables, cereals, crisp-bread, cottage cheese, rice, pasta, egg, fish and fruit juice also ranked comparatively high.

All clusters were significantly different from one another on total energy and fat intake (Table 3). The 'low fat and high fibre' cluster had the lowest mean intake of total energy and per cent energy from fat, while the 'milk fat' cluster had the highest energy and fat intakes. The difference in per cent energy from fat was 10.2% between these two clusters. The per cent energy from fat and protein was comparatively low in the 'sweets and cakes' cluster, but was high from carbohydrate. On the other

**Table 3** Nutrient intake differences across food pattern clusters generated from unstandardized food group variables\*

	1 Many foods and drinks (n = 150)	2 Fibre bread (n = 510)	3 Low fat and high fibre (n = 814)	4 White bread (n = 758)	5 Milk fat (n = 675)	6 Sweets and cakes (n = 1092)	F-test (P value)	Multiple comparisons†
Total energy (MJ)‡	9.93	9.51	8.70	10.23	10.90	10.50	0.000	All different
Carbohydrate (%energy)	42.6	48.1	48.2	46.0	40.6	47.7	0.000	5 < 1, 2, 3, 4, 6; 1 < 2, 3, 4, 6; 4, 6 < 3; 4 < 2, 6
Protein (%energy)	16.5	16.4	17.6	16.0	15.0	15.1	0.000	5, 6 < 1, 2, 3, 4; 4 < 1, 2, 3; 1, 2 < 3
Fat (%energy)	40.9	35.5	34.3	38.0	44.5	37.3	0.000	All different
Fat (g)	100.0	87.7	84.3	94.3	109.9	92.5	0.000	All different
P/S	0.40	0.46	0.46	0.46	0.27	0.40	0.000	5 < 1, 2, 3, 4, 6; 1, 6 < 2, 3, 4
$\omega 3/\omega 6$	0.20	0.20	0.22	0.18	0.26	0.20	0.000	4 < 1, 2, 3, 5, 6; 1, 2, 6 < 3, 5; 3 < 5
Alcohol (g)§	10.8	5.8	5.2	6.0	9.2	5.5	0.000	4, 6 < 1, 2, 3, 5; 2, 3, 5 < 1
Fibre (g)	19.6	26.6	25.7	19.8	17.5	19.4	0.000	5 < 1, 2, 3, 4, 6; 1, 4, 6 < 2, 3; 3 < 2
$\beta$ -carotene (mg)	2.93	3.74	4.55	2.34	2.38	2.67	0.000	4, 5 < 1, 2, 3, 6; 6 < 1, 2, 3; 1 < 2, 3; 2 < 3
Ascorbic acid (mg)	98.9	103.4	136.3	80.0	81.7	90.9	0.000	4, 5 < 1, 2, 3, 6; 6 < 1, 2; 1, 6 < 3; 2 < 3
Vitamin E (mg)	11.1	10.6	11.4	10.6	8.92	10.0	0.000	5 < 1, 2, 3, 4, 6; 6 < 1, 2, 3, 4; 2, 4 < 1, 3; 1 < 3
Folic acid ( $\mu$ g)	243	297	300	246	223	233	0.000	5 < 6; 5, 6 < 1, 2, 3, 4; 1, 4 < 2, 3
Calcium (mg)	1121	1114	1320	1054	1004	1085	0.000	5 < 1, 2, 3, 4, 6; 4 < 1, 2, 3, 6; 1, 2, 6 < 3
Iron (mg)	15.4	17.9	16.0	18.3	14.9	15.3	0.000	5 < 1, 2, 3, 4, 6; 1, 6 < 2, 3, 4; 3 < 2, 4; 2 < 4
Selenium ( $\mu$ g)	41.0	39.3	44.8	36.8	36.9	37.4	0.000	4, 5, 6 < 1, 2, 3; 2 < 1, 3; 1 < 3
Zinc (mg)	11.9	12.5	12.6	12.0	11.1	11.1	0.000	5, 6 < 1, 2, 3, 4; 1, 4 < 2, 3

\* Adjusted for age, gender and energy intake; geometric means are shown for energy and all nutrients except alcohol.

†  $\alpha < 0.05$ .

‡ Adjusted for age and gender.

§ Alcohol values are medians.

**Table 4** Nutrient intake differences across food pattern clusters generated from standardized food group variables\*

	1 Drinks and fries ( <i>n</i> = 182)	2 Ice-cream and cake ( <i>n</i> = 245)	3 Dieters ( <i>n</i> = 47)	4 Healthy ( <i>n</i> = 1317)	5 Traditional ( <i>n</i> = 2493)	6 Mediterranean ( <i>n</i> = 1073)	<i>F</i> -test ( <i>P</i> value)	Multiple comparisons†
Total energy (MJ)‡	9.89	10.40	8.90	9.39	10.74	8.93	0.000	3, 4, 6 < 1, 2, 5; 6 < 4; 1 < 2, 5; 2 < 5
Carbohydrate (%energy)	42.8	44.2	47.0	48.0	44.4	44.5	0.000	1 < 2, 3, 4, 5, 6; 2, 5, 6 < 3, 4
Protein (%energy)	16.0	15.7	16.9	16.7	15.2	17.5	0.000	1, 2, 5 < 3, 4, 6; 5 < 1, 2; 4 < 6
Fat (%energy)	41.2	40.1	36.0	35.3	40.4	38.0	0.000	3, 4, 6 < 1, 2, 5; 3, 4 < 6; 2 < 1
Fat (g)	100.5	99.6	88.5	88.1	99.8	91.3	0.000	3, 4, 6 < 1, 2, 5; 4 < 6
P/S	0.50	0.38	0.48	0.46	0.36	0.46	0.000	2, 5 < 1, 3, 4, 6
$\omega$ 3/ $\omega$ 6	0.19	0.20	0.21	0.20	0.21	0.22	0.000	1, 2, 4 < 5, 6
Alcohol (g)§	11.7	5.5	2.6	3.0	8.7	12.0	0.000	3, 4 < 1, 2, 5, 6; 2, 5 < 1, 6
Fibre (g)	18.5	19.6	23.8	24.2	18.4	23.4	0.000	1, 5 < 2, 3, 4, 6; 2 < 3, 4, 6; 6 < 4
$\beta$ -carotene (mg)	2.85	2.88	4.24	3.66	2.32	4.03	0.000	5 < 1, 2, 3, 4, 6; 1, 2 < 3, 4, 6; 6 < 4
Ascorbic acid (mg)	95.7	95.6	114.0	108.0	83.4	120.5	0.000	5 < 1, 2, 3, 4, 6; 1, 2 < 3, 4, 6
Vitamin E (mg)	10.68	10.38	11.14	10.72	9.87	11.74	0.000	5 < 1, 2, 4, 6; 1, 2, 3, 4 < 6
Folic acid ( $\mu$ g)	244	236	261	276	228	288	0.000	1, 2, 5 < 4, 6; 3, 4 < 6; 5 < 3
Calcium (mg)	1057	1134	1294	1250	1044	1129	0.000	1, 5 < 2, 3, 4, 6; 2, 6 < 3, 4
Iron (mg)	15.4	14.8	15.7	16.6	15.7	16.6	0.000	2 < 1, 3, 4, 5, 6; 1, 3, 5 < 4, 6
Selenium ( $\mu$ g)	40.6	38.4	41.8	39.8	36.4	47.0	0.000	5 < 1, 2, 3, 4, 6; 2 < 1, 3; 1, 2, 3, 4 < 6
Zinc (mg)	11.4	11.4	12.3	12.5	11.3	12.3	0.000	1, 2, 5 < 3, 4, 6; 6 < 4

\* Adjusted for age, gender and energy intake; geometric means are shown for energy and all nutrients except alcohol.

†  $\alpha < 0.05$ .

‡ Adjusted for age and gender.

§ Alcohol values are medians.



hand, individuals of the ‘many foods and drinks’ cluster had low intakes of per cent energy from carbohydrate, and high intakes of fat and protein. The ‘milk fat’ cluster had both the lowest P/S ratio, and the highest  $\omega3/\omega6$  ratio. The ‘many food and drinks’ cluster had the highest mean intake of alcohol (medians are presented in Table 3) and the lowest proportion of zero consumers (5.8%), while the ‘white bread’ and the ‘sweets and cakes’ clusters had the lowest mean intakes. The ‘low fat and high fibre’ cluster had, however, the largest proportion of zero consumers (12.7%) and also the smallest median intake.

Individuals of the ‘low fat and high fibre’ cluster had the highest mean intakes for most of the examined micro-nutrients. Folic acid and zinc were high in the clusters ‘low fat and high fibre’ and ‘fibre bread’. The ‘fibre bread’ cluster also had the highest intake of dietary fibre, while the ‘white bread’ cluster was highest for iron. Overall the ‘milk fat’, and to some extent the ‘white bread’, cluster had the lowest micronutrient intakes.

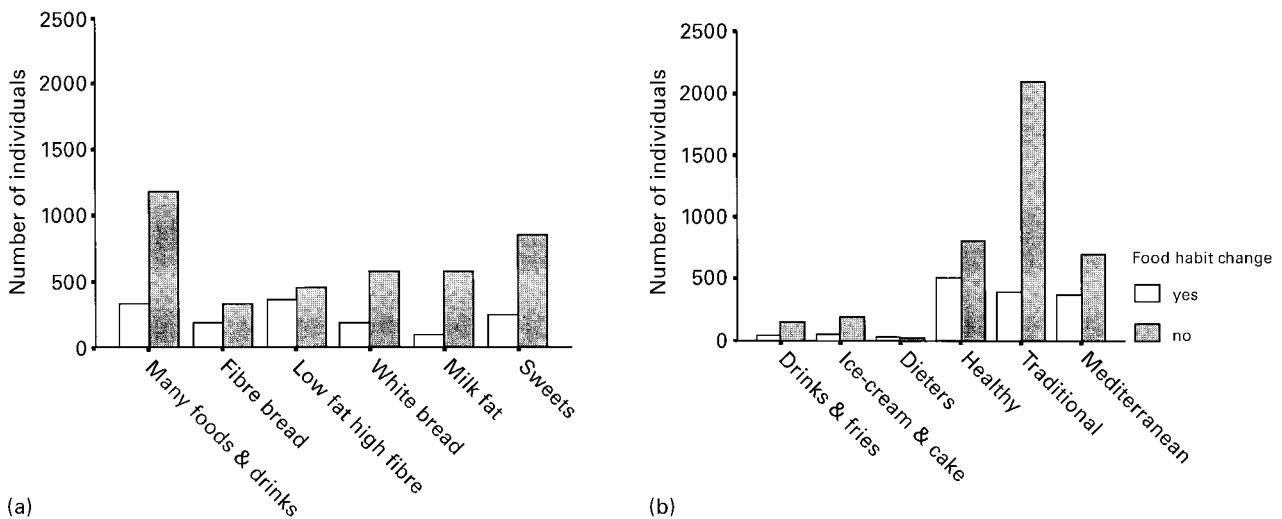
*Z-scored variables*

When transforming per cent energy variables to Z-scores, food groups that contribute very little to total energy intake (e.g. coffee) are given equal weight in analysis to those that contribute large amounts of energy (e.g. sweets). When clusters are formed from Z-scores the emerging cluster may reflect extreme food consumptions (very high or very low), which do not necessarily contribute much to the overall dietary composition (see also Appendix 2). Individuals of the smallest cluster ‘dieters’ ( $n=47$ ) had high Z-scores for low-fat dairy foods, fish, coffee, fruit and vegetables, and low Z-scores for high-fat dairy foods, sweets and tea. One of the larger clusters, ‘healthy’ ( $n=1317$ ), had high scores for low-fat and high-fibre products, but low scores for high-fat dairy

foods and alcoholic drinks. Two clusters emerged with high Z-scores for alcoholic beverages. Individuals of the ‘drinks and fries’ cluster ( $n=182$ ) had high scores for low-fat dressing, hard liquor, fried potatoes, wine and beer. However, individuals of the ‘Mediterranean’ cluster ( $n=1073$ ) ranked highest on wine, oil, vegetables, rice and pasta, and low-fat fish. The largest cluster ‘traditional’ ( $n=2493$ ) was characterized by high scores for white bread, Bregott margarine, sweets and soft drinks, but scores were low for fruits, vegetables and low-fat dairy products. Individuals of the ‘ice-cream and cake’ cluster ( $n=245$ ) ranked high on ice-cream, chocolate, cake and high-fat dairy products, but low on sweets and low-fat dairy foods.

As can be seen from Table 4, the ‘dieter’ cluster showed the lowest mean intake of total energy, while the ‘traditional’ cluster had the highest. Per cent energy from fat, however, was lowest in the ‘healthy’ cluster, and highest in the ‘drinks and fries’ cluster, a difference of 5.9%. The P/S ratio was highest in the ‘drinks and fries’ cluster, while the  $\omega3/\omega6$  ratio was highest in the ‘Mediterranean’ cluster. Alcohol intakes were, as expected, considerably higher in the ‘drinks and fries’ and ‘Mediterranean’ clusters (medians are presented in Table 4), and the proportion of zero consumers was the lowest (6.0% and 4.0%, respectively). The ‘dieters’ and ‘healthy’ clusters had the lowest mean alcohol intakes and the largest proportions of zero consumers (12.8% and 15.4%, respectively).

Overall three clusters of the Z-scored variables emerged with higher micronutrient intakes compared to the rest. Dietary fibre,  $\beta$ -carotene, ascorbic acid and zinc were highest in the ‘dieters’, ‘healthy’ and ‘Mediterranean’ clusters. Vitamin E and selenium were highest in the ‘Mediterranean’ cluster, while iron intakes were high in both the ‘healthy’ and ‘Mediterranean’ clusters.



**Fig. 1** Distribution of individuals reporting ‘past food habit change’, displayed separately for clusters generated from (a) unstandardized and (b) standardized per cent energy food variables

**Table 5** The risk of falling into each food pattern cluster (generated from unstandardized variables) in relation to some potential confounders\*

	Total	Many foods and drinks (n = 1508)		Fibre bread (n = 510)		Low fat and high fibre (n = 814)		White bread (n = 758)		Milk fat (n = 675)		Sweets and cakes (n = 1092)	
		OR	95%CI	OR	95%CI	OR	95%CI	OR	95%CI	OR	95%CI	OR	95%CI
<b>EI/BMR</b>													
≤ 1.35	1623	0.95	0.83–1.09	1.30	1.06–1.59	2.73	2.21–3.22	0.83	0.69–0.99	0.56	0.46–0.69	0.61	0.52–0.72
> 1.35	3671	1.00		1.00		1.00		1.00		1.00		1.00	
<b>Body mass index</b>													
< 23.9	1762	0.84	0.71–0.99	1.33	1.03–1.72	0.70	0.57–0.87	0.79	0.63–0.98	1.33	1.05–1.67	1.41	1.17–1.69
23.9–26.9	1771	0.99	0.85–1.15	1.15	0.91–1.45	0.84	0.69–1.03	0.84	0.70–1.02	1.14	0.92–1.42	1.15	0.96–1.37
> 26.9	1761	1.00		1.00		1.00		1.00		1.00		1.00	
<b>Waist–hip ratio</b>													
< 0.79	1763	0.78	0.59–1.03	0.76	0.50–1.15	1.52	1.04–2.22	0.90	0.63–1.28	1.16	0.80–1.68	1.11	0.82–1.50
0.79–0.90	1778	0.80	0.64–1.01	0.85	0.62–1.16	1.36	0.98–1.90	0.97	0.75–1.26	1.08	0.82–1.43	1.08	0.85–1.37
> 0.90	1753	1.00		1.00		1.00		1.00		1.00		1.00	
<b>Past food habit change</b>													
Yes	1379	0.72	0.62–0.83	1.75	1.44–2.13	2.43	2.06–2.86	0.93	0.77–1.12	0.48	0.38–0.60	0.79	0.68–0.93
No	3915	1.00		1.00		1.00		1.00		1.00		1.00	

\* Adjusted for age, gender, diet assistant and season of data collection.

**Energy underreporting, obesity and past food habit change**

As can be seen from Fig. 1, individuals with past food habit changes were more common relative to those without changes in clusters with 'healthier' food choices, that is the 'fibre bread' and 'low fat and high fibre' clusters from unstandardized variables, and the 'dieters', 'healthy' and 'Mediterranean' clusters from Z-scores. Also, analyses indicated that the average EI/BMR ratios were lower, and mean BMI and WHR were higher, in clusters with 'healthy' food choices for both types of variables. However, differences in EI/BMR between clusters appeared larger with the unstandardized variables, and the differences in obesity indicators appeared larger with the Z-scored variables.

The risk of falling into specific clusters related to past food habit change, energy underreporting, relative weight

(BMI) and body fat distribution (WHR) are shown for unstandardized variables in Table 5 and for Z-scores in Table 6. Individuals falling in both the 'low fat and high fibre' and the 'fibre bread' clusters (Table 5) were likely to have changed their food habits in the past, and to be underreporters of energy. Past food habit change was less likely in the 'many foods and drinks', 'milk fat' and 'sweets and cakes' clusters, and underreporting of energy was less likely in the 'white bread', 'milk fat' and 'sweets and cakes' clusters. Those in the 'low fat and high fibre' cluster were more likely to have a high BMI and low WHR. However, a low BMI was more common among individuals of the 'fibre bread' cluster. Also, a high BMI was more common in the 'many foods and drinks' and 'white bread' clusters, but a low BMI was more common among individuals of the 'milk fat' and 'sweets and cakes' clusters.

**Table 6** The risk of falling into each food pattern cluster (generated from standardized variables) in relation to some potential confounders\*

	Total	Drinks and fries (n = 182)		Ice-cream and cake (n = 245)		Dieters (n = 47)		Healthy (n = 1317)		Traditional (n = 2493)		Mediterranean (n = 1073)	
		OR	95%CI	OR	95%CI	OR	95%CI	OR	95%CI	OR	95%CI	OR	95%CI
<b>EI/BMR</b>													
≤ 1.35	1623	0.94	0.67–1.32	0.69	0.51–0.93	1.93	1.06–3.53	1.47	1.28–1.69	0.40	0.35–0.46	2.35	2.03–2.72
> 1.35	3671	1.00		1.00		1.00		1.00		1.00		1.00	
<b>Body mass index</b>													
< 23.9	1762	1.20	0.79–1.81	0.57	0.40–0.81	0.52	0.22–1.21	1.02	0.85–1.22	1.12	0.95–1.31	0.99	0.82–1.20
23.9–26.9	1771	1.11	0.76–1.63	0.69	0.50–0.95	0.66	0.33–1.33	1.04	0.88–1.22	0.94	0.81–1.09	1.18	0.99–1.40
> 26.9	1761	1.00		1.00		1.00		1.00		1.00		1.00	
<b>Waist–hip ratio</b>													
< 0.79	1763	0.48	0.24–0.95	1.17	0.65–2.09	0.50	0.13–1.84	0.87	0.65–1.17	1.18	0.91–1.53	1.04	0.76–1.43
0.79–0.90	1778	0.58	0.33–1.02	1.18	0.74–1.88	0.65	0.21–2.03	0.88	0.68–1.13	1.16	0.94–1.43	0.97	0.74–1.27
> 0.90	1753	1.00		1.00		1.00		1.00		1.00		1.00	
<b>Past food habit change</b>													
Yes	1379	0.74	0.51–1.07	0.78	0.57–1.07	3.50	1.94–6.31	2.13	1.84–2.44	0.36	0.32–0.42	1.63	1.40–1.90
No	3915	1.00		1.00		1.00		1.00		1.00		1.00	

\* Adjusted for age, gender, diet assistant and season of data collection.

Table 6 shows that past food habit change and underreporting of energy were more common in the 'dieters', 'healthy' and 'Mediterranean' clusters. However, past food habit change was less common among those in the 'traditional' cluster, and energy underreporting was less common in the 'ice-cream and cake' and 'traditional' clusters. Also, those in the 'ice-cream and cake' cluster were less likely to fall in the lower two tertiles of BMI. Individuals in the 'drinks and fries' cluster were less likely to have a low WHR.

## Discussion

### *Stability of clusters and utility of variables*

Cluster analysis is exploratory in nature. The relevance of emerging clusters is commonly evaluated by comparing them on external variables not used in analysis<sup>12,14</sup>. For instance, in this study differences in both macro- and micronutrient intakes indicated that clusters represented distinct dietary exposure categories. This study also used discriminant analysis to examine the stability of the emerging clusters. In contrast to cluster analysis (which uses distance measures between observations to categorize individuals) this methodology bases group prediction on correlations between variables using a regression model<sup>29</sup>. Thus the theories behind the two methodologies are different. With unstandardized variables the distinction between methodologies is quite clear. However, the predictions obtained with standardized variables in methods using distance measures are very similar to the predictions obtained with methods using correlations. (In fact with standardized variables (Z-scores), the Euclidean distance between two observations is a monotone transformation of their correlation<sup>30</sup>.) Therefore it is not surprising that only 18 of the unstandardized variables contributed significantly in the discriminant analysis, but as many as 31 of the Z-scored variables did contribute significantly. Still, the correspondence between the original clusters and groups predicted with discriminant analysis was very high for both variable types.

Studies using cluster analysis have often chosen 'density' variables indicating the per cent contribution of energy from macronutrients<sup>7</sup>, or from specific food groups<sup>10,15-17</sup>, to total energy. Z-score transformation is recommended in cluster analysis to hinder variables with larger values from overwhelming those with much smaller values<sup>27</sup>. However, transformation of per cent energy food variables will give equal weight to foods contributing large amounts of energy to those that contribute much less. As can be seen from Appendix 2, clusters generated from Z-scores in this study did not reflect how food energy was distributed in the over-all diet, but rather separated out individuals with extreme percentages for some foods. Consequently, the dietary compositions for these food patterns were more similar to one another, and appear less useful, compared to food patterns from unstandardized

variables. It can be argued that per cent energy variables are in a sense standardized and that no further standardization is needed. The two standardization procedures have, however, very different purposes and do not replace one another. Per cent energy variables are commonly used in nutrition studies to describe 'dietary quality'. In this study the quality aspect (or the energy contribution of specific food groups to total energy) of the original variables was removed with Z-score transformation. This implies that Z-scores could give misleading results when the underlying hypothesis of the study is based on arguments related to the macronutrient density of diet. Also, studies using other variables (like food grams or frequencies) need to carefully consider whether the purpose of the study is to give an overall account of diet, or bring forward 'odd' exposures, which may have very little to do with the over-all dietary composition.

### *Comparison with other studies of food patterns*

Published work on food patterns has mostly been descriptive, and only to a limited extent has discussed the relation between analysis procedure, variable selection and study outcome. Also, because of major differences in design and analysis it is difficult to compare findings from different dietary pattern studies. For instance, the comparison of results from cluster analysis with those from factor analysis is not straightforward. The latter technique reduces large data sets by searching for highly correlated variables, and identifies a few uncorrelated factors, that may not be overtly observable<sup>11,29</sup>. Although factor analysis has been cited for its statistical properties<sup>31</sup>, a major disadvantage is that individuals will belong to many of the identified factors. Moreover, there is no guarantee that factors will have any conceptual meaning<sup>11,32</sup>, nor is it clear how factor scores relate to total dietary composition or how findings best translate into dietary recommendations. Factor analysis has been successful in identifying factors that predict risk of colon cancer<sup>6,9</sup>. However, in order to determine the relation between these factors and the dietary composition, additional data examination is needed. Cluster analysis on the other hand creates non-overlapping exposure categories, which are easy to interpret, and potentially allows identification of subgroups with specific dietary behaviours. Also, cluster analysis is truly exploratory and does not require data with specific conditions. Other studies using per cent energy food variables to generate clusters have found that food patterns were related to differences in nutritional status and health-disease indicators<sup>10,15-17</sup>, and provided easy to understand descriptions of over-all diets.

Whatever analysis approach is used, the emerging food patterns will potentially be influenced by the measurement errors of dietary data collection methods, and will probably vary between countries and populations. In addition, the grouping of foods, and thus the number of

variables entered into the analysis procedure, will influence the degree of detail captured by the emerging food patterns<sup>33,34</sup>. In this study the major food energy sources were basically the same as those found in similar gender and age groups of the Swedish Food Consumption Survey (i.e. milk and cheese, meat, bread, fats/oils and sweets)<sup>35</sup>, although the dietary proportion of fats/oils and sweets appeared slightly greater and that of bread and cereals slightly smaller in the examined population. This study aggregated foods to separate out food groups with low and high fat content of dairy products, dietary fats and oils, meats, fish, cakes, candy and potatoes. Also, carbohydrate containing foods high in fibre were separated from those low in fibre and high in sucrose. Nutrient intake comparisons across clusters (generated from unstandardized variables) indicated several variations in nutrient quality of diet which may not have been expressed if broader more generic food groups had been used. For instance, three different ways of consuming bread and dietary fats emerged, which clearly resulted in varied intakes of fat and micronutrients. Individuals in the 'fibre bread' and 'white bread' clusters selected low-fat margarine with their bread, had comparatively lower fat intakes and higher P/S ratios. Those in the 'milk fat' cluster preferred a butter-based dietary spread, which resulted in high fat intakes and low P/S ratios, accompanied by lower intakes of several micronutrients. However,  $\omega$ -3 fatty acid intakes were, in this cluster, comparatively higher. Also, the cluster high in sweets had comparatively higher carbohydrate intakes and medium–low fat intakes, but intakes of fibre and several micronutrients were lower. This illustrates that low-fat diets are not necessarily healthier diets. The findings highlight the potential confounding and counteractive effects of various dietary components, which may contribute to conflicting results in nutrition epidemiology studies.

### ***Underreporting of energy***

General and selective underreporting are potential limitations to this study. Energy underreporting appears to be food-pattern specific. The results suggest that food selection was influenced by past change of food habits and the degree of obesity. This is in line with other studies that previously have described underreporting of energy as more common among the obese, and related to dieting behaviour<sup>36–38</sup>. Underreporting, which is a common concern in dietary studies<sup>24,25</sup>, has been reported as more common in certain socioeconomic and lifestyle strata<sup>37–41</sup>. Food- and nutrient-specific underreporting has been observed, both in 'current' diet studies (i.e. underreporting of snacks or foods eaten between meals)<sup>42</sup> and 'usual' diet studies (i.e. carbohydrate- and fat-containing foods)<sup>36,43</sup>. Others have indicated that energy underreporters specifically may overreport nutrient dense foods<sup>44,45</sup>. The food-pattern-specific underreporting of this study may be related either to past change of diet,

or to common social norms of an acceptable healthy diet. This study had no information about prescribed diets (e.g. due to weight reduction or diabetes). However, food choices reported by individuals of the 'low fat and high fibre' cluster may reflect dietary change and dieting in overweight individuals. Also, the higher risk of past food habit change and underreporting in the 'fibre bread' cluster may indicate current weight maintenance, or dieting and health concerns, in non-obese individuals.

Other studies have indicated that both dietary under- and overreports may be method specific<sup>34,46–51</sup>. Previous studies have generally reported that diet history methods overestimate intakes by 10–20%<sup>52</sup>. When the relative validity of the MDC diet history method was compared with 18 days of weighed food records both over- and underestimations of the reported food amounts were observed<sup>21</sup>. For instance, added fats and oils, milk, fruits and potatoes were overestimated both in men and women. Women tended to underestimate alcoholic beverages and cream, and men underestimated fish and cream. Although, the possibility of a method effect on the self-reported food selection can not be ignored, the validation study does not suggest that such an effect specifically is related to 'healthy' food choices, nor does it invalidate the observations in this study that past diet change, energy underreporting and obesity are related to specific food patterns. However, future studies of the MDC cohort need to carefully consider the potential confounding associated with past diet change and energy underreporting. Studies of diet–disease relations may need to limit analysis to those that report no diet change in the past. Other studies may find it necessary to stratify analysis on past food habit change, energy underreporting or obesity status, or control for these factors in analysis.

To conclude, when dissecting dietary data to identify complex food combinations for studies of disease aetiology, the specific selection of food variables will undoubtedly be guided by the underlying diet–disease hypothesis. This study suggested that Z-scores of per cent energy food variables were less successful in discriminating individuals in dietary exposure groups, and may provide misleading results, compared to unstandardized per cent energy food variables. Also, in order to ensure the expression of specific food behaviours, the groupings of foods need to be carefully selected. Measurement errors and group-specific biases, associated with self-reported dietary intakes, may not be equally distributed across emerging food patterns, and need to be accounted for in data analysis and interpretation of results.

### **Acknowledgements**

The authors wish to acknowledge the funding from the Swedish Cancer Society, the Swedish Medical Research Council, the European Commission and the City of Malmö.

## References

- 1 Schatzkin A, Dorgan J, Swanson C, Potischman N. Diet and cancer: future etiologic research. *Environ. Health Perspect.* 1995; **103** (Suppl. 8): 171–5.
- 2 Kohlmeier L, Simonsen N, Mottus K. Dietary modifiers of carcinogenesis. *Environ. Health Perspect.* 1995; **103** (Suppl. 8): 177–84.
- 3 Potter JD. Food and phytochemicals, magic bullets and measurement error: a commentary. *Am. J. Epidemiol.* 1996; **144**(11): 1026–7.
- 4 Byers T, Gieseke K. Issues in the design and interpretation of studies of fatty acids and cancer in humans. *Am. J. Clin. Nutr.* 1997; **66** (Suppl.): S1541–7.
- 5 Palmgren J. Controlling for total energy intake in regression models for assessing macronutrient effects on disease. *Eur. J. Clin. Nutr.* 1993; **47** (Suppl. 2): S46–50.
- 6 Randall E, Marshall JR, Brasure J, Graham S. Dietary patterns and colon cancer in western New York. *Nutr. Cancer* 1992; **18**: 265–76.
- 7 Farchi G, Mariotti S, Menotti A, Seccareccia F, Torsello S, Fidanza F. Diet and 20-y mortality in two rural population groups of middle-aged men in Italy. *Am. J. Clin. Nutr.* 1989; **50**: 1095–103.
- 8 Huijbregts P, Feskens E, Kromhout D. Dietary patterns and cardiovascular risk factors in elderly men: the Zutphen Elderly Study. *Int. J. Epidemiol.* 1995; **24**: 313–20.
- 9 Slattery ML, Boucher KM, Caan BJ, Potter JD, Ma K-N. Eating patterns and colon cancer. *Am. J. Epidemiol.* 1998; **148**(1): 4–16.
- 10 Tucker K, Dallal G, Rush D. Dietary patterns of elderly Boston-area residents defined by cluster analysis. *J. Am. Diet. Assoc.* 1992; **92**: 1487–91.
- 11 Jacobson H, Stanton J. Pattern analysis in nutrition. *Clin. Nutr.* 1986; **5**: 249–53.
- 12 Aldenderfer MS, Blashfield RK. *Cluster Analysis. Quantitative Applications in the Social Sciences.* Newbury Park, CA: Sage Publications, 1984.
- 13 Hulshof K, Wedel M, Löwik M, *et al.* Clustering of dietary variables and other lifestyle factors (Dutch Nutritional Surveillance System). *J. Epidemiol. Community Health* 1992; **46**: 417–24.
- 14 Patterson RE, Haines PS, Popkin BM. Health lifestyle patterns of US adults. *Prev. Med.* 1994; **23**: 453–60.
- 15 Tucker K, Hannan MT, Chen H, *et al.* Diet pattern groups relate to bone mineral density (BMD) among elders in the Framingham heart study. *Eur. J. Clin. Nutr.* 1998; **52** (Suppl. 2): S78 (abstract).
- 16 Wirfält AKE, Jeffery RW, Elmer PJ. Using cluster analysis to examine dietary patterns: nutrient intakes, gender, and weight status differ across food pattern clusters. *J. Am. Diet. Assoc.* 1997; **97**: 272–9.
- 17 Haveman-Nies A, de Groot LCPGM, van Staveren WA. Snack patterns of older Europeans. *J. Am. Diet. Assoc.* 1998; **98**: 1297–302.
- 18 Berglund G, Elmståhl S, Janzon L, Larsson SA. The Malmö Diet and Cancer Study: design and feasibility. *J. Intern. Med.* 1993; **233**: 45–51.
- 19 Callmer E, Riboli E, Saracci R, Åkesson B, Lindgärde F. Dietary assessment methods evaluated in the Malmö food study. *J. Intern. Med.* 1993; **233**: 53–7.
- 20 Elmståhl S, Gullberg B, Riboli E, Saracci R, Lindgärde F. The Malmö Food Study: the reproducibility of a novel diet history method and an extensive food frequency questionnaire. *Eur. J. Clin. Nutr.* 1996; **50**: 134–42.
- 21 Elmståhl S, Riboli E, Lindgärde F, Gullberg B, Saracci R. The Malmö Food Study: the relative validity of a modified diet history method and an extensive food frequency questionnaire for measuring food intake. *Eur. J. Clin. Nutr.* 1996; **50**: 143–51.
- 22 Riboli E, Elmståhl S, Saracci R, Gullberg B, Lindgärde F. The Malmö Food Study: validity of two dietary assessment methods for measuring nutrient intake. *Int. J. Epidemiol.* 1997; **26** (Suppl. 1): S161–73.
- 23 Willett W. *Nutritional Epidemiology*, 2nd edn., New York: Oxford University Press, 1998.
- 24 Goldberg GR, Black AE, Jebb SA, *et al.* Critical evaluation of energy intake data using fundamental principles of energy physiology: 1. Derivation of cut-off limits to identify under-reporting. *Eur. J. Clin. Nutr.* 1991; **45**: 569–81.
- 25 Black AE, Goldberg GR, Jebb SA, Livingstone MBE, Cole TJ, Prentice AM. Critical evaluation of energy intake data using fundamental principles of energy physiology: 2. Evaluating the results of published surveys. *Eur. J. Clin. Nutr.* 1991; **45**: 583–99.
- 26 FAO/WHO/UNU. *Energy and Protein Requirements Report of a Joint FAO/WHO/UNU Expert Consultation.* Technical Report Series No. 724. Geneva: World Health Organization, 1985.
- 27 Norusis MJ. *SPSS for Windows Professional Statistics Release 6.0.* Chicago: SPSS Inc., 1993.
- 28 Willett WC, Howe GR, Kushi LH. Adjustment for total energy intake in epidemiologic studies. *Am. J. Clin. Nutr.* 1997; **65** (Suppl.): S1220–8.
- 29 Kim J-O, Mueller CW. *Factor Analysis Statistical Methods and Practical Issues.* Newbury Park, CA: Sage Publications, 1978.
- 30 Coombs CH. *A Theory of Data.* New York: John Wiley & Sons, 1964.
- 31 Herbert JR, Stoddard AM, Harris DR. Measuring the effect of a worksite-based nutrition intervention on food consumption. *Ann. Epidemiol.* 1993; **3**: 629–35.
- 32 Herbert JR, Kabat GC. Implications for cancer epidemiology of differences in dietary intake associated with alcohol consumption. *Nutr. Cancer* 1991; **15**: 107–19.
- 33 Krebs-Smith SM, Cronin FJ, Haytowitz DB, Cook DA. Contributions of food groups to intakes of energy, nutrients, cholesterol, and fiber in women's diets: effect of method of classifying food mixtures. *J. Am. Diet. Assoc.* 1990; **90**: 1541–6.
- 34 Serdula M, Byers T, Coates R, Mokdad A, Simoes E, Eldridge L. Assessing consumption of high-fat foods: the effect of grouping foods into single questions. *Epidemiology* 1992; **3**(6): 503–8.
- 35 Becker W. *Food Habits and Nutrient Intake in Sweden 1989.* Uppsala, Sweden: The National Food Administration, 1994.
- 36 Johansson L, Solvoll K, Björneboe G-EA, Drevon CA. Under- and overreporting of energy intake related to weight status and lifestyle in a nationwide sample. *Am. J. Clin. Nutr.* 1998; **68**: 266–74.
- 37 Braam LAJL, Ocké MC, Bueno-de-Mesquita HB, Seidell JC. Determinants of obesity-related underreporting of energy intake. *Am. J. Clin. Nutr.* 1998; **147**(11): 1081–6.
- 38 Lafay L, Basdevant A, Charles M-A, *et al.* Determinants and nature of delivery underreporting in a free-living population: the Fleurbaix Laventie Ville Santé (FLVS) study. *Int. J. Obes.* 1997; **21**: 567–73.
- 39 Stallone DD, Brunner EJ, Bingham S. Dietary assessment in Whitehall II: the influence of reporting bias on apparent socioeconomic variation in nutrient intakes. *Eur. J. Clin. Nutr.* 1997; **51**: 1–11.
- 40 Pryer JA, Vrijheid M, Nichols R, Kiggins M, Elliott P. Who are the 'low energy reporters' in the Dietary and Nutritional Survey of British Adults? *Int. J. Epidemiol.* 1997; **26**(1): 146–54.
- 41 Klesges RC, Eck LH, Ray JW. Who underreports dietary intake in a dietary recall? Evidence from the second National Health and Nutrition Examination Survey. *J. Consult. Clin. Psychol.* 1995; **63**(3): 438–44.
- 42 Poppitt SD, Swann D, Black AE, Prentice AM. Assessment of selective under-reporting of food intake by both obese and

- non-obese women in a metabolic facility. *Int. J. Obes.* 1998; **22**: 203–11.
- 43 Lilienthal Heitmann B, Lissner L. Dietary underreporting by obese individuals – is it specific or non-specific? *BMJ* 1995; **311**: 986–9.
- 44 Price GM, Paul AA, Cole TJ, Wadsworth MEJ. Characteristics of the low-energy reporters in a longitudinal national dietary survey. *Br. J. Nutr.* 1997; **77**: 833–51.
- 45 Hirvonen T, Männistö S, Roos E, Pietinen P. Increasing prevalence of underreporting does not necessarily distort dietary surveys. *Eur. J. Clin. Nutr.* 1997; **51**: 297–301.
- 46 Krebs-Smith SM, Heimdinger J, Subar AF, Patterson BH, Pivonka E. Using food frequency questionnaires to estimate fruit and vegetable intake: association between the number of questions and total intake. *J. Nutr. Educ.* 1995; **27**: 80–5.
- 47 Boutron MC, Faivre J, Milan C, Lorcerie B, Esteve J. A comparison of two diet history questionnaires that measure usual food intake. *Nutr. Cancer* 1989; **12**: 83–91.
- 48 Wolk A, Bergström R, Adami H-O, *et al.* Self-administered food frequency questionnaire: the effect of different designs on food and nutrient intake estimates. *Int. J. Epidemiol.* 1994; **23**(3): 570–6.
- 49 Kuskowska-Wolk A, Holte S, Ohlander E-M, *et al.* Effects of different designs and extension of a food frequency questionnaire on response rate, completeness of data and food frequency responses. *Int. J. Epidemiol.* 1992; **21**(6): 1144–50.
- 50 Körtzinger I, Bierwag A, Mats M, Müller MJ. Dietary underreporting: validity of dietary measurements of energy intake using a 7-day dietary record and a diet history in non-obese subjects. *Nutr. Metab.* 1997; **41**: 37–44.
- 51 Wirfält AKE, Jeffery RW, Elmer PJ. Comparison of food frequency questionnaires: the reduced Block and Willett questionnaires differ in ranking on nutrient intakes. *Am. J. Epidemiol.* 1998; **148**(12): 1148–56.
- 52 Margetts BM, Nelson M. *Design Concepts in Nutritional Epidemiology*, 2nd edn. Oxford: Oxford University Press, 1998.

### Appendix 1: per cent energy contribution from food groups by food patterns (unstandardized)

Food groups	Total sample (n = 5357)	Many foods and drinks (n = 1508)	Fibre bread (n = 510)	Low fat and high fibre (n = 814)	White bread (n = 758)	Milk fat (n = 675)	Sweets and cakes (n = 1092)
Sweets	8.5	5.0	6.0	5.1	6.6	7.6	18.0
High-fat meats	6.7	7.1	6.4	5.4	7.9	6.8	6.1
White bread	6.6	5.1	2.8	3.5	16.0	7.3	5.9
Fruits	5.7	5.3	5.9	10.5	4.4	4.0	4.7
Cheese	5.3	7.4	4.3	3.1	4.9	6.0	4.3
Low-fat meats	4.9	5.0	5.1	5.5	5.1	4.8	3.9
Cake, pastry	4.4	5.5	3.5	3.7	3.6	4.1	4.6
Fibre bread	4.3	3.4	15.0	3.7	2.1	3.2	3.1
Low-fat milk	4.2	3.6	3.9	8.0	3.6	2.4	3.7
Boiled potatoes	3.7	3.7	4.1	3.7	4.0	3.5	3.3
Whole milk	3.4	3.5	2.6	1.8	3.4	5.0	3.9
Low-fat spread	3.4	3.5	4.5	3.3	5.3	0.3	3.4
Medium-fat spread	3.0	3.7	2.6	2.7	3.2	2.5	2.9
Vegetables	2.8	2.8	3.1	3.8	2.4	2.3	2.4
Cereals	2.6	2.4	2.6	3.5	2.1	2.1	2.7
Bregott spread	2.4	0.7	1.0	0.6	0.7	12.3	1.6
Beer	2.0	2.6	2.0	1.4	1.9	2.4	1.5
Fibre crisp-bread	2.0	2.1	1.9	3.3	1.2	1.6	1.6
Rice, pasta	1.9	2.0	2.1	2.4	1.8	1.4	1.7
Chocolates	1.8	2.2	1.3	1.6	1.3	1.9	2.0
Low-fat cake	1.8	1.9	1.8	1.8	1.8	1.4	2.0
Wine	1.6	2.3	1.3	1.8	1.1	1.5	1.0
Egg	1.5	1.7	1.4	1.8	1.3	1.4	1.3
High-fat fish	1.5	1.7	1.5	1.9	1.2	1.2	1.3
Low-fat fish	1.3	1.5	1.4	1.8	1.0	1.1	1.1
Fried potatoes	1.2	1.5	1.0	0.9	1.6	1.4	1.0
Fruit juice	1.1	1.2	1.1	1.4	0.8	1.1	1.1
Cottage cheese	1.1	0.6	1.7	2.6	1.1	0.5	0.8
Soft drinks	1.1	1.1	0.8	0.7	1.2	1.1	1.2
Ice-cream	1.0	1.1	0.8	1.0	0.9	0.9	0.9
Dressing	0.9	1.1	0.7	0.8	0.9	0.9	0.7
White crisp-bread	0.9	0.9	0.7	1.0	0.8	0.8	0.9
Liquor	0.7	0.9	0.6	0.4	0.6	0.8	0.5
Butter	0.5	0.6	0.4	0.3	0.5	0.3	0.6
Coffee	0.5	0.5	0.4	0.6	0.5	0.4	0.4
Oil	0.4	0.5	0.5	0.6	0.4	0.3	0.4
Low-fat dressing	0.3	0.4	0.2	0.3	0.2	0.3	0.2
Snacks	0.3	0.3	0.2	0.2	0.2	0.3	0.2
Sherbet	0.1	0.1	0.1	0.2	0.09	0.06	0.08
Tea	0.05	0.05	0.07	0.06	0.04	0.04	0.05

**Appendix 2: per cent energy contribution from food groups by food patterns (standardized)**

Food groups	Total sample (n = 5357)	Drinks and fries (n = 182)	Ice-cream and cake (n = 245)	Dieters (n = 47)	Healthy (n = 1317)	Traditional (n = 2493)	Mediterranean (n = 1073)
Sweets	8.5	6.9	6.8	5.4	8.0	10.6	4.7
High-fat meats	6.7	6.7	6.2	6.1	6.7	7.2	5.4
White bread	6.6	6.7	5.5	5.5	5.3	8.7	4.3
Fruits	5.7	4.8	6.1	8.3	7.1	4.2	7.5
Cheese	5.3	5.6	6.0	2.9	3.2	5.6	6.2
Low-fat meats	4.9	5.2	4.8	5.4	4.6	4.5	6.0
Cake, pastry	4.4	4.2	5.2	4.0	4.3	4.9	3.2
Fibre bread	4.3	3.1	3.5	4.0	2.7	3.4	4.5
Low-fat milk	4.2	3.9	3.4	6.0	6.7	2.7	4.5
Boiled potatoes	3.7	3.2	3.3	3.3	4.2	3.6	3.3
Whole milk	3.4	3.1	4.0	1.7	2.0	4.6	2.4
Low-fat spread	3.4	3.2	3.1	4.1	5.3	2.8	2.4
Medium-fat spread	3.0	3.4	2.8	3.0	2.6	3.3	3.0
Vegetables	2.8	2.6	2.5	3.2	3.1	2.2	3.9
Cereals	2.6	2.4	2.3	2.7	2.9	2.2	3.0
Bregott spread	2.4	1.7	1.8	0.8	0.6	3.8	1.5
Beer	2.0	2.6	1.6	1.7	1.3	2.3	2.3
Fibre crisp-bread	2.0	1.7	1.9	2.6	2.7	1.3	2.7
Rice, pasta	1.9	2.0	1.8	2.3	1.5	1.5	3.3
Chocolates	1.8	1.9	2.8	1.9	1.5	2.1	1.3
Low-fat cake	1.8	1.0	2.2	1.5	2.5	1.8	1.1
Wine	1.6	2.1	1.4	1.2	0.8	1.2	3.4
Egg	1.5	1.6	1.4	1.5	1.5	1.3	2.0
High-fat fish	1.5	1.3	1.5	2.5	1.5	1.2	2.1
Low-fat fish	1.3	1.5	1.3	1.5	1.3	1.0	2.0
Fried potatoes	1.2	1.8	1.3	1.1	0.8	1.4	1.3
Fruit juice	1.1	1.3	1.1	0.8	1.3	1.0	1.3
Cottage cheese	1.1	1.2	0.7	2.9	2.5	0.3	1.1
Soft drinks	1.1	1.1	1.1	0.8	0.9	1.3	0.6
Ice-cream	1.0	0.9	5.8	1.2	0.7	0.7	0.7
Dressing	0.9	1.2	1.1	0.6	0.6	0.9	1.4
White crisp-bread	0.9	0.8	0.7	1.0	1.4	0.7	0.7
Liquor	0.7	1.1	0.4	0.3	0.3	0.8	0.8
Butter	0.5	0.3	0.6	0.01	0.2	0.7	0.5
Coffee	0.5	0.5	0.5	0.7	0.5	0.4	0.5
Oil	0.4	0.4	0.4	0.5	0.3	0.3	1.1
Low-fat dressing	0.3	3.7	0.2	0.4	0.2	0.2	0.2
Snacks	0.3	0.3	0.3	0.1	0.1	0.3	0.3
Sherbet	0.1	0.07	0.2	3.4	0.1	0.1	0.08
Tea	0.05	0.03	0.04	0.03	0.06	0.04	0.07