Identifying dietary patterns using a normal mixture model: application to the EPIC study

Michael T Fahey, 1 Pietro Ferrari, 2 Nadia Slimani, 3 Jeroen K Vermunt, 4 Ian R White, 1 Kurt Hoffmann, 5 Elisabet Wirfält, 6 Christina Bamia, 7 Mathilde Touvier, 8, 9 Jakob Linseisen, 10 Miguel Rodríguez-Barranco, 11 Rosario Tumino, 12 Eiliv Lund, 13 Kim Overvad, 14 Bas Bueno de Mesquita, 15 Sheila Bingham, 16 Elio Riboli 17

ABSTRACT
Background Finite mixture models posit the existence of a latent categorical variable and can be used for probabilistic classification. The authors illustrate the use of mixture models for dietary pattern analysis. An advantage of this approach is taking classification uncertainty into account.

Methods Participants were a random sample of women from the European Prospective Investigation into Cancer. Food consumption was measured using dietary questionnaires. Mixture models identified latent classes in food consumption data, which were interpreted as dietary patterns.

Results Among various assumptions examined, models allowing the variance of foods to vary within and between classes fit better than alternatives assuming constant variance (the K-means method of cluster analysis also makes the latter assumption). An eight-class model was best fitting and five patterns validated well in a second random sample. Patterns with lower classification uncertainty tended to be better validated. One pattern showed low consumption of foods despite being associated with moderate body mass index.

Conclusion Mixture modelling for dietary pattern analysis has advantages over both factor and cluster analysis. In contrast to these other methods, it is easy to estimate pattern prevalence, to describe patterns and to use patterns to predict disease taking classification uncertainty into account. Owing to substantial error in food consumptions, any analysis will usually find some patterns that cannot be well validated. While knowledge of classification uncertainty may aid pattern evaluation, any method will better identify patterns from food consumptions measured with less error. Mixture models may be useful to identify individuals who under-report food consumption.

INTRODUCTION
Most of the many recent dietary pattern analyses 1–8 have used one of two apparently contrasting statistical methods to find patterns in food consumption data. The first approach using factor analysis finds dimensions in the diet that represent an individual’s tendency to eat in certain ways. For example, reports on the Health Professionals Study 5, 6 and the Nurses’ Health Study 7, 8 have described ‘Western’ and ‘prudent’ dimensions in the diets of their participants. A disadvantage, however, is that an additional step of cross-classifying the dimensions in some way is necessary if one wishes to compute pattern prevalence or risk of disease for one group of individuals compared with another group.

The second approach classifies individuals into mutually exclusive groups according to how (dis)similar they are with respect to their food consumption using (non-parametric) cluster analysis, for example, the K-means method used by Chen et al. 9 A disadvantage of cluster analysis is that each individual is assigned to one dietary pattern with a probability of 1 and all others to a dietary pattern with a probability of 0. Thus, classification uncertainty is assumed to be 0. Other disadvantages stemming from non-parametric approaches include the difficulty in taking into account covariates and the lack of a convenient way to compare the many different clustering criteria.

The primary objectives of a dietary pattern analysis are to characterise the eating habits of a population and to associate diet with disease 10, 11. A finite mixture model (FMM) can be used to achieve these objectives with additional advantages as outlined by Fahey et al. 12 Classification uncertainty is measured by the posterior probability of pattern membership given the data, which for each individual, may take values between zero and one. It is also easy to adjust for energy and to choose among different clustering criteria. An FMM is analogous to a factor analysis with a categorical latent variable and can be used to create mutually exclusive groups. However, it can also be used to estimate dietary pattern prevalence and to describe patterns without ‘hard’ classification of individuals to clusters. Instead, classification is ‘soft’ with estimates weighted by the posterior probabilities. We adapt the general approach outlined by Fahey et al. 12 in this paper for a very large multi-centre cohort study and show how an FMM restricted to be a mixture of multivariate normal (MVN) distributions can be used to find, interpret and validate dietary patterns.

MATERIALS AND METHODS
Study design, measures and subjects
The European Prospective Investigation into Cancer (EPIC) cohort study consists of approximately 520,000 individuals recruited into subcohorts by 23 research centres in 10 countries. 13 A subsample of the female participants was used here because dietary patterns have been shown to be gender specific. 12 Two random samples stratified by EPIC centre, each having 6009 women, were drawn for numbered affiliations see end of article.

Correspondence to Michael T Fahey, Private Bag 33, Clayton South, VIC 3169 Australia michael.fahey@csiro.au

The co-authors Kurt Hoffmann and Sheila Bingham are now deceased.

Accepted 15 June 2011 Published Online First 28 August 2011
from the population of EPIC subcohorts after excluding women in the top and bottom percentiles of energy intake. The model was developed on the ‘estimation’ sample and validated on the other. Self-reported diet was assessed using country-specific dietary questionnaires (DQs). Food group consumption (g/d) was calculated from the DQs. The 24 food groups listed in table 1 are the slightly modified versions of the food groups used by Slimani et al.14 Log-transformed values of consumptions (g/d + 1) were used in all analyses and back-transformed to the original scale for presentation.

The finite mixture model
We define dietary patterns as unobserved classes in a sample having different food consumption probability distributions. They are identified by decomposing the aggregate distribution into a sum of class-specific food consumption distributions as per equation (1)15:

\[ f_0 = \pi_1 f_1 + \pi_2 f_2 + \cdots + \pi_K f_K \]  

The multivariate probability density function for the observed food consumptions, denoted \( f_0 \), is a mixture of \( K \) class-specific probability densities, \( f_1, \ldots, f_K \). A normal mixture model postulates that the \( K \) probability densities each have an MVN distribution. Thus, the \( k \)th density \( f_k \), is defined by mean food consumptions and a covariance matrix \( \Sigma_k \), containing parameters for the vari-ances and covariances among the food consumptions. The parameters \( \pi_1, \ldots, \pi_K \) are the class prevalences, indicating the proportion of the aggregate data described by each of the \( K \) probability densities. Identification of the class-specific densities is done after fixing their number (\( K \)) and, in principle, is achieved by finding the \( K \)-category latent variable that ‘best’ explains associations among the observed food consumptions within classes. See McLachlan et al15 for examples using other types of data.

### Table 1
Content of food groups and abbreviations (EPIC, 1992–2000)

| Food groups (abbreviation) | Content |
|----------------------------|---------|
| Vegetables (veget)         | Leafy, fruiting, root, grain, pod and stalk vegetables, mushroom, allium, cruciferous, sprouts and mixed salad/vegetables |
| Fruits (fruit)             | Fresh fruits, nuts, seeds, stewed fruit, mixed fruits and olives |
| Potatoes (potat)           | Potatoes and potato products, except chips |
| Legumes (legum)            | Dried peas, lentils and beans, except soya |
| Cereals and cereal products (cerea) | Flour, flakes, starches, pasta, rice and other grains, bread, crispbread, rusks, breakfast cereals, salty and aperitif biscuits, dough and pastry (puff, short-crust, pizza) |
| Cakes (cake)               | Cakes, pies, pastries, puddings (non-milk-based), dry cakes, biscuits |
| Sugar and confectionery (sugar) | Sugar, jam, marmalade, honey, chocolate and products, candy bars, confetti/flarex/drops, boiled sweets, chewing gum, nougat, cereal bars, marzipan, syrup, water ice and ice cream |
| Vegetable oils (oil)       | Vegetable oils |
| Margarines (marga)         | Margarines, mixed dairy margarines, baking fat |
| Butter (butter)            | Butter, herbal butter, butter concentrate |
| Milk and milk beverages (milk) | Liquid milk (eg, cow, goat), processed milk (condensed, dried), whey and milk beverages |
| Cheese (chees)             | Cheese, fromage blanc, petites suisses |
| Yogurt (yogurt)            | Yogurt |
| Eggs (egg)                 | Eggs (eg, chicken, turkey, duck, goose, quail) and egg products, except if used for bread and bakery products |
| Fresh red meat (rmeat)     | Beef, veal, pork, lamb/mutton, horse, goat and offal |
| Fresh white meat (wmeat)   | Poultry, game |
| Processed meat (pmeat)     | Processed meat from red meat or poultry (eg, ham, bacon, sausages, pates, etc) |
| Fish and shellfish (fish)  | Fish and fish products, crustaceans and molluscs |
| Sauces (sauce)             | Tomato sauces, dressing sauces, mayonnaises and similars dressings |
| Tea (tea)                  | Tea (with and without caffeine), iced tea: infusion, powder, instant beverage |
| Coffee (coffee)            | Coffee (with and without caffeine): infusion, powder, instant beverage |
| Soft drinks (soda)         | Carbonated/soft/isotonic drinks, diluted syrups |
| Fruit/vegetable juices (juice) | Fruit and/or vegetable juices and nectars, freshly squeezed juices: pure or diluted with water |
| Alcohol (alcoh)            | Total ethanol intake from all beverages |

### Parameter estimation and posterior classification
Parameter estimates and posterior probability of class membership were obtained using Latent GOLD 3.0 and 4.0.16 Software default values were used for posterior mode estimation and to choose starting values for all models.12 16 Posterior probabilities were used to assign women to their most likely class.

### Clustering criterion and choice of covariance matrix
Non-parametric cluster analysis can be done using optimisation methods that partition a sample into \( K \) clusters by minimising some criterion. For example, the K-means method that has been used by other authors to find dietary patterns9 minimises the multivariate analogue of the within-cluster sums of squares. There is a correspondence between the choice of the minimising criterion in the non-parametric approach and the parameterisation of the covariance matrix in an FMM. The K-means method implicitly assumes that food consumptions have constant variances within and across clusters and that they are uncorrelated within clusters. This assumption is equivalent to restricting the covariance matrix of the \( k \)th class in an FMM, \( \Sigma_k \), such that \( \Sigma_k = \sigma^2 I_k \), where \( I \) is the identity matrix.17 Our previous work has shown that these assumptions are too restrictive for human food consumption.12 Thus, we compared various structures for the covariance matrix \( \Sigma_k \) to make an empirical choice for its form using an approach similar to Fahey et al.12 Included for comparison were models with food consumption variances allowed to vary within and/or across classes as well as models with or without covariances among food consumptions.

### Model comparison and number of classes (parity)
The Bayesian information criterion (BIC) was used to compare models that differed in covariance matrix structure and in number of fitted classes15 18 and to choose a final model for further analysis. The number of classes, \( K \), was determined as follows. We first fit the trivial 1-class model, where all
individuals belong to the same class. Class parity was successively increased by 1 in each subsequent model until the value of the BIC ceased to monotonically decrease or until parity reached 10. This value was chosen as a maximum to ensure substantial dimension reduction from 24 food groups to the number of latent classes.

Covariate adjustment
Food consumptions and class prevalence were conditioned on three categorical covariates: region, age, and total energy intake as per Fahey et al. EPIC recruitment centres were categorised according to their location in one of five regions differing in proximity, type of DQ or cohort, and between-country variation in food consumption: France or Spain; Italy or Greece; Denmark, Norway or Sweden; Germany; the Netherlands or the UK general population; UK health-conscious population. Age at recruitment was categorised into four groups: <35, 35–54, 55–64 and ≥65 years. Total energy intake (kcal) was estimated from the country-specific DQs and categorised into EPIC-wide quintiles.

Model interpretation and validation
To interpret results for the final model, predicted class prevalence and food consumption means were computed as functions of estimated parameters. Dietary patterns were interpreted by examining the predicted geometric mean consumption of each food group by class. To describe dietary patterns in terms of the covariates age, energy and body mass index (BMI), and reflect an individual’s partial contribution to each class, mean or percentage values were computed by weighting them by the posterior probabilities. Regional distribution within classes was examined after assigning women to their most likely class.

Split-sample validation of the final model was done in two ways. First, the model was fitted to the validation sample and interpreted in the same way as described above for the estimation sample. Second, the model fitted to the validation sample was used to assign women to their most likely class. Then, the parameters obtained from the model fitted to the estimation sample were used to predict a second set of assignments of women into their most likely class on the validation sample. Agreement between assignment sets was measured by two statistics. First, we calculated $\kappa$ to adjust individual agreement for chance. To do so, it was necessary to find a correspondence between the two assignment sets, because, for example, class 1 in set 1 has no intrinsic correspondence to class 1 in set 2. Therefore, $\kappa$ was computed from a cross-tabulation of assignment sets after matching the class labels based on interpretation. Second, assignment agreement among pairs was measured using the Rand index, which is invariant to permutation of class labels. It can be interpreted as the probability of a randomly chosen pair of women being assigned the same way in both sets.

RESULTS
Characteristics of the data
Analysis of the logged data showed that mean consumption of food groups in the samples was very similar to the EPIC study proper (not shown) and that by EPIC region the standard deviations of food groups were proportional to mean intakes. The proportion of the total sample that indicated non-consumption of a particular food group was greater than or equal to 15% for nine food groups: soft drinks (41%), tea (40%), butter (31%), margarine (26%), legumes (25%), fruit juice (23%), yogurt (19%), alcohol (15%) and white meat (15%). When non-consumption was low, the distributions of logged food consumptions were often approximately symmetric.

Choice of covariance matrix and class parity
These results are not reported in detail, but we note that given parity greater than 2, model fit was always better when the covariance matrix was allowed to vary by class, and structures with constant variance across classes fit poorly. Adding additional classes was more effective in improving model fit than adding correlations among food consumptions. Over the range of covariance structures and class parities examined, the smallest BIC corresponded to an eight-class model having a class-specific diagonal covariance matrix. This structure allowed variances to be class specific and assumed that food consumptions were uncorrelated within classes. All following results are based on this (final) model. Average posterior probabilities for the final model obtained after classifying each woman into her most likely class ranged from 0.903 to 0.999 and indicate classification uncertainty (table 2).

Class characteristics and dietary patterns
Class characteristics are reported in table 2 and patterns are presented in figure 1. From bottom to top, food groups in figures begin with plant products, followed by refined foods, fats, dairy products, animal products and drinks. The values plotted are, for each food group and class, predicted geometric means expressed as percentage deviation from the mean over classes. Wald test results showed very strong associations between class prevalence and age, energy and region, and between many food groups and these same covariates (p values not shown).

The shapes of the figures indicated that, with the exception of class 6, patterns exhibited relatively strong preferences for or against several foods. Four patterns were strongly associated with region. The fourth and fifth patterns comprised entirely Nordic women but were distinguished by their relative preferences (figure 1, classes 4 and 5). The former showed preferences for fish and processed meat and avoidance of tea and soft drinks. The latter showed strong preferences for fresh meat, tea and alcohol and avoidance of cakes and added fats. The seventh class comprised women from France and Spain. Their profile resembled a ‘Mediterranean’ diet (figure 1, class 7). The eighth pattern included the youngest women (table 2, figure 1, class 8) and had obvious vegetarian tendencies. Ninety per cent of these women were from the UK health-conscious cohort and 10% from other regions.

The sixth pattern was notable for having all food consumptions below the average (with the exception of vegetables) and for not having strong food preferences (figure 1, class 6). This class had the highest proportion of women with reported energy intake below the EPIC-wide second quintile. They were not the smallest women though, and three other classes had more women with normal or low BMI (table 2). The third pattern showed preferences for dairy products, margarine and non-alcoholic drinks. Almost 60% of these women were Dutch, Germans or British. The two largest classes (1 and 2) were similar with regard to many food preferences. However, women in class 1 preferred coffee and yogurt, but avoided soft drinks, while the opposite trend was apparent for women in class 2. Class 1 was about 70% southern European, while class 2 had the most uniform geographical distribution of the eight patterns.
Model validation

Dietary patterns were also identified by fitting the final model to the validation data (results not shown). Five validation data patterns corresponded very closely to classes 4−8 in table 2, four of which had very low classification uncertainty. Two other validation data patterns agreed with estimation classes 1 and 3, each with the exception of three food groups. The profile of the last validation data pattern was dissimilar, with respect to eight food groups, to estimation class 2, which had the second highest classification uncertainty among classes (table 2).

The above correspondence between classes was used to compute κ equal to 58% between the two sets of assignments obtained on the validation sample. Individual agreement depends on matching classes in one set with classes in the other set. For example, switching the correspondence between classes will, in general, give different agreement. By considering pairs, the ARI is invariant to this problem of label switching and is therefore useful when the correspondence between class labels is subjective. The value of the ARI for the two sets of assignments on the validation sample was 37%.

DISCUSSION

The application of a mixture model to EPIC women has shown some of the advantages of this approach for dietary pattern identification. In particular, pattern prevalence is estimated directly from the model parameters. There is no need to classify individuals or to arbitrarily categorise factor scores, and individuals contribute proportionately to pattern prevalence and predicted means for all classes. Moreover, the mixture modelling analogue of the clustering criterion can be chosen objectively by comparing alternative parameterisations of the covariance matrix. In this regard, our results were consistent with the findings from an analysis of the 2000−2001 British National Diet and Nutrition Survey. Both studies indicated preference for a model allowing food variances to vary within and between classes over approaches like K-means that assume variances are constant.

Our eight-class solution included plausible dietary patterns that differed with respect to their total energy intake, their sources of protein and added fats, and their consumption of plant foods, dairy products and alcohol content. Five of these patterns constituting 40% of the data were well validated in an independent sample and were strongly related to age, energy and region. The ‘low consuming’ pattern has not been found previously in EPIC and we know of no other study reporting it in DQ data, although Fahey et al12 found a very similar pattern in British food diary data. Interpretation of this pattern as a class of under-reporters was supported by the women’s BMI. Further analysis of this class could focus on physical activity and metabolic factors.

Two other well-validated patterns comprised Nordic women only. An analysis of average food consumption measured by a dietary recall in the same EPIC centres has also distinguished between at least two patterns among Nordic women.14 Given the nature of the EPIC cohorts,15 we expected to identify Mediterranean and vegetarian dietary patterns. It was surprising, however, that Italian and Greek women were not placed in an obvious Mediterranean-type pattern.

Split-sample validation also indicated that classification could be improved for many women. In particular, the second pattern found here constituting 17% of the sample had relatively high classification uncertainty and was poorly validated. One reason for lack of validation is measurement error in food consumption. It is also likely that differential measurement error among DQs obscured pattern identification.

Several aspects of our approach deserve comment. First, an FMM takes measurement error into account by assuming that each food measures a latent variable, that is, a dietary pattern, and predicted means for all classes. Moreover, the mixture modelling analogue of the clustering criterion can be chosen objectively by comparing alternative parameterisations of the covariance matrix. In this regard, our results were consistent with the findings from an analysis of the 2000−2001 British National Diet and Nutrition Survey. Both studies indicated preference for a model allowing food variances to vary within and between classes over approaches like K-means that assume variances are constant.

Our eight-class solution included plausible dietary patterns that differed with respect to their total energy intake, their sources of protein and added fats, and their consumption of plant foods, dairy products and alcohol content. Five of these patterns constituting 40% of the data were well validated in an independent sample and were strongly related to age, energy and region. The ‘low consuming’ pattern has not been found previously in EPIC and we know of no other study reporting it in DQ data, although Fahey et al12 found a very similar pattern in British food diary data. Interpretation of this pattern as a class of under-reporters was supported by the women’s BMI. Further analysis of this class could focus on physical activity and metabolic factors.

Two other well-validated patterns comprised Nordic women only. An analysis of average food consumption measured by a dietary recall in the same EPIC centres has also distinguished between at least two patterns among Nordic women.14 Given the nature of the EPIC cohorts,15 we expected to identify Mediterranean and vegetarian dietary patterns. It was surprising, however, that Italian and Greek women were not placed in an obvious Mediterranean-type pattern.

Split-sample validation also indicated that classification could be improved for many women. In particular, the second pattern found here constituting 17% of the sample had relatively high classification uncertainty and was poorly validated. One reason for lack of validation is measurement error in food consumption. It is also likely that differential measurement error among DQs obscured pattern identification.

Several aspects of our approach deserve comment. First, an FMM takes measurement error into account by assuming that each food measures a latent variable, that is, a dietary pattern, with individual agreement depending on matching classes in one set with classes in the other set. For example, switching the correspondence between classes will, in general, give different agreement. By considering pairs, the ARI is invariant to this problem of label switching and is therefore useful when the correspondence between class labels is subjective. The value of the ARI for the two sets of assignments on the validation sample was 37%.

**Theory and methods**

**Table 2** Characteristics of latent classes in normal mixture model (EPIC, 1992−2000)

| Characteristic* | Class number |
|----------------|--------------|
|                | 1       | 2       | 3       | 4       | 5       | 6       | 7       | 8       |
| Class prevalence (% of sample) | 26     | 17     | 17     | 13     | 8      | 8      | 6      | 5      |
| Average modal posterior probability | 0.983  | 0.906  | 0.919  | 0.998  | 0.999  | 0.903  | 0.911  | 0.993  |
| Average age (yrs) | 53      | 50      | 50      | 51      | 57      | 53      | 49      | 40      |
| Distribution of energy intake (%)† | **Low to < Q1** | | | | | | | |
| Q1 to < Q2 | 17     | 19     | 16     | 28     | 10     | 35     | 18     | 27     |
| Q2 to < Q3 | 21     | 18     | 21     | 26     | 15     | 23     | 21     | 25     |
| Q3 to < Q4 | 22     | 21     | 22     | 16     | 25     | 14     | 19     | 15     |
| Q4 to high | 22     | 25     | 20     | 8      | 26     | 16     | 20     | 17     |
| Proportion with BMI < 25 kg m−2 (%) | 65     | 57      | 49     | 65      | 52      | 61      | 52      | 78      |
| Region (% of class) | | | | | | | | |
| France, Spain | 50     | 23      | 5      | 0      | 0      | 41      | 100     | 3      |
| Italy, Greece | 19     | 21      | 23     | 0      | 0      | 12      | 0       | 0      |
| Netherlands, Germany, UK general population | 23     | 32      | 42     | 0      | 0      | 24      | 0       | 7      |
| UK health conscious | 6      | 15      | 16     | 0      | 0      | 5       | 0       | 90     |
| Norway, Denmark, Sweden | 1      | 9       | 13     | 100     | 100     | 19      | 0       | 0      |

*Entries are means or percentages as indicated.
†Q1−Q4 are thresholds corresponding to EPIC-wide quintiles of total energy intake.
probabilistic classification of individuals to classes, model checking, including normality and outlier detection, cannot be evaluated using standards methods. These problems will be the focus of a subsequent paper. Fourth, owing to the size of the data, we fitted models to random samples drawn from the EPIC study. Classification could be predicted for all women using the estimated parameters.

Mixture models are a flexible alternative to other multivariate methods for finding and describing dietary patterns. The estimated patterns could be used to predict disease after hard

| Class 1 | Class 2 | Class 3 | Class 4 | Class 5 | Class 6 | Class 7 | Class 8 |
|---------|---------|---------|---------|---------|---------|---------|---------|
| JUICE  | SODA  | COFFE | FISH  | PMEAT  | YOGUR | CHEES | YOGUR |
| ALCOH | TEA  | 52    | 11    | 140   | 31    | 5     | 15    |
| SODA  | COFFE | 12    | 13    | 31    | 136   | 31    | 15    |
| TEA  | SODA  | 9     | 7     | 12    | 136   | 31    | 15    |
| FISH  | PMEAT | 11    | 14    | 31    | 136   | 31    | 15    |
| PMEAT | YOGUR | 14    | 31    | 136   | 31    | 15    | 15    |
| YOGUR | CHEES | 31    | 136   | 31    | 15    | 15    | 15    |
| CHEES | SUGAR | 31    | 136   | 31    | 15    | 15    | 15    |
| SUGAR | OIL  | 31    | 136   | 31    | 15    | 15    | 15    |
| OIL  | BUTTE| 31    | 136   | 31    | 15    | 15    | 15    |
| BUTTE| PMEAT| 31    | 136   | 31    | 15    | 15    | 15    |
| PMEAT| FISH | 31    | 136   | 31    | 15    | 15    | 15    |
| FISH | SODA | 31    | 136   | 31    | 15    | 15    | 15    |
| SODA | JUICE| 31    | 136   | 31    | 15    | 15    | 15    |

Figure 1 Predicted dietary patterns* from an eight-class normal mixture model (EPIC, 1992–2000). *Values plotted for each food group are model-predicted geometric mean consumptions expressed as percentage deviation from the mean over classes. See table 1 for food group abbreviations.
assignment of individuals to a single class or, taking classification uncertainty into account, by using the posterior probabilities directly in a Cox regression model. Future work could focus on using mixture models to identify individuals who under-report food consumption.

Author affiliations
1Biostatistics Unit, Medical Research Council, Cambridge, UK
2Data Collection and Exposure Unit, European Food Safety Authority, Parma, Italy
3Dietary Exposure and Assessment Group, IARC, Lyon, France
4Department of Methodology, University of Tilburg, The Netherlands
5German Institute of Human Nutrition, Postdam-Rehbrücke, Germany
6Department of Community Medicine, Lund University, Malmö, Sweden
7Department of Hygiene and Epidemiology, University of Athens Medical School, Greece
8INSERM ERI-20, Institut Gustave-Roussy, Villejuif, France
9AFSSA, Maisons Alfort, France
10Institute of Epidemiology, Helmholtz Centre Munich, Neuhberg, Germany
11Andalusian School of Public Health, Granada, Spain
12Cancer Registry, Azienda Ospedaliera Civile-M.P, Arezzo, Ragusa, Italy
13Institute of Community Medicine, University of Tromsø, Norway
14Department of Clinical Epidemiology, Aarhus University Hospital, Aalborg, Denmark
15National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands
16MRC Dunn Human Nutrition Unit and MRC Centre for Nutritional Epidemiology and Cancer Prevention, UK
17Division of Epidemiology, Public Health and Primary Care, Imperial College London, UK

Funding The EPIC study was supported by the ‘Europe Against Cancer’ Programme of the European Commission (SANCO); Ligue contre le Cancer (France); Société 3M (France); Mutuelle Générale de l’Éducation Nationale; Institut National de la Santé et de la Recherche Médicale (INSERM); German Cancer Aid; German Cancer Research Center; German Federal Ministry of Education and Research; Danish Cancer Society; Health Research Fund (FIS) of the Spanish Ministry of Health; Spanish regional governments of Andalusia, Asturias, Basque Country, Murcia, Navarra and ISCIII; Red de Centros RCESP, C03/09; Cancer Research UK; Medical Research Council, UK; the Stroke Association, UK; British Heart Foundation; Department of Health, UK; Food Standards Agency, UK; the Wellcome Trust, UK; Greek Ministry of Health; Greek Ministry of Education; Italian Association for Research on Cancer; Italian National Research Council; Dutch Ministry of Public Health, Welfare and Sports; Dutch Ministry of Health; Dutch Prevention Funds; UK Research Funds; Dutch ZON (Zorg Onderzoek Nederland); World Cancer Research Fund (WCRF); Swedish Cancer Society; Swedish Scientific Council; Regional Government of Skane, Sweden; and the Norwegian Cancer Society. IRW was supported by Medical Research Council grant U.1052.00.006.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. Terry P, Suzuki R, Hu FB, et al. A prospective study of major dietary patterns and the risk of breast cancer. Cancer Epidemiol Biomarkers Prev 2001;10:1281–9.
2. Terry P, Hu FB, Hansen H, et al. Prospective study of major dietary patterns and colorectal cancer risk in women. Am J Epidemiol 2001;154:1143–9.
3. Kim MK, Sasakii S, Sasazuki S, et al. Prospective study of three major dietary patterns and risk of gastric cancer in Japan. Int J Cancer 2004;110:435–42.
4. Tseng M, Breslow RA, DeVellis RF, et al. Dietary patterns and prostate cancer risk in the National Health and Nutrition Examination Survey Epidemiological Follow-up Study cohort. Cancer Epidemiol Biomarkers Prev 2004;13:71–7.
5. Hu FB, Rimm EB, Stampfer MJ, et al. Prospective study of major dietary patterns and risk of coronary heart disease in men. Am J Clin Nutr 2003;72:912–21.
6. van Dam RM, Rimm EB, Willett WC, et al. Dietary patterns and risk for type 2 diabetes mellitus in U.S. men. Ann Intern Med 2002;136:201–9.
7. Fung TT, Willett WC, Stampfer MJ, et al. Dietary patterns and the risk of coronary heart disease in women. Arch Intern Med 2001;161:1857–62.
8. Fung T, Hu FB, Fuchs C, et al. Major dietary patterns and the risk of colorectal cancer in women. Arch Intern Med 2003;163:109–14.
9. Chen H, Ward MH, Graubard BI, et al. Dietary patterns and adenocarcinoma of the esophagus and distal stomach. Am J Clin Nutr 2002;75:137–44.
10. Waijers PM, Dcké MIK, van Rossum CT, et al. Dietary patterns and survival in older Dutch women. Am J Clin Nutr 2003;83:1170–6.
11. Bamilia C, Orfanos P, Ferrari P, et al. Dietary patterns among older Europeans: the EPIC-Leiden study. Br J Nutr 2005;94:100–13.
12. Fahey MT, Thane CW, Bramwell GD, et al. Conditional Gaussian mixture modelling for dietary pattern analysis. J R Stat Soc A Ser A Stat Soc 2007;170:149–66.
13. Riboli E, Hunt KJ, Slimani N, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutr 2002;5:1113–24.
14. Slimali N, Fahey M, Welch AA, et al. Diversity of dietary patterns observed in the European Prospective Investigation into Cancer and Nutrition (EPIC) project. Public Health Nutr 2002;5:1311–28.
15. McLachlan GJ, Peel D. Finite Mixture Models. New York: Wiley, 2000.
16. Vermunt JK, Magidson J. Latent GOLD’s User’s Guide. Boston: Statistical Innovations Inc, 2000.
17. Banfield JD, Raftery AE. Model-based Gaussian and non-Gaussian clustering. Biometrics 1993;49:832–7.
18. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 2002;97:611–31.
19. Slimali N, Fahey M, Welch A, et al; EPIC Working Group on Dietary Patterns. Do dietary patterns actually vary within the EPIC study? IARC Sci Publ 2002;159:49–52.
20. Hubert LJ, Arabie P. Comparing partitions. J Classification 1985;2:193–218.
21. Fisher DG, Hoffman P. The adjusted Rand statistic: a SAS macro. Psychometrika 1986;51:417–23.