NUTRITION RESEARCH 35 (2015) 265–268

Available online at www.sciencedirect.com

ScienceDirect

www.nrjournal.com

Communication

Modest validity and fair reproducibility of dietary patterns derived by cluster analysis

Anna N. Funtikova [a,b,c], *Alejandra A. Benítez-Arciniega* [d], *Montserrat Fitó* [a,e], *Helmut Schröder* [a,b,*]

[a] *Cardiovascular Risk and Nutrition Research Group (CARIN-ULEC), Program of Research in Epidemiology and Public Health, IMIM (Institut de Recerca Hospital del Mar), Barcelona, Spain*
[b] *CIBER Epidemiology and Public Health (CIBERESP), Instituto de Salud Carlos III, Spain*
[c] *Food and Nutrition PhD program, University of Barcelona, Barcelona, Spain*
[d] *Faculty of Medicine, Autonomous University of Mexico State, Toluca, Mexico*
[e] *CIBER Physiopathology of Obesity and Nutrition (CIBEROBN), Instituto de Salud Carlos III Spain*

ARTICLE INFO

Article history:
Received 29 July 2014
Revised 30 December 2014
Accepted 30 December 2014

Keywords:
Validation
A posteriori dietary patterns
Repeatability
Cluster analysis
Food frequency questionnaire
Reproducibility

ABSTRACT

Cluster analysis is widely used to analyze dietary patterns. We aimed to analyze the validity and reproducibility of the dietary patterns defined by cluster analysis derived from a food frequency questionnaire (FFQ). We hypothesized that the dietary patterns derived by cluster analysis have fair to modest reproducibility and validity. Dietary data were collected from 107 individuals from population-based survey, by an FFQ at baseline (FFQ1) and after 1 year (FFQ2), and by twelve 24-hour dietary recalls (24-HDR). Repeatability and validity were measured by comparing clusters obtained by the FFQ1 and FFQ2 and by the FFQ2 and 24-HDR (reference method), respectively. Cluster analysis identified a "fruits & vegetables" and a "meat" pattern in each dietary data source. Cluster membership was concordant for 66.7% of participants in FFQ1 and FFQ2 (reproducibility), and for 67.0% in FFQ2 and 24-HDR (validity). Spearman correlation analysis showed reasonable reproducibility, especially in the "fruits & vegetables" pattern, and lower validity also especially in the "fruits & vegetables" pattern. κ statistic revealed a fair validity and reproducibility of clusters. Our findings indicate a reasonable reproducibility and fair to modest validity of dietary patterns derived by cluster analysis.

1. Introduction

Dietary patterns are increasingly used by researchers studying the relationship between diet and diseases [1]. Cluster analysis is one of the methods used to define dietary patterns. Although this method easily defines comprehensive dietary patterns [2], it has a limited comparability of patterns between different data sources, time points, and study samples [3]. This requires careful analysis of the validity and reproducibility of the patterns defined using cluster analysis.

Abbreviations: 24-HDR, 24-hour dietary recalls; FFQ, food frequency questionnaire.

* *Corresponding author.* Cardiovascular Risk and Nutrition Research Group (CARINULEC), Institut de Recerca Hospital del Mar (IMIM), Biomedical Research Park, c/Doctor Aiguader 88, 08003 Barcelona, Spain. Tel.: + 34 933 160 720; fax: +34 933 160 796.
E-mail address: hschroeder@imim.es (H. Schröder).

http://dx.doi.org/10.1016/j.nutres.2014.12.011

The hypothesis of the study was that the dietary patterns derived by cluster analysis have fair to modest reproducibility and validity. In the present study, we aimed to analyze the validity and reproducibility of dietary patterns derived from a food frequency questionnaire (FFQ) and 24-hour recall as a reference method.

2. Methods and materials

This validation study included 150 men and women, aged 30 to 80 years, a consecutively selected sample from a population-based cross-sectional survey carried out in Girona (Spain) in 2005. A total of 107 (71.3%) participants completed the FFQ at baseline (FFQ1) and at 1-year follow-up (FFQ2) and provided at least ten 24-hour dietary recalls (24-HDR). The general characteristics of participants did not differ from the initial sample [4]. All participants were duly informed and provided signed consent to participate in the study. The project was approved by the local Ethics Committee (CEIC-PSMAR, Barcelona, Spain).

Dietary intake data were determined by the FFQ [5] at baseline and follow-up. The 24-HDR (reference method for validity analysis) were collected monthly over a 12-month period by a trained telephone interviewer. At least ten 24-HDR, including minimum 5 weekdays and 1 weekend day, were required for inclusion in analysis. Participants were not alerted to the dates when they would be interviewed.

The K-mean cluster algorithm was used to derive dietary patterns from FFQ1, FFQ2, and 24-HDR. All food items from the FFQ and 24-HDR were combined into 45 food groups according to similarities in their nutritional content. The larger number of food items contained in the 24-HDR was grouped according to the FFQ food groups to be able to run the same cluster analysis for both instruments. Clusters were based on consumption of food groups in grams. Several runs of cluster formation were performed to establish the best cluster configurations. Criteria for cluster solutions were nutritional meaningfulness and a reasonable sample size (ie, every cluster contained at least 5% of the study sample). This solution was confirmed by the tree diagram resulting from the Ward method of cluster analysis. Finally, discriminant function analysis was carried out to examine the stability and classification ability of the cluster solution. The optimal cluster solutions derived from FFQ1and FFQ2 contained 3 clusters. The 24-HDR solution contained 2 clusters. One of the FFQ clusters had less than 5% of the population (1 participant in FFQ1 and 4 participants in FFQ2); these 5 individuals were not included in the cluster analysis of FFQ data.

2.1. Statistical analyses

Mean values and proportions of general characteristics are presented according to cluster membership (Table 1). To test the hypothesis of the study, we used contingency tables for cross-tabulation analysis between clusters derived from (*a*) the FFQ1 and FFQ2 (cluster reproducibility) and (*b*) the FFQ2 and 24-HDR (cluster validity) to identify participants who were categorized consistently (the same cluster) and inconsistently (the opposite cluster). Relative agreement was appraised by Spearman correlation coefficient and Cohen κ coefficient [6]. Spearman correlation analysis of food group consumption was used to determine reproducibility (FFQ1-FFQ2) and validity (FFQ2–24-HDR) of dietary patterns among participants with concordant classification between the dietary data sources (Table 2). κ statistic values were calculated, comparing clusters derived from FFQ1, FFQ2, and 24-HDR. Differences were considered significant if $P < .05$. The SPSS for Windows version 15 (SPSS, Inc., Chicago, IL) was used for statistical analysis.

3. Results

We identified 2 dietary patterns, "fruits & vegetables" and "meat," common to all dietary data sources, FFQ1, FFQ2, and

Table 1 – General characteristics of the clusters from FFQ1 at baseline, FFQ2 at follow-up, and multiple 24-HDR [a]

Variables	FFQ1			FFQ2			24-HDR		
	Fruits & vegetables n = 66	Meat n = 40	P [b]	Fruits & vegetables n = 72	Meat n = 30	P [b]	Fruits & vegetables n = 72	Meat n = 34	P [b]
Women (%)	65.2	22.5	<.001	55.6	40.0	.152	62.5	20.6	<.001
Age (y)	61.7 (1.4)	52.5 (1.8)	<.001	60.0 (1.4)	54.1 (2.1)	.022	61.0 (1.3)	52.5 (2.0)	.001
Education [c] (%)	56.1	75.0	.050	59.7	71.0	.278	54.8	79.4	.014
LTPA	273	236	.752	264	276	.681	263	282	.538
(METs ·min/d)	164, 414	139, 481		159, 419	161, 457		160, 409	152, 457	
Smokers [d] (%)	6.1	33.3	<.001	12.5	24.1	.147	9.7	30.3	.008
BMI (kg/m^2)	27.6 (0.52)	27.8 (0.68)	.829	27.3 (0.5)	28.4 (0.8)	.264	27.6 (0.5)	27.7 (0.7)	.903
Obesity (%)	30.3	20.5	.273	26.8	30.0	.740	26.8	26.5	.975

BMI, body mass index; LTPA, leisure-time physical activity; METs, metabolic equivalents.

[a] Means and SDs for continues variables (age and BMI); proportions for categorical variables (women, education, smokers, obesity); median and 25th and 75th percentiles in LTPA.

[b] P values were obtained by analysis of variance, Mann-Whitney *U*, and Pearson χ^2 for normal continuous, nonnormal continuous, and categorical variables, respectively.

[c] More than secondary school education.

[d] Active smokers or ex-smokers less than 1 year.