A treelet transform analysis to relate nutrient patterns to the risk of hormonal receptor-defined breast cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC)

Nada Assi1,2, Aurelie Moskal1, Nadia Slimani1, Vivian Viallon3,4,5, Veronique Chajes1, Heinz Freising1, Stefano Monni2, Sven Krueppel7, Jana Förster7, Elisabeth Weiderpass8,9,10,11, Leila Lujan-Barroso12, Pilar Amiano13,14, Eva Ardanaz13,15, Esther Molina-Montes13,16, Diego Salmeron13,17,18, José Ramón Quirós19, Anja Olsen20, Anne Tjønneland20, Christina C Dahm21, Kim Overvad21, Laure Dossus22,23,24, Agnès Fournier22,23,24, Laura Baglietto25,26, Renee Turzanski Fortner6, Rudolf Kaaks6, Antonia Trichopoulou27,28, Christina Bamia29,2,30, Maria Santucci De Magistris30, Giovanna Masala31, Claudia Agnoli32, Fulvio Ricceri33, Rosario Tumino34, H Bas Bueno de Mesquita35,36,37, Marije F Bakker38, Petra HM Peeters38, Guri Skeie8, Tonje Braaten8, Anna Winkvist39, Ingegerd Johansson40, Kay-Tee Khaw41, Nicholas J Wareham42, Tim Key43, Ruth Travis43, Julie A Schmidt43, Melissa A Merritt27, Elio Riboli37, Isabelle Romieu1 and Pietro Ferrari1,*

1International Agency for Research on Cancer, 150 Cours Albert Thomas, 69372 Lyon Cedex 08, France; 2Université Claude-Bernard Lyon 1, Villeurbanne, France; 3Université de Lyon, Lyon, France; 4Université Lyon 1, UMRSEST, Lyon, France; 5ESTAR, UMRC, Bron, France; 6Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany; 7Department of Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nutheatal, Germany; 8Department of Community Medicine, Faculty of Health Sciences, University of Tromsø, The Arctic University of Norway, Tromsø, Norway; 9Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; 10Cancer Registry of Norway, Oslo, Norway; 11Department of Genetic Epidemiology, Folkhälsoanstalt Research Center, Helsinki, Finland; 12Unit of Nutrition, Environment and Cancer, Cancer Epidemiology Research Program, Catalan Institute of Oncology (ICO), Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Spain; 13CIBER de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain; 14Public Health Division of Gipuzkoa, BioDonostia Research Institute, Health Department, San Sebastian, Spain; 15Navarre Public Health Institute, Pamplona, Spain; 16Escuela Andaluza de Salud Pública, Instituto de Investigación Biosanitaria de Granada (Granada.ibs), Granada, Spain; 17Department of Epidemiology, Murcia Regional Health Council, Murcia, Spain; 18Department of Health and Social Sciences, Universidad de Murcia, Murcia, Spain; 19Public Health Directorate, Asturias, Oviedo, Spain; 20Danish Cancer Society Research Center, Copenhagen, Denmark; 21Section for Epidemiology, Department of Public Health, Aarhus University, Aarhus, Denmark; 22Inserm, Centre for Research in Epidemiology and Population Health (CEP), Nutrition, Hormones and Women’s Health Team, Villejuif, France; 23Université Paris Sud, UMR Villejuif, France; 24IGR, Villejuif, France; 25Cancer Epidemiology Centre, Cancer Council of Victoria, Melbourne, Australia; 26Centre for Epidemiology and Biostatistics, School of Population and Global Health, University of Melbourne, Melbourne, Australia; 27Neale Health Foundation, Athens, Greece; 28Bureau of Epidemiologic Research, Academy of Athens, Athens, Greece; 29Department of Hygiene, Epidemiology and Medical Statistics, University of Athens Medical School, Athens, Greece; 30Azienda Ospedaliera Universitaria (AOU) Federico II, Naples, Italy; 31Molecular and Nutritional Epidemiology Unit, Cancer Research and Prevention Institute – ISPO, Florence, Italy; 32Epidemiology and Prevention Unit, Fondazione IRCCS, Istituto Nazionale dei Tumori, Milan, Italy; 33Unit of Cancer Epidemiology – CERMS, Department of Medical Sciences, University of Turin and Città della Salute e della Scienza Hospital, Turin, Italy; 34Cancer Registry and Histopathology Unit, ‘Civile M.P. Arezzo’ Hospital, Ragusa, Italy; 35Department for Determinants of Chronic Diseases (DCD), National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands; 36Department of Gastroenterology and Hepatology, University Medical Centre, Utrecht, The Netherlands; 37Department of Epidemiology and Biostatistics, The School of Public Health, Imperial College London, London, UK; 38Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands; 39Department of Internal Medicine and Clinical Nutrition, The Sahlgrenska Academy, Göteborg, Sweden; 40Department of Odontology, Umeå University, Umeå, Sweden; 41Department of Public Health and Primary Care, University of Cambridge School of Clinical Medicine, Cambridge, UK; 42MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge, UK; 43Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK

Submitted 9 September 2014: Final revision received 12 January 2015: Accepted 20 January 2015: First published online 23 February 2015
Nutrient patterns and breast cancer in EPIC

Abstract

Objective: Pattern analysis has emerged as a tool to depict the role of multiple nutrients/foods in relation to health outcomes. The present study aimed at extracting nutrient patterns with respect to breast cancer (BC) aetiology.

Design: Nutrient patterns were derived with treelet transform (TT) and related to BC risk. TT was applied to twenty-three log-transformed nutrient densities from dietary questionnaires. Hazard ratios (HR) and 95% confidence intervals computed using Cox proportional hazards models quantified the association between quintiles of nutrient pattern scores and risk of overall BC, and by hormonal receptor and menopausal status. Principal component analysis was applied for comparison.

Setting: The European Prospective Investigation into Cancer and Nutrition (EPIC).

Subjects: Women (n = 334,850) from the EPIC study.

Results: The first TT component (TC1) highlighted a pattern rich in nutrients found in animal foods loading on cholesterol, protein, retinol, vitamins B12 and D, while the second TT component (TC2) reflected a diet rich in β-carotene, riboflavin, thiamin, vitamins C and B6, fibre, Fe, Ca, K, Mg, P and folate. While TC1 was not associated with BC risk, TC2 was inversely associated with BC risk overall (HRQ5 < 0.89, 95% CI 0.83, 0.95, PTrend < 0.01) and showed a significantly lower risk in oestrogen receptor-positive (HRQ5 v. Q1 = 0.89, 95% CI 0.81, 0.98, PTrend = 0.02) and progesterone receptor-positive tumours (HRQ5 v. Q1 = 0.87, 95% CI 0.77, 0.98, PTrend < 0.01).

Conclusions: TT produces readily interpretable sparse components explaining similar amounts of variation as principal component analysis. Our results suggest that participants with a nutrient pattern high in micronutrients found in vegetables, fruits and cereals had a lower risk of BC.

Breast cancer (BC) remains the highest incident cancer affecting women worldwide, with almost 1,670,000 cases registered in 2012. It is a major public health concern with mortality from BC accounting for over 522,000 deaths in 2012. It is a major public health concern with almost 198,000 deaths in Western countries and about 324,000 in less developed regions. Established BC risk factors include age, genetic mutations, ethnicity, height, reproductive history, breast-feeding, hormone therapy and diabetes. Besides these, a number of modifiable lifestyle factors are associated with BC such as smoking, body fat and obesity, physical inactivity, alcohol consumption and diet. Diet has been suggested to account for up to 25–40% of preventable causes of cancers; in particular, 50% of BC deaths are linked to diet, although the consensus around this estimate is not unanimous. Standard approaches customarily evaluate the risk of BC associated with one or a group of dietary items, i.e. food(s) or nutrient(s). Nevertheless, associations between diet and disease might be missed when one parses the effect of a limited list of dietary constituents. Although this simplified approach of examining a single food or nutrient at a time has led to important results on the role of an individual dietary component in BC aetiology, such as fibre from vegetables, alcohol, tea consumption, folate and other micronutrients, research might benefit from a more comprehensive approach by exploring BC aetiology in terms of an integrated ensemble of dietary characteristics.

To capture the complexity of individuals’ dietary habits, dietary pattern analysis has emerged as a complementary holistic methodology focusing on sets of dietary variables and addressing their inherent interrelations. This approach is justified as components of dietary exposure are not independent and because it allows to account for complex relationships between nutrients in biological pathways. In addition, BC is a multifactorial disease, the aetiology of which possibly depends on more than a restricted list of dietary items.

Recent investigations carried out in Western populations have consistently identified two main dietary patterns: the prudent/healthy and the Western/unhealthy. While diet is related to cultural background, common nutrients are present in different combinations of foods; hence looking into diet–disease associations on the nutrient scale could lead to the identification of specific nutritional profiles relevant to BC aetiology.

In the present study, nutrient patterns within the European Investigation into Cancer and Nutrition (EPIC) were related to BC risk. Nutrient patterns were obtained by applying the treelet transform (TT) that has recently been introduced into nutritional epidemiology and the well-known principal component analysis (PCA) was used for the sake of comparison. TT yields sparse components and reveals the intrinsic structure of the data, thus simplifying interpretability. Aspects related to the application of TT to dietary data in the context of a multi-centre study are described and discussed. The association between nutrient...
patterns and BC was evaluated using all BC cases and by taking into account the heterogeneity of BC subtypes by integrating information on menopausal and hormone receptor status.

Materials and methods

Study population and exclusion criteria

EPIC is a large prospective cohort of 521,330 healthy men and women designed to evaluate the relationships between dietary habits, nutrition, lifestyle factors and the incidence of cancer. The EPIC cohort includes participants from twenty-three centres in France, Germany, Denmark, Sweden, Norway, Greece, Italy, the Netherlands, Spain and the UK. In most centres, participants were recruited from the general population, the exceptions being France (women were enrolled from a national health insurance scheme covering teachers in the French education system employees), Italy (Turin and Ragusa: blood donors; Florence: screening programme participants), Spain (blood donors) and the Netherlands (Utrecht: women participating in BC screening). In Norway, only women from the general population were recruited and in the UK, one-half of the cohort (the Oxford sub-cohort) consisted of ‘health-conscious’ individuals from England, Wales, Scotland and Northern Ireland. The design of the study and its rationale along with the recruitment process have been described elsewhere. Among the 521,330 EPIC participants, men were first removed (n 153,427). Women with prevalent cancers at any site at baseline (other than non-melanoma skin cancer; n 19,853) or lost to follow-up (n 2892) were excluded, as were women who did not complete any dietary questionnaire (n 3315) and those who did not complete a lifestyle questionnaire (n 26). To avoid including extreme values, participants in the top and bottom 1% of the distribution of the ratio of reported total energy intake to energy requirement (n 6753) were excluded. After exclusion of non-first BC cases (n 2) the cohort included 335,062 women upon whom the dietary patterns were derived. An additional number of women (n 212) with missing information on BC status were excluded, which left 334,850 women retained for the statistical analyses.

Cancer assessment

Incident BC cases were identified through population cancer registries (Denmark, Italy, Netherlands, Norway, Spain, Sweden and UK) or through active follow-up (France, Germany, Naples and Greece), as detailed in Ferrari et al. Information on oestrogen receptor (ER) and progesterone receptor (PR) statuses was provided by each centre on the basis of pathology reports.

Dietary assessment

Long-term usual dietary intake was assessed at baseline using country-specific and validated dietary questionnaires (self-administered FFQ, semi-quantitative or interviewer-performed). In the validation studies, the dietary questionnaires were compared with a reference method which was in most centres 24 h dietary recalls, except in Sweden and the UK, where food records were used. Generally, the correlation coefficients were between 0·40 and 0·70 for all nutrients examined which was considered satisfactory. Individual intakes of twenty-three nutrients and total energy were estimated using a common food composition database, the EPIC Nutrient Database (ENDB), which was compiled from national food composition databases of the ten countries represented in EPIC following standardized procedures.

Lifestyle questionnaires

Information on sociodemographic characteristics, including education, and lifestyle habits such as levels of physical activity, tobacco smoking, as well as consumption of alcohol and drinking habits, were collected using lifestyle questionnaires. In addition, anthropometric measures and past medical information were gathered at recruitment.

Nutrient pattern assessment

EPIC-wide nutrient patterns were derived among female participants in EPIC using TT in the main analysis and PCA in the sensitivity analysis. The sample covariance matrix of twenty-three log-transformed nutrient densities, computed using alcohol-free energy intake, was consistently used. The use of the sample covariance matrix allows variability to be informative in the pattern discovery phase. The distribution of nutrient consumption tends to be log-normal and may not be best described by the mean and variance on the original scale. Moreover micro- and macronutrients are expressed on different scales (micrograms, milligrams or grams). The nutrient densities were log-transformed to remove scale dependence and render their variance (or covariance) independent of the unit of measure. In line with previous work, alcohol intake was not included and was considered as a lifestyle factor. Total fat was divided into MUFA, PUFA and SFA, and total carbohydrates were broken down into starch and sugar. The micro- and macronutrients studied were Ca, β-carotene, cholesterol, MUFA, PUFA, SFA, Fe, fibre, K, Mg, P, protein, riboflavin, starch, sugar, thiamin, vitamins B₆, B₁₂, C, D, E and folate. The list of nutrients as well as the approach described for their handling is consistent with the nutrient patterns initiative within EPIC described by Moskal et al.

Pattern extraction

The TT method used for pattern extraction is described in detail by Gorst-Rasmussen and co-workers. Briefly, TT is a dimension reduction technique aimed at converting a set of observations of possibly correlated variables into orthogonal components. TT scores, corresponding to
the projection of data onto components, generally have a small degree of correlation, unlike PCA scores that are always uncorrelated. The number of retained components was based on the percentage of explained variance, scree plots and interpretability. The nutrient patterns were defined after the inspection of factor loadings, i.e. eigenvectors, expressing the contribution of nutrients to a given component. Score variables were determined for each component of TT and reflected adherence to a given type of diet/nutrient profile. TT combines the quantitative pattern extraction capabilities of PCA with interpretational advantages of hierarchical clustering of variables. In TT, the two variables displaying the highest correlation (or covariance) are identified, and a PCA is performed on them. The two variables are then replaced with the score of their first PCA component and a merge is indicated in the cluster tree. This operation is re-iterated until all variables have joined the cluster tree. In this way, TT produces a hierarchical grouping of variables which may reveal intrinsic characteristics of data structure. An important feature of TT is that it introduces sparsity into factors, making many factors loadings exactly equal to zero, potentially simplifying the interpretation. Alongside the cluster tree dendrogram produced by TT (as exemplified in Fig. 1), TT yields a coordinate system for the data at each level of the cluster tree. Selecting a cluster tree level (cut-level) for the TT cluster tree amounts to choosing the level of detail desired in the dimension reduction of data. More variation can be explained at the cost of factor sparsity when the cluster tree is cut near its ‘root’. If the data have \( p \) variables, there are \( p - 1 \) possible cut-levels. After deciding on the number of components to retain, we performed a tenfold cross-validation to identify the optimal cut-level, i.e. the point at which increasing the cut-level does not substantially increase the variation of the retained patterns. We also performed a sensitivity analysis to assess the effect of different cut-levels.

Consistently, a PCA was also applied for the sake of comparison. This technique yields orthogonal components that are invariant to the number of subsequent components retained. PCA identifies the best linear combination of the variables accounting for the most variance observed in the original data, producing components with uncorrelated scores. Results of TT analysis were compared with findings obtained with the more classic PCA method. To make the comparison easier, and because TT returns sparse vectors, only nutrients with absolute loadings greater than 0.2 were retained to identify a given pattern in PCA.

**Patterns and breast cancer risk**

The associations between nutrient patterns and risk of BC were investigated by using Cox proportional hazards regression models to estimate hazard ratios (HR) and 95% confidence intervals. Breslow’s method was adopted for handling time ties. The time at entry was the age at recruitment and the time of exit was the age at cancer diagnosis, death, loss or end of follow-up, whichever happened first. Models were stratified by centre, to control for differences in questionnaire designs, follow-up procedures and other centre-specific effects, as well as for age at recruitment (1-year categories). Analyses were performed by considering the TT (and principal component (PC)) scores in quintiles to appreciate potential departure from linearity. Statistical analyses were adjusted for baseline menopausal status (premenopausal and perimenopausal (reference) or postmenopausal and women who underwent an ovariectomy), baseline alcohol intake (never drinkers (reference), former drinkers, drinkers only at recruitment, lifetime drinkers, unknown), height (continuous), BMI (below (reference) or above 25 kg/m\(^2\)), schooling level (none, primary (reference), technical/
professional/secondary, longer education, unknown/unspecified), age at first full-term pregnancy (nulliparous (reference), ≤21 years, 21–30 years, >30 years, unknown or missing), age at menarche (≤12 years (reference), 12–14 years, >14 years, missing), age at menopause (≤50 years (reference), >50 years, premenopausal or missing), use of hormone replacement therapy (never (reference), ever, unknown), level of physical activity (categorical, metabolic equivalents of task (MET)/h: inactive (reference), moderately inactive, moderately active, active, unknown) and alcohol-free energy (continuous). Use of oral contraceptive pills (never (reference), ever or unknown) and smoking status (never smokers (reference), ex-smokers, current smokers, unknown) were evaluated but not retained in the final models, due to limiting confounding exerted by these variables.

The overall significance of a score variable in categories was evaluated using the likelihood ratio test statistics ($P_{LRT}$) with $df = 4$. Additionally, $P$ values for trend ($P_{trend}$) were computed by modelling a score variable with quintile-specific medians as continuous. The association between nutrient patterns and BC risk was evaluated in pre- and postmenopausal women and according to BC hormonal receptor status (ER/PR status). Interaction between menopausal status and pattern scores was explored. In addition, tests of heterogeneity of associations between nutrient patterns and BC risk was evaluated in pre- and postmenopausal women and according to BC hormonal receptor status (ER/PR status). Interaction between menopausal status and pattern scores was explored. In addition, tests of heterogeneity of associations according to receptor status were performed using the data-augmentation method ($P_{LRT}$) by comparing the difference in the log likelihood between a model with receptor status-specific variable and a model with a single HR estimate for the two categories of receptor status to a $X^2$ distribution with $df = 1$ ($P_{heterogeneity}$).

Departure from linearity was explored with restricted cubic splines ($P_{med}$), using five knots corresponding to the 1st and 99th percentiles and medians of the centred scores of quintiles 1, 3 and 5. Spline plots were produced by taking the log-transformed nutrient densities. The loadings of components 1 and 2 are shown in Table 2. TT yielded a dendrogram shown in Fig. 1, with numbered nodes indicating the four highest-variance factors, where factors 1 and 2 were identified as the first two components after setting the cut-level to 16. Loadings of components 1 and 2 are shown in Table 2. TT yielded a dendrogram shown in Fig. 1, with numbered nodes indicating the four highest-variance factors, where factors 1 and 2 were identified as the first two components after setting the cut-level to 16. Departures from linearity were assessed via an evaluation of the joint significance of variables other than the linear one included in the model using Wald's test on $df = 3$. Associations between all of the PC and BC were investigated in a consistent way.

Statistical tests were two-sided, the per-test significance level was set to $\alpha = 0.05$. All analyses were performed using the SAS statistical software package version 9.3; the ‘t’ package in the STATA statistical software package release 12 was used to perform TT.

### Results

A total of 11 576 BC cases were recorded in 11.5 years of median follow-up time and 3 670 439 person-years. Based on the information obtained at baseline, 2827 cases were premenopausal, 5872 were postmenopausal, 2548 were perimenopausal and 328 cases had a bilateral ovariectomy. Among incident cases, information on hormone receptor status for ER and PR was available only in 62% and 52% of total cancer cases, respectively, and was distributed as follows: 81% ER+ and 19% ER- tumours and 63% PR+ and 37% PR- tumours. Descriptive information of the study sample by EPIC country is available in Table 1.

### Identification of nutrient patterns

Inspection of factor loadings allowed an initial identification of four nutrient patterns with TT, explaining 62% of total nutrient intake variability within individuals. After a tenfold cross-validation along with a sensitivity analysis strategy and after evaluating the interpretability of each pattern, we chose to cut the cluster tree at level 16. Loadings of components 1 and 2 are shown in Table 2. TT yielded a dendrogram shown in Fig. 1, with numbered nodes indicating the four highest-variance factors, where factors 1 and 2 were identified as the first two components after setting the cut-level to 16 indicated by the dashed line. This dendrogram reveals the correlation structure of the log-transformed nutrient densities. The first treelet component (TC1) loaded on vitamin D, vitamin B12, cholesterol, protein and retinol, suggesting a diet rich in animal products. The second treelet component (TC2)

### Table 1 Numbers of women and breast cancer (BC) cases (first tumours only) in the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort by country

<table>
<thead>
<tr>
<th>Country</th>
<th>No. of women</th>
<th>Person-years</th>
<th>No. of BC cases</th>
<th>Follow-up time (years)*</th>
<th>Age at enrolment (years)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>67 356</td>
<td>699 216</td>
<td>3187</td>
<td>11.8</td>
<td>51.5</td>
</tr>
<tr>
<td>Italy</td>
<td>30 498</td>
<td>341 417</td>
<td>1047</td>
<td>11.7</td>
<td>50.9</td>
</tr>
<tr>
<td>Spain</td>
<td>24 846</td>
<td>299 575</td>
<td>495</td>
<td>12.6</td>
<td>47.7</td>
</tr>
<tr>
<td>UK general population</td>
<td>17 145</td>
<td>200 812</td>
<td>719</td>
<td>12.3</td>
<td>55.6</td>
</tr>
<tr>
<td>UK health-conscious</td>
<td>35 368</td>
<td>385 533</td>
<td>761</td>
<td>11.3</td>
<td>41.5</td>
</tr>
<tr>
<td>Netherlands</td>
<td>26 839</td>
<td>315 554</td>
<td>916</td>
<td>12.2</td>
<td>52.7</td>
</tr>
<tr>
<td>Greece</td>
<td>15 224</td>
<td>148 594</td>
<td>198</td>
<td>10.7</td>
<td>53.6</td>
</tr>
<tr>
<td>Germany</td>
<td>27 390</td>
<td>272 011</td>
<td>834</td>
<td>10.9</td>
<td>48.4</td>
</tr>
<tr>
<td>Sweden</td>
<td>26 339</td>
<td>349 110</td>
<td>1095</td>
<td>13.9</td>
<td>50.6</td>
</tr>
<tr>
<td>Denmark</td>
<td>28 933</td>
<td>316 601</td>
<td>1340</td>
<td>11.6</td>
<td>56.3</td>
</tr>
<tr>
<td>Norway</td>
<td>35 152</td>
<td>342 195</td>
<td>984</td>
<td>10.1</td>
<td>48.0</td>
</tr>
<tr>
<td>Total</td>
<td>334 850</td>
<td>3 670 439</td>
<td>11 576</td>
<td>11.5</td>
<td>51.0</td>
</tr>
</tbody>
</table>

*Median is given for follow-up time and age at enrolment.
presented high positive loadings on β-carotene, thiamin, fibre, vitamin C and folate, and singled out some nutrients with mild loadings (<0.2), i.e. Fe, Ca, K, Mg and P (Table 2). TC2 may evoke a diet rich in vegetables, fruits and cereals. While the third treelet component (TC3) was largely driven by vitamin D, the fourth treelet component (TC4) was less straightforward to characterize, as displayed in the online supplementary material, Supplemental Table 1. Distributions of known risk factors for BC by quintiles of TT scores for the first two components are displayed in Table 3.

PC loadings are displayed in the online supplementary material, Supplemental Table 2. PCA produced patterns similar to TT with respect to the amount of variability explained and the nutrients contributing to the definition of each component: with PC1 displaying high loadings for cholesterol, retinol, vitamin B12 and vitamin D and negative loadings for vitamin C and β-carotene; and PC2 suggesting a micronutrient-dense pattern rich in fruits, vegetables, plant foods and dairy. The first two components (in TT and PCA) explained the most variability and were the most informative with respect to capturing meaningful nutrient patterns, and thus were further related to BC risk in disease models.

**Nutrient patterns and breast cancer risk**

Scores of nutrient patterns were related to BC risk in disease models. TC1 showed no statistically significant association with BC risk with HR_{TC1 Q5 v. Q1} = 1.05 (95% CI 0.98, 1.13, P_{trend} = 0.36, P_{LRT} = 0.39), while TC2 was significantly associated with BC risk with HR_{TC2 Q5 v. Q1} = 0.89 (95% CI 0.83, 0.95, P_{trend} < 0.001, P_{LRT} = 0.02), as shown in Table 4. The relationship between TT scores and BC risk was modelled through restricted cubic splines (RCS) and is presented in Fig. 2. Overall, there was a significant progressive decrease in BC risk for the second component. TC2 scores showed a linear decrease in BC risk (RCS_{TC2} P_{trend} = 0.02). However, no departure from linearity was observed (P_{Wald non-linearity} = 0.94 and 0.77, respectively, in TC1 and TC2; Fig 2). Analyses of interaction between TC (or PC) scores and menopausal status were not statistically significant (results not shown).

**Hormonal receptor status**

In ER+ tumours, no significant association with BC risk was observed for TC1 and TC2 scores (Table 4). For ER+ tumours there was a decrease in BC risk in the fourth and fifth quintiles of TC2 scores with HR_{Q4 v. Q1} = 0.90 (95% CI 0.83, 0.99) and HR_{Q5 v. Q1} = 0.89 (95% CI 0.81, 0.98, P_{trend} = 0.02; Table 4). Regarding PR+ tumours (see online supplementary material, Supplemental Table 3), the second component TC2 showed a decreased BC risk with HR_{Q5 v. Q1} = 0.84 (95% CI 0.72, 0.98). For PR+ tumours, TC2 was linked with a decreased BC risk in participants in the fifth quintile with HR_{Q5 v. Q1} = 0.87 (95% CI 0.77, 0.98). No significant association was seen for ER+/PR- tumours (Table 5). TC2 was linked with a decreased BC risk trend in ER+/PR+ tumours with HR_{Q5 v. Q1} = 0.86 (0.76, 0.98, P_{trend} < 0.01; Table 5). Tests of heterogeneity yielded no significant results.

PCA derived components displayed a significant increase in BC risk for PC1 in participants in the highest quintile and a decreasing trend of BC risk for PC2, as shown in the online supplementary material, Supplemental Table 4 and Supplemental Fig. 1. Results of associations of PC with tumours by hormone receptor status are displayed in the online supplementary material, Supplemental Tables 4 and 5.

**Discussion**

In the present study, the role of nutrient patterns in the aetiology of BC was explored through the use of TT, a multivariate method recently introduced to the landscape of nutritional epidemiology. The association was evaluated in the context of the EPIC study, characterized by large variability of dietary habits and by a large number of incident cancer cases across participating centres.

In recent years, dietary pattern analysis has emerged as a promising technique, complementary to methods focusing on individual foods or food components, to investigate the relationships between diet and risk of disease. A systematic review and meta-analysis on dietary patterns in BC aetiology selected eighteen
relevant studies from case-control and cohort studies that used combinations of foods and micronutrients to identify dietary patterns. Two a posteriori defined patterns emerged consistently: the Western/unhealthy (in seventeen studies) and the prudent/healthy (eighteen studies). In the aforementioned meta-analysis, the prudent/healthy dietary pattern, rich in intakes of vegetables, leafy vegetables, legumes and fish, was associated to decreased BC risk (relative risk comparing top vs. bottom categories = 0.89, 95% CI 0.82, 0.99), while the Western/
unhealthy pattern, characterized by intakes of high-fat dairy products, red meat, processed meats and French fries, was not associated with BC risk. A recent study of the California Teachers Cohort identified a plant-based pattern, which was related to a reduction of BC risk \(^{(67)}\). In parallel, increasing evidence is accumulating that adherence to the \textit{a priori} defined Mediterranean pattern is associated with a decreased BC risk \(^{(66-70)}\), although results from these studies are not totally consistent, particularly for premenopausal women \(^{(70,71)}\).

The dimension reduction techniques used herein were applied to nutrient densities. Nutrients are present in different combinations of foods, are less country-specific and are directly involved in biological reactions \(^{(72)}\). By exploring macro- and micronutrients, the present study aimed to provide an exhaustive representation of individuals’ diet. Log-transformation was used to address scaling issues that can arise because macro- and micronutrients are expressed in different units. In this way, the variance and the components’ decomposition are invariant to the unit of measure. Dietary normalization was achieved using equal energy, i.e. by dividing nutrient intakes by energy intake, minus energy from alcohol intake \(^{(44)}\). Most nutrients are associated with total energy because either they contribute to total energy directly or because people with higher energy values tend to display larger intakes of specific nutrients \(^{(44,73)}\).

The first two patterns were retained as they were the most interpretable and depicted realistic nutrient patterns that could ultimately be linked with disease risk. The first pattern identified a diet characterized by animal products as opposed to a vegetarian diet, and was associated with a non-significant increase of 5% in BC risk (TT). TC1 was quite comparable to a Western pattern. Two recent reviews on dietary patterns and BC \(^{(74,75)}\) showed that diets rich in high-fat foods and processed meats were associated with an increased BC risk, although the findings described in both reviews have not been conclusive in this respect with most results reporting a positive association between Western-like dietary pattern and BC being not statistically significant \(^{(74,75)}\). In our study, the micronutrient-dense pattern characterized by a diet rich in vitamins and minerals, akin to a prudent pattern, was associated with an 11% reduction in BC risk (TT), in line with previous findings \(^{(33,74,75)}\). The protective effect may come from the anti-carcinogenic properties of nutrients such as \(\beta\)-carotene, vitamins C and E, that may exert an antioxidant effect on oestrogen metabolism and reduce cell proliferation \(^{(75)}\). The TT components were
Analyses carried out by hormonal receptor status showed differential association in pre- and postmenopausal women. The present study relied on dietary questionnaires to assess nutrient intakes, which are prone to measurement errors and may lack information on some relevant nutrients. Questionnaires were country-specific, potentially

tumours. These results are complementary to previous literature findings on dietary patterns and hormonal defined risk of BC. Indeed, Fung et al. found that a prudent dietary pattern was linked with decreased ER" risk (relative risk = 0.62, 95% CI 0.45, 0.91) ER"/PR" tumour risk was reduced in postmenopausal women among participants in the highest quintiles of a plant-based pattern and an a priori defined Mediterranean diet by 34% and 20%, respectively. Results from the Pooling Project of Prospective Studies of Diet and Cancer found a protective association between total fruit or fruit and vegetable consumption in ER" tumours but not in ER" tumours or overall BC risk.

Whereas a large portion of the scientific literature on dietary patterns has used factor analysis or principal component factor analysis (PCA), the current paper promotes the use of TT. While PCA produces patterns that are eigenvectors of a covariance/correlation matrix of starting variables, TT is a multivariate technique that yields components by aggregating variables according to covariance/correlation, while at the same time exploring the clustering structure of variables, combining features of PCA with those of cluster analysis. Eventually, TT produces a cluster tree revealing the hierarchical grouping structure of variables. The dendrogram allows a visual inspection of the way different nutrients cluster, possibly easing interpretability of patterns. In addition, loadings are sparse, i.e. some of them are equal to zero as they do not pertain to the clustering node of the component so that a limited number of variables contributes to each treelet component.

In line with other clustering techniques (PCA), TT users are confronted with subjective decisions to select the appropriate cut-level for the cluster tree. Information on the grouping structure of variables that have joined (or not) the tree are specific to each level of the TT tree. By choosing a cut-level, the user decides on how much information to extract and the degree of sparsity of the components. If the tree is cut near the root, all nutrient variables join the tree. The information would be comparable to PCA output, i.e. all variables would contribute to treelet components. If the tree is cut closer to the leaves, i.e. when the cut-level is lower, loadings are sparse as many are equal to zero, possibly making the interpretation easier. By contrast, this may lead to components that do not capture dietary complexity and are therefore not informative. As pointed out by Meinshausen and Bühlmann, the use of TT leads to a trade-off between amount of variability explained and sparsity. The objective is to ‘make the results as sparse as possible but not any sparser’. To identify an optimal cut-level, cross-validation can be used. Once the cut-level is chosen, the loadings computed are invariant to the number of components to be retained, hence keeping n components is an a priori parameter to be specified in the cross-validation step.

The analyses carried out by hormonal receptor status showed that the second TT nutrient pattern was related to a significant decrease in BC risk for ER" and PR" and ER"/PR" highly correlated with those of PCA (\(\rho_{\text{PC1,PC2}} = 0.91\), \(\rho_{\text{PC2,PC2}} = 0.86\)). TT and PCA provided overall consistent findings in terms of pattern identification and amount of total variability explained. Further analyses were conducted by menopausal status at cohort enrolment, showing no differential association in pre- and postmenopausal women.

Fig. 2 Relationship between nutrient patterns derived from treelet transform and breast cancer risk (HR; ——, hazard ratio (HR); ———, associated 95% CI), obtained by using restrictive cubic splines with values of 1st and 99th percentiles and medians of quintiles 1, 3 and 5 used as knots, among 334,850 women in the European Prospective Investigation into Cancer and Nutrition (EPIC): (a) first treelet component (TC1), \(P_{\text{non-linearity}} = 0.94\), \(P_{\text{trend}} = 0.88\); (b) second treelet component (TC2), \(P_{\text{non-linearity}} = 0.77\), \(P_{\text{trend}} = 0.02\). Models were stratified by study centre and age in 1-year categories and adjusted for baseline tumour risk was reduced in postmenopausal women among participants in the highest quintiles of a plant-based pattern and an a priori defined Mediterranean diet by 34% and 20%, respectively. Results from the Pooling Project of Prospective Studies of Diet and Cancer found a protective association between total fruit or fruit and vegetable consumption in ER" tumours but not in ER" tumours or overall BC risk.

Whereas a large portion of the scientific literature on dietary patterns has used factor analysis or principal component factor analysis (PCA), the current paper promotes the use of TT. While PCA produces patterns that are eigenvectors of a covariance/correlation matrix of starting variables, TT is a multivariate technique that yields components by aggregating variables according to covariance/correlation, while at the same time exploring the clustering structure of variables, combining features of PCA with those of cluster analysis. Eventually, TT produces a cluster tree revealing the hierarchical grouping structure of variables. The dendrogram allows a visual inspection of the way different nutrients cluster, possibly easing interpretability of patterns. In addition, loadings are sparse, i.e. some of them are equal to zero as they do not pertain to the clustering node of the component so that a limited number of variables contributes to each treelet component.

In line with other clustering techniques (PCA), TT users are confronted with subjective decisions to select the appropriate cut-level for the cluster tree. Information on the grouping structure of variables that have joined (or not) the tree are specific to each level of the TT tree. By choosing a cut-level, the user decides on how much information to extract and the degree of sparsity of the components. If the tree is cut near the ‘root’, all nutrient variables join the tree. The information would be comparable to PCA output, i.e. all variables would contribute to treelet components. If the tree is cut closer to the ‘leaves’, i.e. when the cut-level is lower, loadings are sparse as many are equal to zero, possibly making the interpretation easier. By contrast, this may lead to components that do not capture dietary complexity and are therefore not informative. As pointed out by Meinshausen and Bühlmann, the use of TT leads to a trade-off between amount of variability explained and sparsity. The objective is to ‘make the results as sparse as possible but not any sparser’. To identify an optimal cut-level, cross-validation can be used. Once the cut-level is chosen, the loadings computed are invariant to the number of components to be retained, hence keeping n components is an a priori parameter to be specified in the cross-validation step.

The analyses carried out by hormonal receptor status showed that the second TT nutrient pattern was related to a significant decrease in BC risk for ER" and PR" and ER"/PR" highly correlated with those of PCA (\(\rho_{\text{PC1,PC2}} = 0.91\), \(\rho_{\text{PC2,PC2}} = 0.86\)). TT and PCA provided overall consistent findings in terms of pattern identification and amount of total variability explained. Further analyses were conducted by menopausal status at cohort enrolment, showing no differential association in pre- and postmenopausal women. Analyses carried out by hormonal receptor status showed that the second TT nutrient pattern was related to a significant decrease in BC risk for ER" and PR" and ER"/PR" highly correlated with those of PCA (\(\rho_{\text{PC1,PC2}} = 0.91\), \(\rho_{\text{PC2,PC2}} = 0.86\)). TT and PCA provided overall consistent findings in terms of pattern identification and amount of total variability explained. Further analyses were conducted by menopausal status at cohort enrolment, showing no differential association in pre- and postmenopausal women. Analyses carried out by hormonal receptor status showed that the second TT nutrient pattern was related to a significant decrease in BC risk for ER" and PR" and ER"/PR"
Table 5 Hazard ratios (HR) and 95% confidence intervals for breast cancer (BC) by quintiles of pattern scores (first and second components of treelet transform, cut-level 16) for oestrogen receptor-positive + progesterone receptor-positive (ER+/PR+) and oestrogen receptor-negative + progesterone receptor-negative (ER-/PR-) tumours in 334 850 women in the European Prospective Investigation into Cancer and Nutrition (EPIC)

<table>
<thead>
<tr>
<th>Model*</th>
<th>TC1</th>
<th>TC2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Person-years</td>
<td>No. of BC cases</td>
</tr>
<tr>
<td>ER+/PR+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q1</td>
<td>723 508</td>
<td>568</td>
</tr>
<tr>
<td>Q2</td>
<td>728 884</td>
<td>811</td>
</tr>
<tr>
<td>Q3</td>
<td>725 948</td>
<td>750</td>
</tr>
<tr>
<td>Q4</td>
<td>726 667</td>
<td>751</td>
</tr>
<tr>
<td>Q5</td>
<td>717 569</td>
<td>773</td>
</tr>
<tr>
<td>ER-/PR−</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q1</td>
<td>720 830</td>
<td>172</td>
</tr>
<tr>
<td>Q2</td>
<td>724 871</td>
<td>235</td>
</tr>
<tr>
<td>Q3</td>
<td>722 003</td>
<td>207</td>
</tr>
<tr>
<td>Q4</td>
<td>722 988</td>
<td>222</td>
</tr>
<tr>
<td>Q5</td>
<td>713 798</td>
<td>214</td>
</tr>
</tbody>
</table>

TC1, treelet component 1; TC2, treelet component 2; Q1, quintile 1; Q2, quintile 2; Q3, quintile 3; Q4, quintile 4; Q5, quintile 5; Ref., reference category.

* Models were stratified by study centre and age in 1-year categories and adjusted for baseline menopausal status (premenopausal and perimenopausal (reference) or postmenopausal and women who underwent an ovariectomy), baseline alcohol intake (never drinkers (reference), former drinkers, drinkers only at recruitment, lifetime drinkers, unknown), height (continuous), BMI (below (reference) or above 25 kg/m²), schooling level (none, primary (reference), technical/professional/secondary, longer education, unknown/unknown), age at first full-term pregnancy (nulliparous (reference), ≤21 years, 21–30 years, >30 years, unknown or missing), age at menarche (≤12 years (reference), 12–14 years, >14 years, missing), age at menopause (≤50 years (reference), >50 years, premenopause or missing), use of hormone replacement therapy (never (reference), ever, unknown), level of physical activity (inactive (reference), moderately inactive, moderately active, active, unknown) and alcohol-free energy (continuous).

† $P_{\text{LRT}}$, P values for the likelihood ratio test (LRT) that was used to evaluate the overall significance of a score variable in quintile categories compared with a χ² distribution with df = 4.

‡ $P_{\text{trend}}$, P values obtained by modelling score variables with quintile-specific medians as continuous variables.

§ $P_{\text{heterogeneity}}$, P values for BC risks across ER/PR status with df = 1 obtained using a data augmentation method.

Introducing systematic between-country differences in nutrient assessment. However, in the EPIC study, harmonized composition tables across European countries were used to translate food into nutrient intakes(42), thus sizeably improving the comparability of nutrient intakes.

One key element in pattern literature is reproducibility of patterns across populations. With twenty-three centres from ten countries, EPIC accounts for a wide heterogeneity in diet(90,91). Previous findings in Moskal et al.’s study(45) on the EPIC data showed that more than 75% of the variance that would be captured by centre-specific PC was captured by PC from overall PCA. This evidence suggested that overall PCA combining data from all EPIC centres allows capturing a good proportion of the variance explained by each EPIC centre. This motivated the choice of applying pattern decomposition on the overall data.

Conclusion

The current study presented results of a nutrient pattern analysis in an international setting using a new tool, TT, and subsequently related the patterns to risk of developing BC. TT is a complementary method to PCA in nutritional epidemiology as it produces readily interpretable sparse components. In the EPIC study, nutrient patterns characterized by a diet rich in macronutrients of animal origin, such as cholesterol or SFA, were associated with a non-significant increase in BC risk while a diet rich in vitamins, minerals and β-carotene, indicating a more plant-based diet, was associated with a significant decreased BC risk. This decrease was also significant for ER+, PR+, PR− and ER+/PR− tumours.

Acknowledgements

Acknowledgement: The authors thank Dr Anders Gorst-Rasmussen (Department of Cardiology, Aalborg University Hospital) for his critical input and useful discussions about the manuscript. Financial support: The coordination of the EPIC study is financially supported by the European Commission (Directorate General for Health and Consumer Affairs) and the International Agency for Research on Cancer (IARC). The national cohorts are supported by: the Health Research Fund (FIS) of the Spanish Ministry of Health RTICC ‘Red Temática de Investigación Coordinada’ (grant numbers RD06/0020/0001 and RD12/0036/0018), the Regional Governments of Andalucía, Asturias, Basque Country, Murcia (project 6236) and Navarra, and the Instituto de Salud Carlos III, Redes de Investigación Cooperativa (RD06/0020) (Spain); the Danish Cancer Society (Denmark); the Ligue Contre le Cancer, the Institut Gustave Roussy, Mutuelle Générale de l’Education Nationale and the Institut National...
Supplementary material

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S1368980015000294

References

Nutrient patterns and breast cancer in EPIC


