**Table 1** Comparison of best fit indicators for Latent Class solutions of 2 to 6 classes

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Loglikelihood | AIC | BIC | Entropy | LMR | Cases per class(n) |
| 2-class | -14675.586 | 29535.172 | 29966.615 | 0.719 |  | 369/435 |
| 3-class | -14402.630 | 29083.260 | 29735.114 | 0.748 | 272.956\*\* | 293/266/245 |
| 4-class | -14199.969 | 28771.939 | 29644.204 | 0.836 | 202.661\*\* | 241/257/125/181 |
| 5-class | -14078.180 | 28622.359 | 29715.036 | 0.842 | 121.789\*\* | 92/180/213/144/175 |
| 6-class | -13916.011 | 28392.023 | 29705.111 | 0.872 | 162.169\*\* | 90/169/183/77/142/143 |

\*\**P*<0.01

Note: In order to identify the best classification, we start by assuming that there is only two class for all subjects, and gradually increase the number of classes until we find the best model for fitting data. We compared and analyzed three different indicators: fitting index, classification validity and the number of classes. The information evaluation indicators AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) has always been used as fitting indexes, the lower these values are, the better the model fits. Because the AIC index does not consider the influence of sample size, when the sample size is large will lack asymptotic probabilistic derivation of AIC and as a disadvantage of the AIC index. BIC takes the influence of the number of samples into consideration, so when the number of samples reaches more than 1000, the index of BIC performs is better than AIC. In addition, the Lo-Mendel-Rubin (LMR) likelihood ratio test can also be used to compare the fitting differences of the potential class models. If the *P-*value reaches or is close to a significant level, it indicates that the K class models are better than K-1 category model. Further, the entropy index is used to assess the accuracy of the classification and its value ranges from 0 to 1. As the number of classes increases, entropy tends to decrease. The larger value of entropy means smaller classification errors; when the entropy reaches 0.80, the classification accuracy is over 90%. From **Table 1**, when the model retains 4 classes, the value of BIC is the smallest and the Lo-Mendall-Rubin test reaches a significant level, while the entropy value is ideal. And the proportion of each class after dividing into four categories is also balanced. Therefore, based on the balance of statistical fit and parsimony, we believe that the four classes are the most appropriate.

**Table 2** Food consumption level conditional probabilities of dietary pattern classes

|  |  |  |
| --- | --- | --- |
|  | High consumption (third tertile) conditional probability  | No consumption conditional probability |
| Food group | Chinese traditional | Western | Prudent | Picky | Chinese traditional | Western | Prudent | Picky |
| Rice/Flour | 0.0960 | 0.1590 | 0.0800 | 0.2370 | 0.2580 | 0.4010 | 0.2520 | 0.1670 |
| Cereals | 0.2470 | 0.2770 | 0.3520 | 0.1680 | 0.1200 | 0.1510 | 0.2390 | 0.5610 |
| Fried food | 0.0000 | 0.4570 | 0.1310 | 0.0150 | 0.5000 | 0.1190 | 0.2440 | 0.7320 |
| Meat | 0.0270 | 0.7930 | 0.4660 | 0.1460 | 0.0010 | 0.0080 | 0.0280 | 0.1030 |
| Poultry | 0.2880 | 0.7050 | 0.4560 | 0.2330 | 0.0530 | 0.0000 | 0.2560 | 0.3550 |
| Aquatics | 0.2940 | 0.5540 | 0.5600 | 0.2660 | 0.0080 | 0.0000 | 0.0180 | 0.1320 |
| Eggs | 0.0580 | 0.2210 | 0.0460 | 0.0670 | 0.0310 | 0.0070 | 0.2480 | 0.3640 |
| Milk | 0.0120 | 0.0400 | 0.0080 | 0.0060 | 0.3810 | 0.2640 | 0.4720 | 0.8450 |
| Fruits | 0.2970 | 0.1830 | 0.2450 | 0.1360 | 0.0060 | 0.0000 | 0.0410 | 0.3180 |
| Vegetables | 0.1710 | 0.2730 | 0.2560 | 0.3720 | 0.0000 | 0.0000 | 0.0100 | 0.0400 |
| Soy foods | 0.2510 | 0.4860 | 0.3960 | 0.1810 | 0.0550 | 0.0880 | 0.1670 | 0.3770 |
| Nuts | 0.1440 | 0.2530 | 0.2630 | 0.1110 | 0.1490 | 0.0780 | 0.3400 | 0.6680 |
| Cakes | 0.1310 | 0.3330 | 0.2270 | 0.1030 | 0.3090 | 0.1520 | 0.5060 | 0.7590 |
| SSB | 0.1130 | 0.2510 | 0.0190 | 0.0470 | 0.8870 | 0.7490 | 0.9810 | 0.9530 |
| Fresh juice | 0.1120 | 0.2740 | 0.0590 | 0.0210 | 0.8880 | 0.7260 | 0.9410 | 0.9790 |
| Soft drink | 0.1770 | 0.4720 | 0.0570 | 0.0730 | 0.8230 | 0.5280 | 0.9430 | 0.9270 |
| Pickled foods | 0.1640 | 0.2510 | 0.1170 | 0.2440 | 0.3150 | 0.1880 | 0.4400 | 0.3250 |
| Coffee | 0.0800 | 0.2600 | 0.0810 | 0.0150 | 0.9200 | 0.7400 | 0.9190 | 0.9850 |

**Table 3** Mean intake for each dietary pattern

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Dietary patterns’ characteristics** |  |  |
| **Food group** |  | **Class 1****(Prudent)** | **Class 2****(Chinese** **traditional)** | **Class 3****(Western)** | **Class 4****(Picky)** | ***P*-value** |  |
| Rice/Flour (g/day) |  | 272.23  | 280.10  | 258.24  | **319.89**  | 0.00 |  |
| Cereals (g/day) |  | **26.89**  | 18.44  | 23.24  | 11.07  | 0.00 |  |
| Fried food (g/day) |  | 1.70  | 1.93  | **4.97**  | 0.57  | 0.00 |  |
| Red meat (g/day) |  | 77.27  | 23.91  | **99.62**  | 29.63  | 0.00 |  |
| Poultry (g/day) |  | 15.90  | 9.35  | **24.26**  | 7.27  | 0.00 |  |
| Aquatics (g/day) |  | **49.07**  | 29.69  | 48.77  | 26.50  | 0.00 |  |
| Eggs (g/day) |  | 33.78  | 33.58  | **36.54**  | 24.66  | 0.00 |  |
| Milk (ml/day) |  | 97.91  | 91.35  | **107.73**  | 24.97  | 0.00 |  |
| Fruits (g/day) |  | **133.02**  | 132.38  | 129.89  | 75.41  | 0.00 |  |
| Vegetables (g/day) |  | 69.46 | 57.912 | 61.234 | **76.54** | 0.00 |  |
| Soy foods (g/day) |  | **28.92**  | 19.48  | 24.06  | 16.10  | 0.00 |  |
| Nuts (g/day) |  | 18.01  | 11.46  | **20.98**  | 7.51  | 0.00 |  |
| Cakes (g/day) |  | 13.07  | 10.05  | **18.82**  | 8.48  | 0.00 |  |
| SSB (g/day) |  | 0.32  | 2.40  | **12.32**  | 0.70  | 0.00 |  |
| Fresh juice (g/day) |  | 8.18  | 5.65  | **13.83**  | 3.01  | 0.00 |  |
| Soft drink (g/day) |  | 1.30  | 3.93  | **15.83**  | 3.26  | 0.00 |  |
| Pickled foods (g/day) |  | 2.51  | 3.43  | **5.74**  | 4.31  | 0.01 |  |
| Coffee (50ml/day) |  | 2.52  | 1.75  | **7.70**  | 0.71  | 0.00 |  |
| Energy (g/day) |  | 1906.30  | 1597.00  | 2208.39  | 1585.80  | 0.00 |  |

**Table 4** Characteristics distributed within the identified latent classes

|  |  |  |
| --- | --- | --- |
|  | **Percentage**  |  |
| **Variable** | **Class 1****(Prudent)** | **Class 2****(Chinese traditional)** | **Class 3****(Western)** | **Class 4****(Picky)** |  |  |
| **Educational level** |  |  |  |  |  |  |
| illiterate and primary  | 22.4 | 20.6 | 13.6 | 41.4 | 0.000 |  |
| middle and high school | 35.7 | 36.6 | 31.2 | 39.2 |  |  |
| university and above | 41.9 | 42.8 | 55.2 | 19.3 |  |  |
| **Area** |  |  |  |  |  |  |
| City | 66.0 | 59.1 | 60.8 | 37.0 | 0.000 |  |
| Rural | 34.0 | 40.9 | 39.2 | 63.0 |  |  |
| **Smoking** |  |  |  |  |  |  |
| No  | 41.5 | 45.5 | 31.2 | 50.8 | 0.006 |  |
| Yes1 | 58.5 | 54.5 | 68.8 | 49.2 |  |  |
| **Moderate physical** **activity (minutes/day)** |  |  |  |  |  |  |
| Mean ± SD | 17.49±15.17 | 16.31±13.94 | 15.40±11.18 | 15.47±12.36 | 0.382 |  |
| **Oral contraceptives use** |  |  |  |  |  |  |
| No | 19.5 | 19.8 | 16.8 | 15.5 | 0.615 |  |
| Yes  | 80.5 | 80.2 | 83.2 | 84.5 |  |  |
| **Hormone replacement therapy** |  |  |  |  |  |  |
| No | 95.4 | 96.9 | 93.6 | 98.3 | 0.146 |  |
| Yes4 | 4.6 | 3.1 | 6.4 | 1.7 |  |  |
| **Age at** **menarche** |  |  |  |  |  |  |
| 10-14 | 33.6 | 32.3 | 38.4 | 17.7 | 0.000 |  |
| 15-16 | 36.1 | 38.9 | 38.4 | 38.7 |  |  |
| 17-22 | 30.3 | 28.8 | 23.2 | 43.6 |  |  |
| Mean ± SD | 15.49±1.91 | 15.47±1.81 | 15.31±1.75 | 16.31±1.88 | 0.000 |  |
| **Number of full-term births** |  |  |  |  |  |  |
| 0 | 2.1 | 1.2 | 1.6 | 0.6 | 0.000 |  |
| 1 | 58.5 | 62.6 | 68.0 | 44.2 |  |  |
| 2 | 27.0 | 27.2 | 25.6 | 33.7 |  |  |
|  3 | 12.4 | 8.9 | 4.8 | 21.5 |  |  |
| **Age at first full-term delivery**  |  |  |  |  |  |  |
|  25 | 43.2 | 51.0 | 51.2 | 59.1 | 0.059 |  |
| 25-29 | 53.3 | 47.5 | 46.4 | 38.1 |  |  |
|  >29 | 3.3 | 1.6 | 2.4 | 2.8 |  |  |
| Mean ± SD | 24.92±2.70 | 24.48±2.33 | 24.58±2.39 | 24.19±2.44 | 0.026 |  |
| **Breastfeeding**  |  |  |  |  |  |  |
| No | 30.3 | 29.6 | 27.2 | 35.9 | 0.364 |  |
| Yes | 69.7 | 70.4 | 72.8 | 64.1 |  |  |
| **Energy intake (kcal/day)** |  |  |  |  |  |  |
| Mean ± SD | 1906.30±586.90 | 1597.00±539.50 | 2208.39±780.79 | 1585.80±546.64 | 0.000 |  |
| **Menopause age** |  |  |  |  |  |  |
| Mean ± SD | 49.78±5.07 | 49.40±5.29 | 49.92±4.22 | 48.89±4.86 | 0.436 |  |

Note: we compared the sociodemographic (SES) distributed within the identified latent classes and found the class

differed significantly across sociodemographic characteristics.

**Comparison between LCA and FA**

**Dietary patterns derived by EFA and CFA**

***EFA.*** We conducted an exploratory factor analysis (EFA) on 18 food groups (as category variable) using weighted least squares and factors were derived orthogonal using Varimax rotation. We decided on the number of factors remaining from a combination of the scree plot and the interpretation of the factor loadings. Dietary patterns’ names were given according to the foods with higher loadings and also based on the literature.

***CFA.*** We performed confirmatory factor analysis (CFA) on the dietary factors derived by EFA including only food groups with loadings in absolute value 0.25, allowing food groups to load on multiple factors. If the loading of food groups was <0.25, it will be excluded from CFA and reconstructed to ensure that the lading of each food group0.25.

It should be explained that food group consumption is regarded as categorical variables modeling EFA and CFA. Although using categorical variables will lose some information, skewed data has a greater impact on the model fit. EFA and CFA based on categorical variables modeling had better goodness of fit than continuous variables (**Supplementary Table 5**)

**Table 5** Model fit indicators for factor analysis

|  |  |
| --- | --- |
| Model Fit Indicators | Confirmatory Factor Analysis (CFA) |
| Variable Type | Continuous | Category |
| Chi-Square Test value | 2586.822 | 3951.469 |
| P-Value | 0.000 | 0.000 |
| RESEA | 0.047 | 0.043 |
| CFI | 0.839 | 0.911 |
| TLI | 0.792 | 0.886 |

`RMSEA, Root mean square error of approximation

Note: in exploratory factor analysis, the Chi-Square Test value for continuous variable is 389.016, P-Value<0.000, for category variable is 388.270, for *P*-value<0.000

**FA.** According to the scree plot from EFA, after extracting 5 factors, factors did not contribute much to explain the variance of the data (the first 6 Eigenvalues were 2.57, 1.66, 1.44, 1.29, 1.18, and 1.01). 1st factor loaded high in fried foods and red meat, named Western; 2nd factor loaded high in poultry, eggs and soy foods, named Chinese traditional (short for Chinese); 3rd factor loaded high in cereals, aquatics, milk, fruits, soy foods, nuts, cakes, fresh juice, named Prudent; 4th factor loaded high in vegetables, soy foods, pickled foods, named Picky; 5th factor loaded high in cakes, SSB, fresh juice, soft drinks, pickled foods and coffee, named Sugar.

In CFA analysis, we excluded food groups with EFA loadings <0.25. The factor loadings from EFA and CFA were similar (**Supplementary Table 6**) except for coffee for FA-Picky and fresh juice for FA-Sugar. Hence, for the dietary patterns assessed by CFA, we kept the names given from EFA. After excluding food groups with factor loading<0.25, the model was more concise and the goodness of fit did not decrease.

**Table 6** Selected exploratory and confirmatory factor loadings for the 5-factor model

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | FA-Western | FA-Chinese | FA-Prudent | FA-Picky | FA-Sugar | R2 |
| Food group | EFA | CFA | EFA | CFA | EFA | CFA | EFA | CFA | EFA | CFA |
| Rice/Flour |  |  |  |  |  |  | 0.46 | 0.54 |  |  | 0.29  |
| Cereals |  |  |  |  | 0.45 | 0.35 |  |  |  |  | 0.13  |
| Fried food | 0.78 | 0.78 |  |  |  |  |  |  |  |  | 0.61  |
| Meat | 0.88 | 0.89 |  |  |  |  |  |  |  |  | 0.80  |
| Poultry |  |  | 0.92 | 0.61 |  |  |  |  |  |  | 0.37  |
| Aquatics |  |  |  |  | 0.24 | 0.3 |  |  |  |  | 0.09  |
| Eggs |  |  | 0.41 | 0.64 |  |  |  |  |  |  | 0.42  |
| Milk |  |  |  |  | 0.48 | 0.49 |  |  |  |  | 0.24  |
| Fruits |  |  |  |  | 0.48 | 0.43 |  |  |  |  | 0.18  |
| Vegetables |  |  |  |  |  |  | 0.41 | 0.35 |  |  | 0.12  |
| Soy foods |  |  | 0.26 | 0.29 | 0.32 | 0.33 | 0.25 | 0.3 |  |  | 0.27  |
| Nuts |  |  |  |  | 0.5 | 0.55 |  |  |  |  | 0.30  |
| Cakes |  |  |  |  | 0.42 | 0.43 |  |  | 0.32 | 0.24 | 0.29  |
| SSB |  |  |  |  |  |  |  |  | 0.76 | 0.72 | 0.52  |
| Fresh juice |  |  |  |  | 0.48 | 0.56 |  |  | 0.31 |  | 0.31  |
| Soft drink |  |  |  |  |  |  |  |  | 0.79 | 0.86 | 0.75  |
| Pickled foods |  |  |  |  |  |  | 0.24 | 0.27 | 0.34 | 0.34 | 0.18  |
| Coffee |  |  |  |  | 0.33 | 0.37 | -0.26 |  | 0.45 | 0.36 | 0.34  |

The LCA derived classes had significantly higher means for the corresponding FA-factor score (**Figure 1**), which means LCA-Western class had the highest means for FA-Western, LCA-Chinese class had had the highest means for FA-Chinese, LCA-Prudent had the highest means for FA-Prudent. Besides, the LCA-Western class also had the highest means for FA-Sugar factor scores. The LCA-Picky class had significantly lower than zero for FA-Western, FA-Chinese, FA-Prudent and FA-Sugar and the lowest means for FA-Prudent. Although The LCA-Prudent had higher means for the FA-Prudent factor score, the factor score was not significantly different among LCA-Chinsed and LCA-Western.

****

**Figure 1** Factor scores’ means by latent class, 4-LCA on 5 factor scores.

However, the overall test for the correlations between the 5 factors being zero was significant (*P* < 0.001) and this model had a better fit than the one with uncorrelated factors. Chinese has a significant correlation with Western (r=0.39; *P* <0.001) and Prudent (r=0.37; *P* <0.001); Picky has a negative correlation with Prudent (r=-0.32; *P* <0.001) and Sugar is positive correlated with Western (0.25; *P* <0.001), **Figure 2**.



**Figure 2** Confirmatory factor analysis (CFA)

Additionally, according to our EFA results, the extracted 5 factors only contribute 45.2% explanation of the original variance. Because this type of method performs on the square of simple correlation coefficients between variables, skewed distribution results in the sum of squares of simple correlation coefficients between variables being much smaller than the sum of squares of partial correlation coefficients, and the lower variance contribution will make it challenging to capture the information about the relationship between the variables of interest(1).

Unlike FA, which assumes the same component structure applies to all people and focuses on correlations between dietary measurement items, LCA focuses on identifying important intraindividual and interindividual differences(2). While in the study of dietary patterns, LCA used to identify subtypes of individuals that exhibit similar dietary patterns of characteristics related. Specifically, heterogeneity of diet may result in attenuation of correlations between items, LCA using an inclusive maximum-probability approach to de-attenuates these correlations. Technically, model distinct configurations of heterogeneity within a given sample, which means that observed outcomes with different distributions can be modeled simultaneously(2). Finite mixture modeling captures unobserved heterogeneity, which has a direct physical explanation and is a latent variable(3). The mixture distribution means the heterogeneous across the sample but homogeneous within sub-samples. Given the sample heterogeneity, the target variables have a different probability distribution within each subgroup. LCA uses maximum likelihood estimation to perform computations, which generates a probabilistic model that makes the data "most likely" for a given distribution and thus best describes the data(4).

Through our study, we found that the classification of dietary patterns was roughly similar to the FA approach, which demonstrates LCA and FA identified similar dietary patterns when presented with the same data set. However, LCA seemed better to reveal the heterogeneity of diet in individuals and to classify each participant by using posterior probability accurately.

**Comparison between LCA and DQI**

When we study the relationship between LCA-derived dietary patterns and breast cancer risk, we used the Prudent class as a reference. Since LCA is a highly data-driven method, we choose one of the classes as a reference with a certain degree of subjective. Alternate Mediterranean Diet Score (aMED) as “a priori” approach, could capture specific diet habits of the population under an identified actual dietary pattern. Some extensive cohort studies reported the inverse association for the dietary pattern characterized by the traditional diet among populations in Mediterranean countries of breast cancer risk(5-7), which implies the Mediterranean diet may be a unique dietary combination for breast cancer prevention. More and more summarized review studies are proving and reinforcing this hypothesis(8). According to DQI, we calculated the aMED score of each subject in this study. Furthermore, we conduct a correlation analysis between food group consumption and aMED score and the post posterior probability of LCA-derived dietary classes, **Figure 3**. We found that the aMED score indices seemed similar to the Prudent dietary pattern in terms of its correlation with specific foods.

**Figure 3** Correlations between food consumption and 'a posteriori' dietary patterns and 'a priori' aMED.

**Note**: Compared with 'a priori' approaches based on data-driven, 'a posteriori' methods better capture specific diet habits of the population under an identified actual dietary pattern. The correlations between food group consumption and the conditional probability of dietary patterns are weak, not suggesting that the patterns are not stable. Because the conditional probability from the LCA method itself uses probability correlation coefficients in the calculation process, when we additionally estimate the correlations between them will week the association. The purpose of comparison with a priori method is to observe the reliability of the data-driven method, and we found that the aMED score indices seemed similar to the Prudent dietary pattern in terms of its correlation with specific food groups.

**Selection bias analysis**

From Nov 2013 to Nov 2014, a total of 1410 newly diagnosed breast cancer cases were identified in Wuxi City, but only 1072 cases meeting the inclusion criteria (excluded secondary or recurrent BC and multiple incident cancers, only included those with BC as the first original malignancy diagnosed) and 818 of them were recruited in this study, the response rate is 76.3% (818/1072). Moreover, 1,072 controls were selected and 935 of them participated, with a response rate of 87.2% (935/1072). Therefore, selection bias is likely present in this study. We analyzed the impact of selection bias on the results, as below:

**For our study:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID |  | actual situation | Observed situation | Missingness |
|  | X(exposure=dietary patterns) | Y (0=control, 1=case) | Y\* (0=control, 1=case,NA=missing) | RY(0=intact,1=missing) |
| 1 | 6 | 1 | 1 | 0 |
| 2 | 7 | 0 | NA | 1 |
| 3 | 5 | 1 | 1 | 0 |
| 4 | 8 | 1 | NA | 1 |
| 5 | 5 | 0 | 0 | 1 |
| . |  |  |  |  |
| . |  |  |  |  |
| . |  |  |  |  |

NOTE:

**Figure 4** Missing data status

1. Missing Completely At Random (MCAR)
2. Missing At Random (MAR)
3. Missing Not At Random (MNAR)

Note:

 means Missingness does not depend on either observed or unobserved quantities

 means Missingness depends on observed quantities but conditionally independent given observed quantities

 means Missingness may depend on unobserved quantities

The effect of selection bias can be recognized by missingness analysis. It is assumed that 1072 cases are the number of cases we want to investigate. Actually, 818 cases were recruited and 254 cases are missing. Missingness can be divided into three types. The first type is **Figure 4 (A)**, the missingness of patients is completely random, which means the occurrence of the missing is not related to the exposure (X) and the outcome (Y). In this case, there will be no bias in the estimation of the association between X and Y. The second type is **Figure 4 (B)**. Missingness is related to exposure (X). In this case, the missingness is conditional independence (the condition is to control the exposure factor), so it is also unbiased after the exposure factor is controlled.

The biased estimate is the third type, **Figure 4 (C)**. The missingness is related to the outcome (Y), which means the cases refuse to participate in the investigation because of their own reasons. In our study, we will provide free medical examination and condition consultation for the participating patients. We believe that it is attractive for most patients. After the telephone follow-up part of the patients that refused to participate in the study, the reasons for their refusal to participate were mostly due to area factor and educational factor (living in the countryside, so it is inconvenient to go too far for medical examination, and the education level was low that they think they could not complete the investigation). Through the existing data analysis, a similar population has the highest probability of belonging to the Picky class. Based on the existing data, we fill the missingness with random sampling from existing samples based on their characteristics (live in a rural area and with primary education level), using fully conditional specification multivariate imputation by the chained equations method(9). We found that the risk effect of the Picky class on breast cancer has been further strengthened.

**Reference**

1. MacCallum RC, Zhang S, Preacher KJ *et al.* (2002) On the practice of dichotomization of quantitative variables. *Psychol Methods* **7**, 19-40.

2. Berlin KS, Williams NA Parra GR (2014) An introduction to latent variable mixture modeling (part 1): overview and cross-sectional latent class and latent profile analyses. *J Pediatr Psychol* **39**, 174-187.

3. Muthen B & Shedden K (1999) Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* **55**, 463-469.

4. Lanza ST, Collins LM, Lemmon DR *et al.* (2007) PROC LCA: A SAS Procedure for Latent Class Analysis. *Struct Equ Modeling* **14**, 671-694.

5. Masala G, Bendinelli B, Assedi M *et al.* (2017) Up to one-third of breast cancer cases in post-menopausal Mediterranean women might be avoided by modifying lifestyle habits: the EPIC Italy study. *Breast Cancer Res Treat* **161**, 311-320.

6. Sieri S, Krogh V, Pala V *et al.* (2004) Dietary patterns and risk of breast cancer in the ORDET cohort. *Cancer Epidemiol Biomarkers Prev* **13**, 567-572.

7. Trichopoulou A, Bamia C, Lagiou P *et al.* (2010) Conformity to traditional Mediterranean diet and breast cancer risk in the Greek EPIC (European Prospective Investigation into Cancer and Nutrition) cohort. *Am J Clin Nutr* **92**, 620-625.

8. Bloomfield HE, Koeller E, Greer N *et al.* (2016) Effects on Health Outcomes of a Mediterranean Diet With No Restriction on Fat Intake: A Systematic Review and Meta-analysis. *Ann Intern Med* **165**, 491-500.

9. Van Buuren S, Brand JP, Groothuis-Oudshoorn CG *et al.* (2006) Fully conditional specification in multivariate imputation. *J Stat Comput Simul* **76**, 1049-1064.