Reproducibility of data-driven dietary patterns in two groups of adult Spanish women from different studies. Adela Castelló, Virginia Lope, Jesús Vioque, Carmen Santamariña, Carmen Pedraz-Pingarrón , Soledad Abad, Maria Ederra , Dolores Salas-Trejo , Carmen Vidal , Carmen Sánchez-Contador , Nuria Aragonés , Beatriz Pérez-Gómez and Marina Pollán.

Br J Nutr. 2016 Aug;116(4):734-42.

Title: Reproducibility of data-driven dietary patterns in two groups of adult Spanish women from different studies.

Authors: Adela Castelló[1,2,3], Virginia Lope[1,2,3], Jesús Vioque[2, 4], Carmen Santamariña[5], Carmen Pedraz-Pingarrón[6], Soledad Abad[7], Maria Ederra[8], Dolores Salas-Trejo[9], Carmen Vidal[10], Carmen Sánchez-Contador[11], Nuria Aragonés[1,2,3], Beatriz Pérez-Gómez[1,2,3], Marina Pollán[1,2,3] **on behalf of DDM-Spain research group^**

1. Cancer Epidemiology Unit. National Center for Epidemiology. Instituto de Salud Carlos III. Madrid, Spain.

2. Consortium for Biomedical Research in Epidemiology & Public Health (CIBERESP). Instituto de Salud Carlos III. Madrid, Spain.

3. Cancer Epidemiology Research Group. Oncology and Hematology Area. IIS Puerta de Hierro (IDIPHIM). Madrid, Spain.

4. Universidad Miguel Hernandez. Sant Joan D'Alacant, Spain.

5. Galician Breast Cancer Screening Program, Galician Regional Health Authority. A Coruña, Spain.

6. Castile-León Breast Cancer Screening Program, General Directorate of Public Health. Burgos, Spain

7. Aragón Breast Cancer Screening Program. Aragon Health Service. Zaragoza, Spain.

8. Navarre Breast Cancer Screening Program, Public Health Institute. Pamplona, Spain.

9. Valencian Breast Cancer Screening Program, General Directorate of Public Health. Valencia, Spain.

10. Cancer Prevention and Control Unit, Catalonian Institute of Oncology (ICO). Barcelona, Spain.

11. Balearic Islands Breast Cancer Screening Program, Regional Authority for Health & Consumer Affairs. Palma de Mallorca, Islas Baleares.

^ Membership of the DDM-Spain research group is provided in the acknowledgments section

CORRESPONDING AUTHOR:

Dr. Adela Castelló
Cancer Epidemiology Unit, National Center for Epidemiology.
Instituto de Salud Carlos III.
Avenida Monforte de Lemos, 5, 28029, Madrid, Spain.
Phone: +34 91 822 2667.
Fax: +34 91 387 7815.
e-mail: acastello@isciii.es

Running Title: Reproducibility of data-driven patterns

Keywords: Dietary patterns; Reproducibility; Congruence Coefficient; Principal Component Analysis; Component loadings; Component scores.

ABSTRACT

The objective of the present study was to assess the reproducibility of data-driven dietary patterns in different samples extracted from similar populations**.** Dietary patterns were extracted by applying principal component analyses to the dietary information collected from a sample of 3550 women recruited in 7 screening centres belonging to the Spanish Breast Cancer (BC) screening network (DDM-Spain study). The resulting patterns were compared with the 3 dietary patterns obtained in a previous Spanish case-control study on female BC (EpiGEICAM study) using the dietary intake data of 973 healthy participants. The level of agreement between patterns was determined using the congruence coefficient (CC) between the pattern loadings, considering patterns with a CC≥0.85 as fairly similar. The conclusions were compared with those reached considering as fairly similar those patterns with a statistically significant linear correlation between patterns scores (the method commonly used). This is the first study exploring the reproducibility of data-driven patterns from two studies and the first using the CC to determine pattern similarity. We were able to reproduce the EpiGEICAM Western pattern in the DDM-Spain sample (CC=0.90). However, the reproducibility of the Prudent (CC=0.76) and Mediterranean (CC=0.77) patterns was not as good. The linear correlation between pattern scores was statistically significant in all cases, highlighting its arbitrariness for determining pattern similarity. We conclude that the reproducibility of widely prevalent dietary patterns is better than the reproducibility of more population specific patterns. More methodological studies are needed to establish an objective measurement and threshold to determine patterns' similarity.

INTRODUCTION

Diet is a key modifiable risk factor, but the exploration of its role in disease occurrence is complicated due to methodological issues related to the dietary assessment method used [1-3], food and nutrient interactions [4,5] and differences in food consumption across populations [6-8]. Traditionally, nutritionists and researchers have explored the effect of individual dietary factors in disease occurrence. However, some authors advocate the use of dietary patterns instead of individual foods and nutrients, arguing that they may better capture variability in the population's diet, while allowing the evaluation of interactions between dietary factors [9-11].

These patterns can be identified with data-driven methods such as Principal Component Analysis (PCA), Factor Analysis (FA) and Cluster Analysis or can be represented by investigator-driven patterns, known as dietary quality indices. Investigator-driven patterns assign a set of scores based on individuals' fulfillment of a set of fixed recommendations. Therefore, they are widely applicable facilitating the exploration of the reproducibility of their association with different diseases in independent populations [12-16]. However, they present the disadvantage of being very disease-dependent given that they are mainly based on existing evidence of the association between diet and cardiovascular disease [17]. On the other hand, data-driven dietary patterns are more representative of the diet of the specific population from which they have been extracted and independent from the diseases, but a number of authors argue that the patterns obtained are very population-dependent and therefore difficult to reproduce in other settings [11,18,19]. The reproducibility of data-driven dietary patterns has been assessed previously by various authors using dietary information obtained with common assessment tools at different moments of time within the same sample [20-23]. However, no prior studies have explored the reproducibility of data-driven dietary patterns extracted from different samples.

The objective of this study was to assess the reproducibility of data-driven dietary patterns in different samples extracted from similar populations. We compared the results from a previous case-control study (EpiGEICAM) on diet and female breast cancer (BC) in Spain [24] with those obtained from a sample of Spanish women attending BC screening programs (DDM-Spain), by evaluating the correlation between pattern scores and the congruence between the composition of patterns in both populations.

METHODS

Study population and data collection

We used information on 3 dietary patterns obtained from a previous case control study on female BC (EpiGEICAM-study) using the dietary intake data of 973 healthy participants aged 22-71 and recruited in 14 Spanish provinces during the period 2006-2011 [24]. These patterns will be used as a reference to explore their reproducibility in a different sample using data from the DDM-Spain participants. DDM-Spain (*Determinantes de la Densidad Mamográfica en España*- Determinants of Mammographic Density in Spain) is a cross-sectional multicentre study carried out in 7 screening centres belonging to the Spanish Breast Cancer Screening network and located throughout the Spanish peninsula [25,26]. In Spain, all women aged 50-69 (45-69 in some regions), regardless of nationality or legal status, are invited to be screened under these government-sponsored programs every 2 years. Women were randomly selected among all screening attendants and invited to participate on a daily basis until the minimum sample size of 500 for each center was reached. A total of 3,550 women were recruited between 2007 and 2008, with an average participation rate of 74.5% (range 64.7–84.0% across centres). Women were interviewed at the screening centres by trained interviewers who collected demographic, anthropometric, physical activity, gynaecologic, obstetric and occupational data, as well as family and personal history (including weight and height at age 18). Information on smoking included current status and months since quitting for ex-smokers. Current smokers were defined as those women who smoked at the time of mammography or had quit less than 6 months before. Dietary intake during the preceding year was collected using a validated 117-item food frequency questionnaire (FFQ) [27,28]. Post-menopausal status was defined as self-reported absence of menstruation in the previous 12 months. Interviewers measured weight, height, waist and hip circumferences twice using the same protocol and identical balance scales, stadiometers and measuring tapes. A third measure was taken when the first two were not equal.

The DDM-Spain study was conducted according to the guidelines laid down in the Declaration of Helsinki and all procedures involving human subjects were approved by the bioethics and animal welfare committee at the Carlos III Institute of Health. All participants signed a consent form, including permission to publish the results from the current research.

Dietary patterns

The FFQs used in both studies were designed to assess the whole diet, had similar structures and were based on a validated FFQ [27,28]. However, the FFQ of the DDM-Spain study included some additional food items that were not contained in the FFQ of the GEICAM-study [25,26]: the FFQ used in the EpiGEICAM study contained 99 items from which 86 were used to create the food groups (after excluding the non-caloric and alcoholic beverages) whereas the FFQ from DDM-Spain included 117 items (the same 99 from DDM-Spain plus 18 additional foods) from which 99 were used to create the food groups (after excluding non-caloric and alcoholic beverages). In both cases, the dietary information collected was grouped into the exact same 26 food groups that are summarized in **Table 1**, where the items only included in the DDM-Spain study are represented in bold.

The EpiGEICAM study identified 3 dietary patterns over 26 food groups: a Western pattern characterized by elevated intake of high-fat dairy products, processed meat, refined grains, sweets, caloric drinks and other convenience foods and sauces and by low intakes of low-fat dairy products and whole grains; a Prudent pattern defined by high intakes of low-fat dairy products, vegetables, fruits, whole grains and juices; and a Mediterranean pattern represented by a high intake of fish, vegetables, legumes, boiled potatoes, fruits, olives and vegetable oil, and a low intake of juices. These patterns explained 16%, 13% and 8% of the total variability in food intake, respectively [(24)]. We assessed the reproducibility of these three patterns by comparing them with the patterns extracted by applying the same PCA analysis to the same 26 food groups from the DDM-Spain sample.

Statistical analysis

Major existing dietary patterns were identified in the DDM-Spain sample using the same technique applied to the EpiGEICAM data [(24)]: applying PCA without rotation to the variance-covariance matrix over 26 inter-correlated food groups that were reduced to a set of principal components (dietary patterns in this case). The first components with eigenvalues higher than 1 were selected for initial exploration. The PCA reports, for a given pattern, a set of weights associated with each food group (commonly called component/pattern weights) that is used to calculate pattern scores, defined, for each individual, as a weighted sum of the food group consumption. Afterwards, these scores were correlated with the food group consumption to calculate the pattern loadings, which indicate the importance of individual food groups in each pattern. Pattern weights and pattern loadings give similar information, except that they are measured on different scales (weights are standardized into Z score form)[(29)]. Since only information on pattern loadings was provided by the EpiGEICAM study, these were used to compare dietary patterns from both studies. For comparison purposes we considered that food groups with pattern loadings ≥|0.3| were the main contributors to a dietary pattern.

To evaluate the level of agreement between the food composition of patterns extracted in the DDM-Spain study and those reported in the EpiGEICAM study, we calculated the Congruence Coefficients (CC) [29,30] between the pattern loadings from both studies. CC represents the correlation between pattern loadings based on their deviations from 0 (instead of being based on the deviations from the mean of the factor loadings as the Pearson Correlation is) and it is the preferred measure for component/factor similarity extracted with PCA/FA[31]. CC ranges from -1 to 1, a value in the range [0.85-0.94] corresponds to a fair similarity, while a value higher or equal to 0.95 implies that the two compared components/factors can be considered equivalent [31-33].

The CCs between the pattern loadings of a given pattern from EpiGEICAM (l_{1j}) and the pattern loadings of a given pattern from DDM-Spain (l_{2j}) for each of the i=1,…26 food groups, were calculated as follows:

$$CC = \frac{\sum_{j=1}^{26} l_{1j} \cdot l_{2j}}{\sqrt[2]{\left(\sum_{j=1}^{26} l_{1j}^2\right) \cdot \left(\sum_{j=1}^{26} l_{2j}^2\right)}}$$

Additionally, to follow the same methodology commonly used in studies exploring the reproducibility of dietary patterns, the Spearman Correlation Coefficients (Corr) between the EpiGEICAM and DDM-Spain pattern scores were calculated. For that purpose, patterns scores (which reflect the level of compliance of each woman with each one of the dietary patterns) were calculated as the linear combination of the consumption of food groups weighted by the pattern loadings from EpiGEICAM Western, Prudent and Mediterranean patterns and from the set of selected patterns resulting from applying PCA to the DDM-Spain data as follows [34]:

$$P_{ki} = \sum_j (L_{kj} \cdot C_{ji})$$

P= Pattern Score; L=Loading Score; C=Centered food consumption

k= Western, Prudent and Mediterranean patterns from EpiGEICAM; Western, Prudent and Mediterranean patterns from DDM-Spain.

i=1,…., 3550 women

j=1,…, 26 food groups

CC is the preferred measure for component/factor similarity extracted with PCA/FA because its validity is supported by methodological research [31-33]. Additionally a recent study has questioned using solely the Pearson correlation (Corr) coefficient to assess pattern similarity [35]. However, the majority of the studies exploring the reproducibility of dietary patterns base their conclusions on the latter measure, considering any significant correlation as being indicative of pattern similarity regardless of its value [20-23]. Here, we provide the correlation coefficient for the sake of comparability to published research, but we will base our final conclusion regarding pattern reproducibility on the CC.

In order to take into account sampling variability in the estimation of the pattern loadings using DDM-Spain data, and subsequently in the estimation of the agreement measurements between the patterns identified within the EpiGEICAM and DDM-Spain studies, we performed a non-parametric bootstrap estimation with 5000 replications. Using sampling replacement, the bootstrap obtained 5000 replicates of the original DDM-Spain dataset. PCA was then applied in each replication and the three principal components that proved to be more similar to those reported in EpiGEICAM were selected, based on the distance between the pattern loadings (more detail is given in **Supplemental Meth. 1**). The 95% percentile confidence intervals for each parameter were represented by percentiles 2.5 and 97.5 of the 5000 bootstrap point estimates' distribution.

Similar analyses were carried out applying the PCA to food groups from DDM-Spain study, that include the same exact 86 items considered in the EpiGEICAM analysis (**Supplemental Tab. 1 and Figure 1)**

Analyses were performed using STATA/MP 14.0.

RESULTS

The anthropometric, reproductive and sociodemographic characteristics of the EpiGEICAM controls (extracted from the Castelló et al. article [24]) and DDM-Spain women are summarized in **Table 2**. The DDM-Spain study recruited a higher percentage of older and postmenopausal women (53% vs 23%), women with higher energy intake (on average, 150 kcal/day more in the DDM-Spain group),higher BMI and a higher percentage of women that practice physical activity with moderate-vigorous intensity (76% vs 63%). On the other hand, these women reported lower intake of alcohol, lower educational level (34% with Primary school or less in DDM and 16% in

EpiGEICAM), lower percentage of family history of BC (7% vs 20%), lower age at first delivery (43% of parous women in the DDM had their first child before 25 years old, while this proportion was 26% in EpiGEICAM), and there was a lower percentage of nulliparous (9% vs 23%). The distribution of age at menarche and smoking appeared to be fairly similar in both studies.

Figures 1, 2 and **3** show the comparison between the original loadings from the EpiGEICAM study with their corresponding values in the DDM-Spain study. Western patterns from both studies were characterized by a high consumption of high-fat dairy products, refined grains, caloric drinks and convenience food and sauces; and a low consumption of low-fat dairy and whole grains. Correlations with the intake of red and/or processed meat and with sweets were also close to the 0.3 threshold. Moreover, the DDM-Spain Western pattern seemed to be negatively correlated with the consumption of white fish, a result that was not observed in EpiGEICAM. Despite these small differences, the elevated CC between patterns (CC=0.90) indicates a fair similarity between the Western patterns extracted from the EpiGEICAM and DDM-Spain data (**Figure 1**).

We did not identify a pattern among women of the DDM-Spain study that was highly congruent with the EpiGEICAM Prudent pattern. The most similar pattern presented a high consumption of whole grains and juices but failed to correlate with low-fat dairy products, vegetables and fruits (**Figure 2**). Something similar happened with the Mediterranean pattern; several high correlations were observed with some vegetables, legumes, potatoes and nuts. However ,the pattern from DDM-Spain study did not include other typical factors of the Mediterranean diet such as fish, olive oil and fruits (even if pattern loadings for these food groups were not low), while other foods more common in the Western diet, such as low-fat dairy products, sweets, sugary and convenience foods, were included with high correlations. According to the CC (0.77), the EpiGEICAM and DDM-Spain Mediterranean patterns cannot be considered similar (**Figure 3**).

Finally, had we considered any significant correlation as being indicative of similarity, we would have concluded that all patterns extracted with the EpiGEICAM data were reproducible in the DDM-Spain study.

DISCUSSION

To the best of our knowledge, this is the first study exploring the reproducibility of data-driven patterns in two different samples extracted from similar populations. We were able to reproduce the Western pattern identified in women from the EpiGEICAM study among women attending BC screening programs who participated in the DDM-Spain study. However, the reproducibility of the Prudent and Mediterranean patterns cannot be considered good.

The association between dietary patterns and BC has been explored in many studies in different settings. Most of these studies identified a Western/Unhealthy pattern, which shares the most important characteristics with the Western patterns identified in EpiGEICAM and DDM-Spain, such as high consumption of fatty dairy products, red/processed meat, refined grains, sweets and convenience foods [36-41]. However, the Mediterranean and Prudent patterns have often been mixed under the names of Vegetable, Prudent, Healthy or Mediterranean diet. These patterns are characterized by a high consumption of vegetables and fruits [36-47] that are an important part of the Mediterranean diet, but fail to include other items such as olive oil [36,38-41,44-47], nuts [36-41,43-47], legumes [37,39-41,44,46,47] or fish [38,41] which are key foods to differentiate the so-called Prudent or Healthy patterns from the Mediterranean.

None of the above-mentioned studies have been able to identify both, a Prudent and a Mediterranean pattern in the same population, probably reflecting the difficulty in differentiating them in contexts where the Mediterranean diet is not very prevalent. On the other hand, the higher agreement in the definition of a Western pattern across studies is consistent with the greater reproducibility of this pattern observed in our study.

As noted earlier in this paper, PCA reduces a set of inter-correlated variables to a group of principal components (dietary patterns in this case) so that the maximum correlation between the variables within components and the minimum correlation among components is obtained [48]. Therefore, the greater the variability in diet, the easier it will be to find clearly differentiated independent patterns. In our study, while EpiGEICAM included women from 14 Spanish provinces (4 of them on the Mediterranean coast), DDM-Spain participants were recruited in screening centres located on 7 provinces (3 of them located on the Mediterranean coast). Therefore, the greater geographical distribution in the EpiGEICAM study may imply a greater representativeness of all diets across the Spanish territory. Additionally, distribution of age among DDM-Spain women was more homogeneous (range=45-69) than that observed in the EpiGEICAM participants (range=22-71). As García-Arenzana et al. previously described, older women tend to have healthier dietary habits than younger women [49], which may have produced a more heterogeneous distribution of dietary habits in the EpiGEICAM study. This heterogeneity might have facilitated the identification of more specific patterns, not only limited to the discrimination of two antagonistic patterns (Western vs Healthy/Prudent/Mediterranean) but also allowing the clear differentiation of patterns with subtle differences such as the Prudent and Mediterranean patterns.

Regarding the pre-established thresholds for the CC that define the similarity of dietary patterns in both studies, we based our decision on three published pieces of research that evaluated concordance coefficients in light of the subjective opinion of several experienced researchers judging the equivalence between different components [31-33]. Haven and Nesselroade [31,33] argue that values over 0.80 are enough to assume fair similarity between components while Lorenzo-Seva et al. [32] maintain a more conservative approach setting the cut point for fair similarity at 0.85 and preventing a CC below this value from being interpreted as indicative of similarity. All three articles agree on the difficulty in setting up a cut point under which patterns should be considered clearly different. Despite the fact that the CC is considered a good measure of agreement between components or factors extracted with PCA or FA [31-33], the existing bibliography evaluating the reproducibility of data-driven dietary patterns does not use this measure and bases its conclusions only on the correlations between pattern scores, considering any significant correlation as being indicative of similarity regardless of its value [20-23] which can be as low as 0.27 [23]. In our case the correlations were significant and high for all three patterns (**Figures 1, 2** and **3)**. However, according to the CC, only the Western pattern can be considered fairly similar between studies, which highlights the arbitrariness of the significance of the linear correlation to define pattern similarity and the need to choose an appropriate measure and a concrete threshold for such measure to determine the level of congruence between patterns. In this regard, we have recently explored the applicability of previously reported dietary patterns in a different setting and we found that, for CC between pattern loadings ≥0.82 or correlations between pattern scores ≥0.57, patterns not only appear to have a very similar composition, but were similarly associated with BC risk [35]. The same direction of the associations but loss of significance was observed for values of the CC between pattern loadings ≤0.77 and values of the correlation between pattern scores ≤0.52. In the present study, taking into account only the methodological studies published regarding the threshold of the CC for pattern similarity [31-33], we followed the most conservative approach and considered dietary patterns to be fairly similar if CC values were ≥0.85.

One of the main limitations of the use of dietary patterns is the potential for subjective interpretations by the investigator to be introduced at various stages of the dietary patterns' construction. Subjective decisions that might affect the comparability between studies are: which foods should be included in each of the defined groups, the thresholds chosen to determine the contribution of food groups to the identified dietary patterns, and the assignation of a label to each of these patterns [9-11,18,19]. However, we have demonstrated that this limitation can be overcome by a detailed analysis when comprehensive information on food grouping and loadings is provided by authors[35]. On the other hand, both FFQs from EpiGEICAM and DDM-Spain collected information on 99 identical foods, except for the fact that DDM-Spain included 18 additional foods that were not included in EpiGEICAM. Additionally, the same group of researchers took principal responsibility for the analysis of the data; therefore, food grouping and labelling was very similar in both studies.

Finally, we would like to summarize the main strengths of the present study. As previously mentioned, various studies have assessed the reproducibility of investigator-driven patterns [12-16]. The reproducibility of data-driven dietary patterns extracted from the same sample using the dietary information obtained with different assessment tools or in different moments of time [20-23] has also been explored. However, to our knowledge this is the first study assessing the reproducibility of data-driven dietary patterns in different samples from similar populations and the first using the CC to evaluate their similarity. In addition, most of the published studies on reproducibility of data-driven dietary patterns based their conclusions on limited sample sizes that range from 124-498 [20-22]. Dietary patterns from EpiGEICAM were extracted over 973 healthy women and for DDM-Spain the sample size was 3550, a size only exceeded by the Newby et al. study [23].

CONCLUSIONS

The reproducibility of widely prevalent dietary patterns, such as the Western pattern, is better than the reproducibility of patterns more specific to certain populations, such as the Mediterranean. More methodological studies exploring the reproducibility of dietary patterns are needed to establish a more objective threshold for the CC between pattern loadings and their equivalent Corr between pattern scores which define pattern similarity.

ACKNOWLEDGEMENTS

The authors wish to thank the DDM-Spain study participants for their contribution to breast cancer research and to thank all collaborator researchers: Pilar Moreo, Mª Pilar Moreno, Mª Soledad Abad, Francisca Collado, Francisco Casanova, Jose Antonio Vázquez, Nieves Ascunce, Milagros García, Manuela Alcaraz, Mª Soledad Laso, Josefa Miranda and Francisco Ruiz Perales.

FINANCIAL SUPPORT

This study was supported by Carlos III Institute of Health FIS (Spanish Public Health Research Fund: PI060386 FIS; PS09/00790 and PI15CIII/0029 research grants); the Spanish Ministry of Health (EC11-273); the Spanish Ministry of Economy and Competitiveness (IJCI-2014-20900); the Spanish Federation of Breast Cancer Patients (FECMA: EPY 1169-10); and the Association of Women with Breast Cancer from Elche (AMACMEC: EPY 1394/15).

None of the funders had any role in the design, analysis or writing of this article.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest

AUTHORSHIP

Virginia Lope, Nuria Aragonés, Beatriz Pérez-Gómez and Marina Pollán designed research; Adela Castelló, Jesús Vioque, Carmen Santamariña, Carmen Pedraz-Pingarrón, Soledad Abad, Maria Ederra, Dolores Salas-Trejo, Carmen Vidal, Carmen Sánchez-Contador collected the data and/or prepared the database. Adela Castelló performed statistical analysis and wrote the initial version of the manuscript that Marina Pollán revised and corrected in its different versions. All authors have read and approved the final manuscript

REFERENCES

1. Bingham SA, Luben R, Welch A *et al.* (2003) Are imprecise methods obscuring a relation between fat and breast cancer? *Lancet* **362**, 212-214.

2. Kelemen LE (2007) GI Epidemiology: nutritional epidemiology. *Aliment Pharmacol Ther* **25**, 401-407.

3. Willett W (2001) Commentary: Dietary diaries versus food frequency questionnaires-a case of undigestible data. *Int J Epidemiol* **30**, 317-319.

4. Jacobs DR, Jr., Steffen LM (2003) Nutrients, foods, and dietary patterns as exposures in research: a framework for food synergy. *Am J Clin Nutr* **78**, 508s-513s.

5. Messina M, Lampe JW, Birt DF *et al.* (2001) Reductionism and the narrowing nutrition perspective: time for reevaluation and emphasis on food synergy. *J Am Diet Assoc* **101**, 1416-1419.

6. Irala-Estevez JD, Groth M, Johansson L *et al.* (2000) A systematic review of socio-economic differences in food habits in Europe: consumption of fruit and vegetables. *Eur J Clin Nutr* **54**, 706-714.

7. Sanchez-Villegas A, Martinez JA, Prattala R *et al.* (2003) A systematic review of socioeconomic differences in food habits in Europe: consumption of cheese and milk. *Eur J Clin Nutr* **57**, 917-929.

8. Teufel NI (1997) Development of culturally competent food-frequency questionnaires. *Am J Clin Nutr* **65**, 1173s-1178s.

9. Barkoukis H (2007) Importance of understanding food consumption patterns. *J Am Diet Assoc* **107**, 234-236.

10. Hu FB (2002) Dietary pattern analysis: a new direction in nutritional epidemiology. *Curr Opin Lipidol* **13**, 3-9.

11. Jacques PF, Tucker KL (2001) Are dietary patterns useful for understanding the role of diet in chronic disease? *Am J Clin Nutr* **73**, 1-2.

12. George SM, Ballard-Barbash R, Manson JE *et al.* (2014) Comparing indices of diet quality with chronic disease mortality risk in postmenopausal women in the Women's Health Initiative Observational Study: evidence to inform national dietary guidance. *Am J Epidemiol* **180**, 616-625.

13. Harmon BE, Boushey CJ, Shvetsov YB *et al.* (2015) Associations of key diet-quality indexes with mortality in the Multiethnic Cohort: the Dietary Patterns Methods Project. *Am J Clin Nutr* **101**, 587-597.

14. Liese AD, Krebs-Smith SM, Subar AF *et al.* (2015) The Dietary Patterns Methods Project: synthesis of findings across cohorts and relevance to dietary guidance. *J Nutr* **145**, 393-402.

15. McCullough ML (2014) Diet patterns and mortality: common threads and consistent results. *J Nutr* **144**, 795-796.

16. Reedy J, Krebs-Smith SM, Miller PE *et al.* (2014) Higher diet quality is associated with decreased risk of all-cause, cardiovascular disease, and cancer mortality among older adults. *J Nutr* **144**, 881-889.

17. Fung TT, McCullough ML, Newby PK *et al.* (2005) Diet-quality scores and plasma concentrations of markers of inflammation and endothelial dysfunction. *Am J Clin Nutr* **82**, 163-173.

18. Martinez ME, Marshall JR, Sechrest L (1998) Invited commentary: Factor analysis and the search for objectivity. *Am J Epidemiol* **148**, 17-19.

19. Slattery ML, Boucher KM (1998) The senior authors' response: Factor analysis as a tool for evaluating eating patterns. *Am J Epidemiol* **148**, 20-21.

20. Hu FB, Rimm E, Smith-Warner SA *et al.* (1999) Reproducibility and validity of dietary patterns assessed with a food-frequency questionnaire. *Am J Clin Nutr* **69**, 243-249.

21. Khani BR, Ye W, Terry P *et al.* (2004) Reproducibility and validity of major dietary patterns among Swedish women assessed with a food-frequency questionnaire. *J Nutr* **134**, 1541-1545.

22. Nanri A, Shimazu T, Ishihara J *et al.* (2012) Reproducibility and validity of dietary patterns assessed by a food frequency questionnaire used in the 5-year follow-up survey of the Japan Public Health Center-Based Prospective Study. *J Epidemiol* **22**, 205-215.

23. Newby PK, Weismayer C, Akesson A *et al.* (2006) Long-term stability of food patterns identified by use of factor analysis among Swedish women. *J Nutr* **136**, 626-633.

24. Castello A, Pollan M, Buijsse B *et al.* (2014) Spanish Mediterranean diet and other dietary patterns and breast cancer risk: case-control EpiGEICAM study. *Br J Cancer* **111**, 1454-1462.

25. Lope V, Perez-Gomez B, Sanchez-Contador C *et al.* (2012) Obstetric history and mammographic density: a population-based cross-sectional study in Spain (DDM-Spain). *Breast Cancer Res Treat* **132**, 1137-1146.
26. Pollan M, Lope V, Miranda-Garcia J *et al.* (2012) Adult weight gain, fat distribution and mammographic density in Spanish pre- and post-menopausal women (DDM-Spain). *Breast Cancer Res Treat* **134**, 823-838.
27. Vioque J, Navarrete-Munoz EM, Gimenez-Monzo D *et al.* (2013) Reproducibility and validity of a food frequency questionnaire among pregnant women in a Mediterranean area. *Nutr J* **12**, 26.
28. Willett WC, Sampson L, Stampfer MJ *et al.* (1985) Reproducibility and validity of a semiquantitative food frequency questionnaire. *Am J Epidemiol* **122**, 51-65.
29. Burt C (1948) Factor Analysis and canonical correlations. *Br J Math Stat Psychol* **1**, 95-106.
30. Tucker LR (1951) A method for the synthesis of factor analysis studies (Personnel Research Section Report No. 984). Washington: Department of the Army., 18-20.
31. Haven S, Berge J (1977) Tucker's coefficient congruence as a measure of factorial invariance: An empirical study. *Heymans Bulletin 290 EX University of Groningen*.
32. Lorenzo-Seva U, Berge J (2006) Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology* **2**, 54-67.
33. Nesselroade J, Baltes P (1970) On a dilemma of comparative factor analysis: A study of factor matching based on random data. *Educ Psychol Meas* **30**, 935-948.
34. Schulze MB, Hoffmann K, Kroke A *et al.* (2003) An approach to construct simplified measures of dietary patterns from exploratory factor analysis. *Br J Nutr* **89**, 409-419.
35. Castello A, Buijsse B, Martin M *et al.* (2016) Evaluating the applicability of data-driven dietary patterns to independent samples with focus on measurement tools for pattern similarity. *J Acad Nutr Diet (in press).*
36. Agurs-Collins T, Rosenberg L, Makambi K *et al.* (2009) Dietary patterns and breast cancer risk in women participating in the Black Women's Health Study. *Am J Clin Nutr* **90**, 621-628.
37. Cottet V, Touvier M, Fournier A *et al.* (2009) Postmenopausal breast cancer risk and dietary patterns in the E3N-EPIC prospective cohort study. *Am J Epidemiol* **170**, 1257-1267.
38. Cui X, Dai Q, Tseng M *et al.* (2007) Dietary patterns and breast cancer risk in the Shanghai breast cancer study. *Cancer Epidemiol Biomarkers Prev* **16**, 1443-1448.
39. Terry P, Suzuki R, Hu FB *et al.* (2001) A prospective study of major dietary patterns and the risk of breast cancer. *Cancer Epidemiol Biomarkers Prev* **10**, 1281-1285.
40. Velie EM, Schairer C, Flood A *et al.* (2005) Empirically derived dietary patterns and risk of postmenopausal breast cancer in a large prospective cohort study. *Am J Clin Nutr* **82**, 1308-1319.
41. Wu AH, Yu MC, Tseng CC *et al.* (2009) Dietary patterns and breast cancer risk in Asian American women. *Am J Clin Nutr* **89**, 1145-1154.
42. Adebamowo CA, Hu FB, Cho E *et al.* (2005) Dietary patterns and the risk of breast cancer. *Ann Epidemiol* **15**, 789-795.
43. Bessaoud F, Tretarre B, Daures JP *et al.* (2012) Identification of dietary patterns using two statistical approaches and their association with breast cancer risk: a case-control study in Southern France. *Ann Epidemiol* **22**, 499-510.
44. De Stefani E, Deneo-Pellegrini H, Boffetta P *et al.* (2009) Dietary patterns and risk of cancer: a factor analysis in Uruguay. *Int J Cancer* **124**, 1391-1397.
45. Demetriou CA, Hadjisavvas A, Loizidou MA *et al.* (2012) The Mediterranean dietary pattern and breast cancer risk in Greek-Cypriot women: a case-control study. *BMC Cancer* **12**, 113.
46. Hirose K, Matsuo K, Iwata H *et al.* (2007) Dietary patterns and the risk of breast cancer in Japanese women. *Cancer Sci* **98**, 1431-1438.
47. Zhang CX, Ho SC, Fu JH *et al.* (2011) Dietary patterns and breast cancer risk among Chinese women. *Cancer Causes Control* **22**, 115-124.
48. Rencher A (2002) Principal Component Analysis. In *Methods of Multivariate Analysis* pp. 380-407: John Wiley & Sons, Inc.
49. Garcia-Arenzana N, Navarrete-Munoz EM, Peris M *et al.* (2012) Diet quality and related factors among Spanish female participants in breast cancer screening programs. *Menopause* **19**, 1121-1129.

Table 1: Description of food groups used in principal component analyses

FOOD GROUP	FOOD[a]
HIGH-FAT DAIRY	Whole-fat milk, w_1[b]·**A+D enriched milk**[c], w_1[b]·**Folate enriched milk**[b], double cream, condensed milk, whole-fat yogurt, semi-cured, cured or creamy cheese, custard, flan, pudding, ice-cream
LOW FAT DAIRY	Semi-skimmed and skimmed milk, **Omega3 enriched milk**[c, d], w_2[b]·**A+D enriched milk**[c], w_2[b]·**Folate enriched milk**[c], **soya milk**[c]**, soya yogurt**[c], skimmed yogurt, cottage or fresh white cheese
EGGS	Eggs
WHITE MEAT	Chicken with skin, skinless chicken, game (turkey, rabbit, etc.)
RED MEAT	Pork, beef, lamb, liver (beef, pork or chicken), entrails, hamburger
PROCESSED MEAT	**Serrano ham**[c] and other cold meat, sausages, bacon, pâte, foie-gras
WHITE FISH	**1/3·all kind of fried fish**[c], Fresh white fish (hake, sea bass, sea bream)
OILY FISH	**1/3· all kind of fried fish**[c], Fresh blue fish (Tuna, swordfish, sardines, anchovies, salmon), canned tuna, canned sardines or mackerel, salted and smoked fish
SEAFOOD/SHELLFISH	**1/3·all kind of fried fish**[c], Clams, mussels, oysters, squid, cuttlefish, octopus, prawn, crab, shrimp, lobster
LEAFY VEGETABLES	Spinach, chard, lettuce, endive, escarole
FRUITING VEGETABLES	Tomato, eggplant, zucchini, cucumber, pepper, artichoke
ROOT VEGETABLES	Carrot, pumpkin
OTHER VEGETABLES	Cooked cabbage, cauliflower or broccoli, onion, green beans, asparagus, **mushrooms**[c], corn, garlic, **vegetable soup**[c]
LEGUMES	Legumes, **soya sprouts**[c]
POTATOES	Roasted or boiled potatoes
FRUITS	Orange, mandarin, banana, apple, pear, peach, nectarine, apricot, watermelon, melon, grapes, plums or prunes (dried or fresh), **strawberries**[c], kiwi
NUTS	Almonds, peanuts, pine nuts, hazelnut
REFINED GRAINS	White-flour bread, rice, pasta
WHOLE GRAINS	Whole-grain bread and partial whole-grain bread, breakfast cereals, **wheat germs**[c]
OLIVES AND VEGETABLE OIL	Olives, added olive oil to salads, bread and dishes, other vegetable oils (sunflower, corn, soybean)
OTHER EDIBLE FATS	Margarine, butter
SWEETS	Chocolate and other sweets, cocoa powder, plain cookies, chocolate cookies, pastries (croissant, donut, cake, pie or similar)
SUGARY	Jam, honey, sugar
JUICES	**Tomato juice**[c], freshly squeezed orange juice, juice (other than freshly squeezed)
CALORIC DRINKS	Sugar-sweetened soft drinks
CONVENIENCE FOOD AND SAUCES	Fried potatoes, crisps, pizza, chicken and Serrano ham croquette, mayonnaise, tomato sauce, ketchup, fish sticks

[a]Log-transformed intake in grams.

[b]Weighted within the high and low fat dairy categories according to the consumption of whole, semi-skimmed and skimmed milk.

w_1 = whole /(whole + semi- skimmed + skimmed)

w_2 = (semi-skimmed + skimmed) /(whole + semi skimmed + skimmed)

w_1 and w_2 were 0.5 if consumption was 0 grams for whole, semi-skimmed and skimmed milk.

[c]In bold, the additional items included only in the FFQ from the DDM-Spain study that were not collected in the FFQ from EpiGEICAM study.

[d]All the omega3 enriched milk brands that have been consulted are skimmed or semi-skimmed

Table 2: Anthropometric, reproductive and sociodemographic characteristics of EpiGEICAM controls and DDM-Spain women.

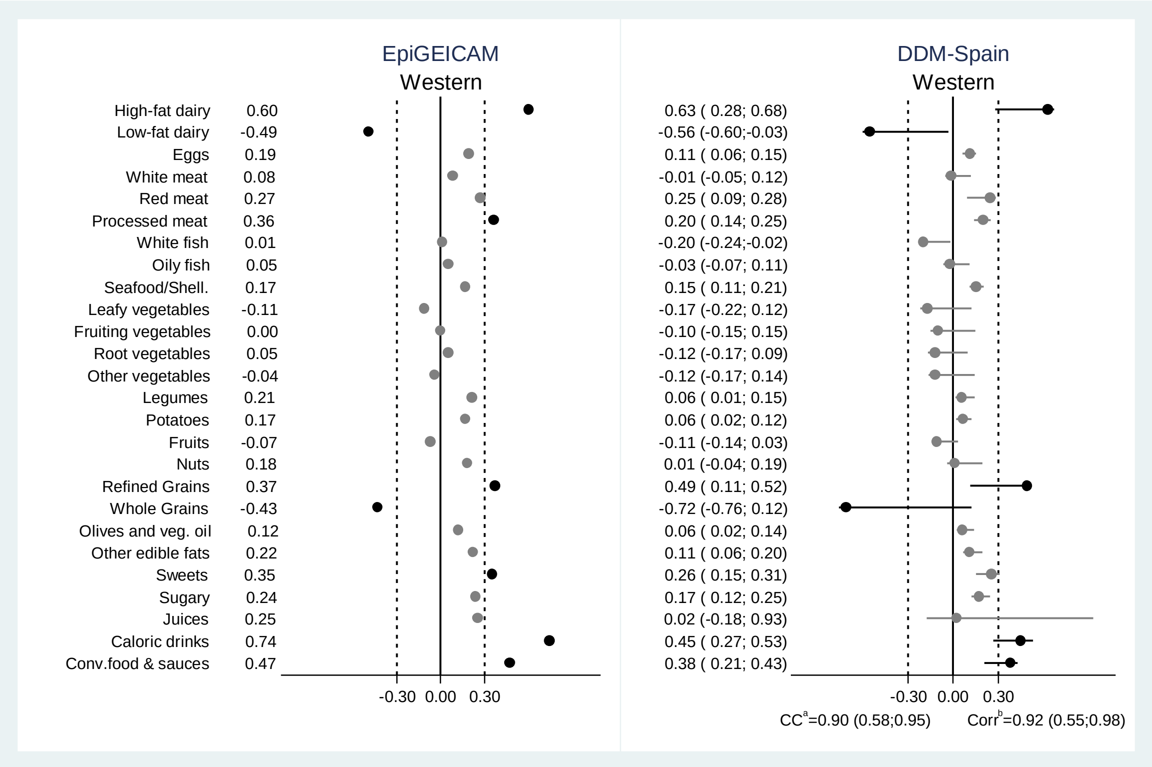
	EpiGEICAM Controls[a] n=973		**DDM-Spain n=3550**	
EpiGEICAM PATTERNS:	**Mean(sd)**	**% v.e.[b]**	**Mean(sd)**	**% v.e.[b]**
Western Pattern	0.00 (3.77)	16%	0.00 (2.31)	16%
Prudent Pattern	0.00 (3.34)	13%	0.34 (-2.21-1.92)[c]	15%
Mediterranean Pattern	0.00 (2.70)	8%	0.00 (1.50)	7%
PARTICIPANTS' CHARACTERISTICS:				
Energy intake (kcal/day), mean (sd)	1897 (628)		2054.15 (481.09)	
Alcohol intake (g/day), median (IQR)	2 (0.04;7.10)		0.85 (0-5.68)	
BMI (Kg/m^2), mean (sd)	25.36 (4.28)		28.03 (4.99)	
Age, mean (sd)	50.63 (9.47)		56.20(5.46)	
Age at menarche, mean (sd)	12.44 (1.52)		13 (12-14)	
Physical activity in the last year, n (%)				
Low	287 (30%)		842 (24%)	
Moderate	368 (38%)		1842 (52%)	
Vigorous	246 (25%)		866 (24%)	
Unknown	72 (7%)		--	
Smoking, n (%)				
Never or Former +6months	645 (67%)		2180 (61%)	
Smoker or former smoker <6 months	325 (33%)		1370 (39%)	
Unknown	3 (0%)		--	
Education, n (%)				
Primary school or less	158 (16%)		1204 (34%)	
Secondary school	489 (50%)		1978 (56%)	
University	318 (33%)		363 (10%)	
Unknown	8 (1%)		5 (0%)	
Family history of BC, n (%)				
No	782 (80%)		3291 (93%)	
Yes	191 (20%)		259 (7%)	
Age at first delivery, n (%)				
<20	45 (5%)		302 (9%)	
20-24	208 (21%)		1194 (34%)	
25-29	266 (27%)		1271 (36%)	
>29	148 (15%)		465 (13%)	
Nulliparous	220 (23%)		316 (9%)	
Unknown	86 (9%)		2 (0%)	
Menopausal status, n (%)				
Premenopausal	513 (53%)		816 (23%)	
Postmenopausal	460 (47%)		2734 (77%)	

[a] Descriptive data extracted from Castello et al. [24] scientific article.

[b] v.e.: Total variability of food groups intake explained by the pattern.

[c] Since distribution of the prudent score was skewed, the median and IQQ range was used to describe this Score.

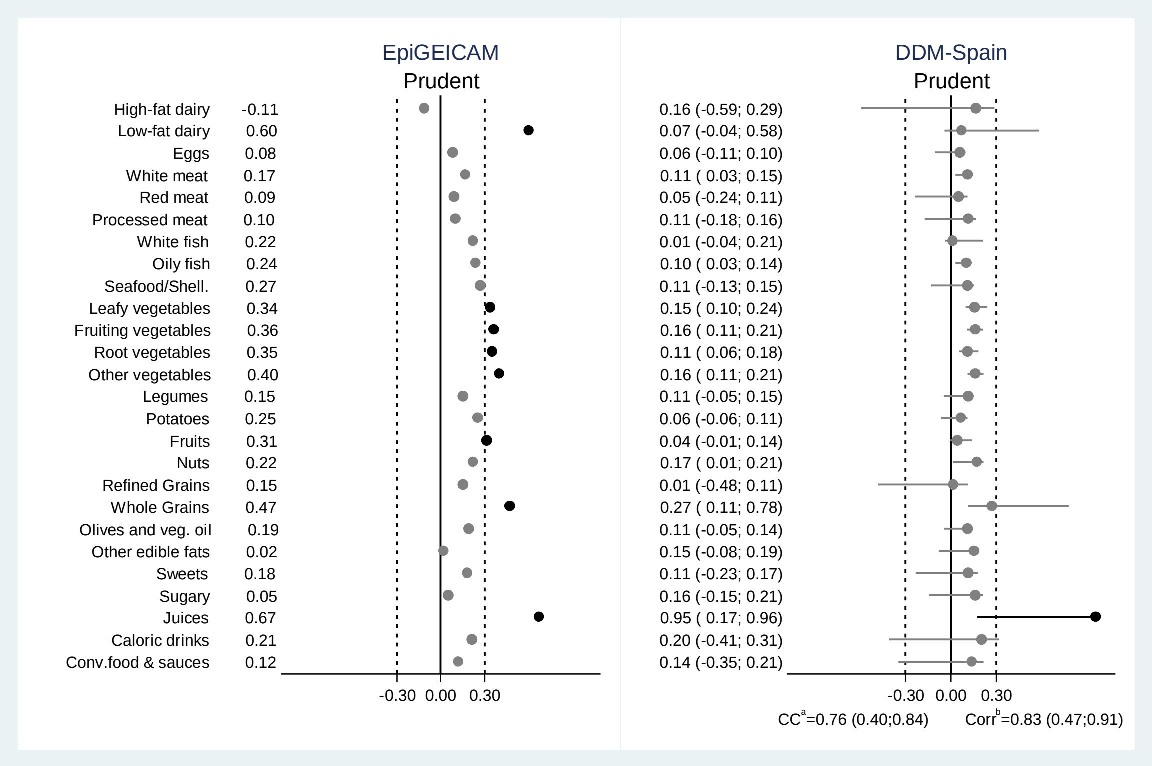
Fig.1 Pattern loadings of the Western dietary pattern extracted from EpiGEICAM study (24) (left) and pattern loadings and 95% percentile confidence intervals of the Western pattern extracted from DDM-Spain data (right).



[a]Congruence coefficient and 95% percentile confidence interval between EpiGEICAM and DDM-Spain pattern loadings.

[b]Correlation coefficient and 95% percentile confidence interval between EpiGEICAM and DDM-Spain pattern scores. All correlations were significant at a 95% confidence level.

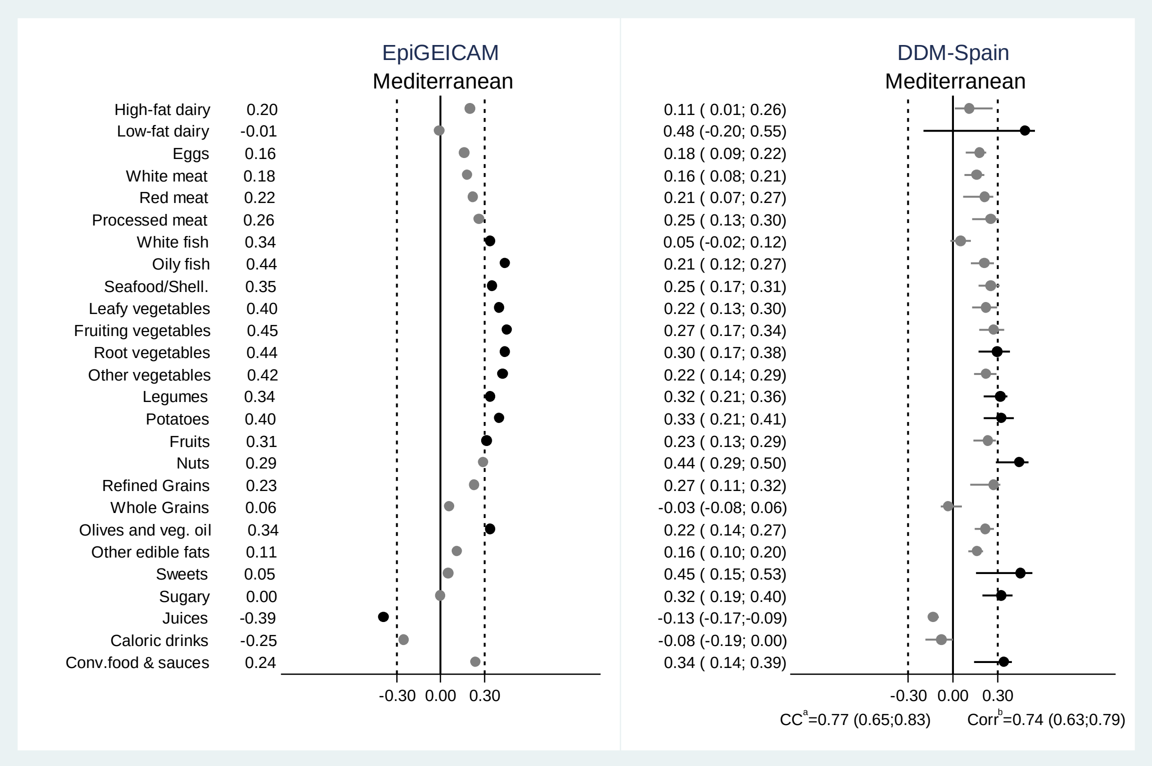
Fig.2 *Pattern loadings of the Prudent dietary pattern extracted from EpiGEICAM study* [(24)] *(left) and pattern loadings and 95% percentile confidence intervals of the Prudent pattern extracted from DDM-Spain data (right).*



[a]Congruence coefficient and 95% percentile confidence interval between EpiGEICAM and DDM-Spain pattern loadings.

[b]Correlation coefficient and 95% percentile confidence interval between EpiGEICAM and DDM-Spain pattern scores. All correlations were significant at a 95% confidence level.

Fig.3 *Pattern loadings of the Mediterranean dietary pattern extracted from EpiGEICAM study* [(24)] *(left) and pattern loadings and 95% percentile confidence intervals of the Mediterranean pattern extracted from DDM-Spain data (right).*



[a]Congruence coefficient and 95% percentile confidence interval between EpiGEICAM and DDM-Spain pattern loadings.

[b]Correlation coefficient and 95% percentile confidence interval between EpiGEICAM and DDM-Spain pattern scores. All correlations were significant at a 95% confidence level.

ONLINE SUPPORTING MATERIAL

Supplemental Methods 1: Methodology used to calculate 95% percentile confidence intervals of the component loadings resulting from the Principal Component Analysis.

PCA presents various difficulties when applying bootstrap that were taken into account:

1. On the one hand, since PCA orders the components by percentage of variability explained, even if the same components are identified over the different bootstrap samples, they might appear in different order in each replication. We selected the most similar component from each replication by calculating the relative difference between all pairs of matrices resulting from comparing EpiGEICAM Western, Prudent and Mediterranean patterns with DDM-Spain components 1-10. For each step and EpiGEICAM pattern we selected the DDM-Spain component showing the smallest relative difference in absolute terms with the original Castelló's et al. (1)
2. On the other hand, two patterns might be explaining the same but be inversely associated with the items they include. For example, we can obtain the Western pattern in one replication and the anti-Western (very similar to Western but with opposite sing in the pattern loadings) in another replication. Those two patterns would explain the same in essence (high adherence to Western≈low adherence to anti-Western) but the change in the sign of the loadings would result in very wide bootstrap percentile confidence intervals that are not capturing the deviation of one result from the other in absolute terms. We overcame this issue by changing the sign of the DDM-Spain components when the relative difference was negative.

1. Castello A, Pollan M, Buijsse B, Ruiz A, Casas AM, Baena-Canada JM, Lope V, Antolin S, Ramos M, Munoz M, et al. Spanish Mediterranean diet and other dietary patterns and breast cancer risk: case-control EpiGEICAM study. British journal of cancer 2014. doi: 10.1038/bjc.2014.434.

ONLINE SUPPORTING MATERIAL

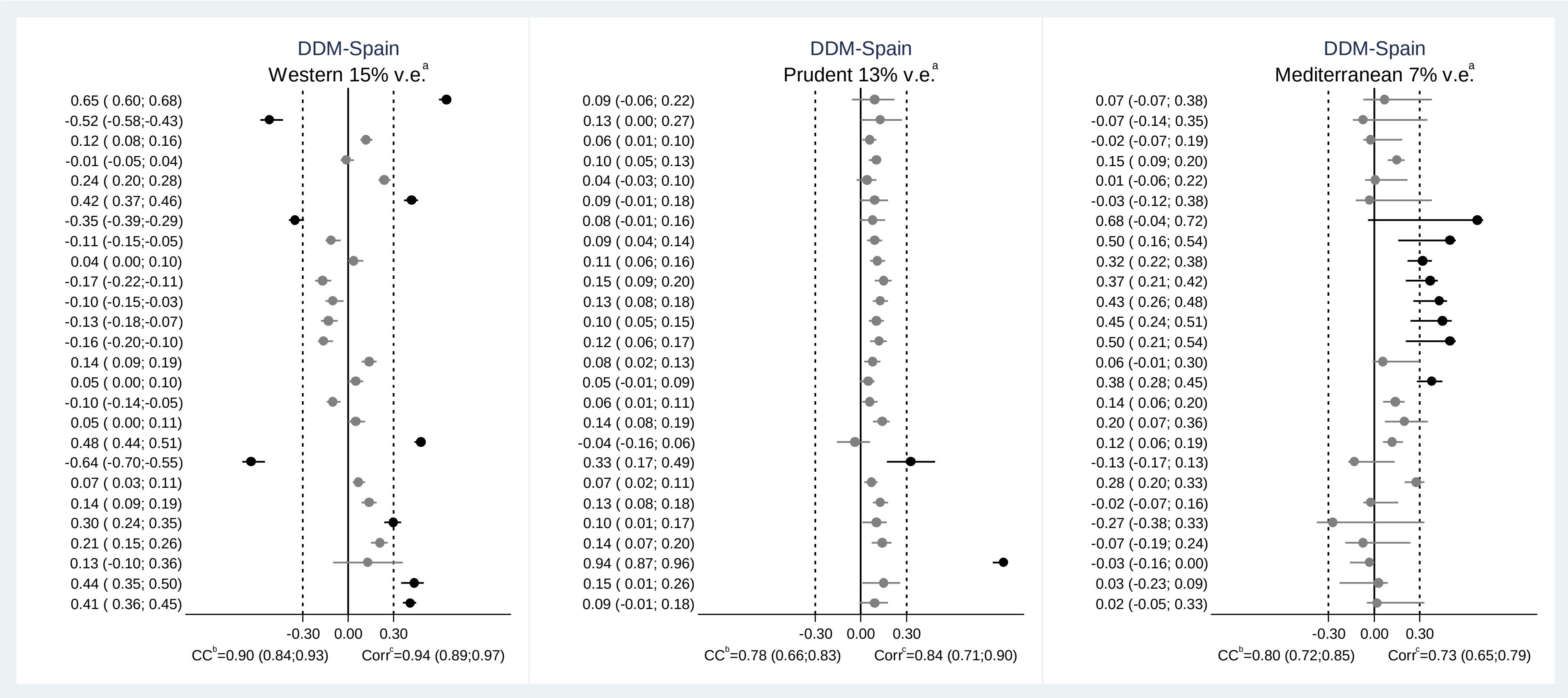
Supplemental Table 1: Food group intakes (gr/day) and component scores from DDM Spain women using all foods from the current FFQ (99 items) and using same foods used in EpiGEICAM (86 items).

	DDM ALL FOODS 99 items		**DDM SAME FOODS 86 items**		
	Median(IQR)		**Median(IQR)**		**Corr[a]**
High fat dairy	48.72(16.03-146.45)		39.69(13.86-132.77)		0.95
Low-fat dairy	332.15(200.00-539.33)		256.98(101.60-521.45)		0.85
Eggs	21.62(7.21-21.62)		21.62(7.21-21.62)		1.00
White meat	38.61(18.90-45.34)		38.61(18.90-45.34)		1.00
Red meat	50.83(26.25-78.20)		50.83(26.25-78.20)		1.00
Processed meat	28.15(21.45-42.90)		6.70(3.26-13.85)		0.64
White fish	19.14(14.36-43.07)		14.36(6.73-43.07)		0.89
Oily fish	30.46(20.14-46.36)		24.62(14.36-39.88)		0.95
Seafood/Shell	14.36(8.60-21.53)		10.55(3.82-14.35)		0.77
Leafy vegetables	60.82(39.40-73.66)		60.82(39.4-73.66)		1.00
Fruiting vegetables	114.21(67.78-156.77)		114.21(67.78-156.77)		1.00
Root vegetables	21.45(7.15-21.45)		21.45(7.15-21.45)		1.00
Other vegetables	133.45(95.85-188.33)		79.89(55.08-111.61)		0.77
Legumes	20.08(20.08-60.23)		20.08(20.08-60.23)		0.89
Potatoes	53.80(17.93-53.8)		53.80(17.93-53.8)		1.00
Fruits	351.11(249.25-507.05)		351.11(249.25-507.05)		1.00
Nuts	2.01(0.00-12.87)		2.01(0.00-12.87)		1.00
Refined grains	93.73(70.49-168.73)		93.73(70.49-168.73)		1.00
Whole grains	0.00(0.00-50)		0.00(0.00-50)		0.99
Olives and veg. oil	30.85(22.87-48.95)		30.85(22.87-48.95)		1.00
Other edible fats	0.00(0-0.68)		0.00(0.00-0.68)		1.00
Sweets	12.50(4.29-30.00)		12.50(4.29-30.00)		1.00
Sugary	10.43(1.79-16.86)		10.43(1.79-16.86)		1.00
Juices	17.88(0.00-99.20)		13.4(0.00-85.8.00)		0.95
Caloric drinks	0.00(0.00-16.75)		0.00(0.00-16.75)		1.00
Conv.food & sauces	24.09(11.50-40.25)		24.09(11.5-40.25)		1.00
Principal components	**Descriptive**	**% Variabilty Explained**	**Descriptive**	**% Variabilty Explained**	
Comp1 Mean(sd)	0.00(2.31)	16%	0.00(2.37)	15%	0.96
Comp2 Median(IQR)	0.34(-2.21-1.92)	15%	0.11(-2.10-1.95)	13%	0.95
Comp5 Mean(sd)	0.00(1.50)	7%	0.00(1.61)	7%	0.85
Comp6 Mean(sd)	0.00(1.27)	5%	0.00(1.42)	6%	0.84

[a]All correlations were significant at a 95% confidence level.

ONLINE SUPPORTING MATERIAL

Supplemental Figure 1: Component loadings resulting from the Principal Component Analysis applied over the DDM-Spain food grouping that uses the same foods included in EpiGEICAM (86 items). Congruence coefficients with the homonymous patterns from EpiGEICAM.



[a]Congruence coefficient between EpiGEICAM and DDM-Spain pattern loadings.
[b]Correlation coefficient between EpiGEICAM and DDM-Spain pattern scores. All correlations were significant at a 95% confidence level.