STRATEGIES FOR ANALYSING NUTRITIONAL DATA FOR EPIDEMIOLOGICAL PURPOSES - FOOD PATTERNS BY MEANS OF CLUSTER-ANALYTICAL PROCEDURES

H. Boeing, German Cancer Research Center*, Heidelberg, FRG, **U. Klein**, **A. Hendrichs**, Federal Health Office*, Berlin, FRG, **U. Oltersdorf**, University of Giessen, FRG, **A. Bodenstedt**, University of Giessen, FRG.

ABSTRACT: Results of a study using cluster analysis to identify structures in daily food consumption is presented. The data originate from nutritional records of 86 individuals collected in two seasons. Three strategies were followed when performing the cluster analysis by (1) using different clustering procedures, (2) different lists of food groups, i.e. by including or excluding beverages, and (3) varying the number of clusters. The investigation lead to relatively stable and plausible clusters representing food patterns prevailing in the study population. Of all cluster solutions a 5-cluster partition based on 16 food groups was given priority. According to their consumption profiles the clusters were labeled as "meat", "bread", "fruit", "wine", and "juice" cluster. For this solution the cluster-specific nutrient intake was calculated revealing two well provided clusters compared to the average, two less well provided cluster, and one cluster corresponding more or less to average nutrient intake.

INTRODUCTION

Nutritionists and specialists in other disciplines are becoming increasingly aware of the complexity of nutrition. In contrast to this fact this complexity is rarely addressed in empirical studies but only specific biological aspects of nutrition such as the intake of nutrients considered. Statistical computations show that considering many nutrients simultaneously in one analysis is often impossible. If single food items or groups are of interest the situation becomes more complicated because of their nearly infinitive number.

A strategy to cope with these problems is to look for inherent structures in the

----

* Current affiliations
requests for reprints should be sent to: Dr. H. Boeing, German Cancer Research Center, Institute of Epidemiology and Biometry, Im Neuenheimer Feld 280, 6900 Heidelberg, FRG

data, i.e. food consumption, and to describe nutrition with a few indices that integrated the various aspects of individual nutrition behavior (1). Such indices can be called food pattern defining group-specific nutrition behavior across individuals and separating the study group in subgroups. Important tools in deriving food pattern from food intake data are explorative multivariate procedures such as factor and cluster analysis (2). The importance of this kind of analysis which permits applications specific for nutrition research has been stressed recently (3,4,5).

In this paper findings of a study using cluster analysis to identify structures in daily food consumption are presented. The data originate from nutritional records collected during a pilot study carried out in Gießen. It was our intention to show and discuss the approach. A much broader data base is needed to come up with detailed results especially when food pattern will be evaluated for different groups (1). We hope that exploring the structure of food intake by multivariate techniques will enhance our understanding of human nutrition and, subsequently, its relation to health.

METHODS AND RESULTS

Food consumption data were collected in 1981 within the framework of a pilot project (EMSIG - Ernährungs-Modell-Studie in Giessen) in which techniques for collecting and evaluating nutritional and activity behavior for a community-oriented longitudinal study were established and tested (6,7).

The EMSIG-sample, a total of 112 participants, consists of adults at Giessen, a university town near Frankfurt. The subjects were 17 to 65 years old. Three quarters of these were volunteers who responded to appeals on notice boards, advertisements in newspapers and handbills. One quarter of the participants originated from a random sample taken in a residential area of the city of Giessen.

The baseline examination began with a structured "nutritional interview" including an extensive list of sociodemographic variables and further information of nutritional relevance. A 24-h recall technique was then used to train the participants in the techniques of recording diet. Subsequently, the participants were asked to keep diaries about their food intake over several days. Information on associated variables such as shopping behavior, knowledge about nutrition, attitude towards nutrition and psycho-social factors was obtained in subsequent interviews. Anthropometric, clinical and biochemical data (8) were obtained by medical examination. In addition many[1] participants kept a three-day activity protocol. The food consumption data were collected at two different occasions. The first collection was performed in April 1981 and the second measurement taken in September 1981. 46 women and 40 men participated in this part of the study and recorded their dietary intake over a period of 11 day on average. They were selected for the following investigation.

For the collection of the data about food consumption, a special meal-oriented recording form was developed (9). The top of this form covered the situation associated with food consumption, the bottom was used to record the consumed foods in an open-ended format. Participants were requested to record their food items in the course of a meal as exactly as possible, e.g. by noting fat content or other features specifying a single food. The portion size was obtained in a semiquantitative manner by using household measures. Graphic models were additionally given.

A 4-digit food code connected with a nutrient data base and a list of common recipes was used to process the food diary information (10). The translation of household measures into grams utilized a list of portion sizes typical for households. After determination of the nutrient intake all data were stored in a SIR data base (11).

A suitable indicator to describe the food consumption of the participants was developed for the subsequent multivariate analysis. Any indicator based on absolute consumption figures would be dominated by the amount. Sex differences or differences in activity may become important discriminators. The relative proportion of a food item on the overall food consumption was therefore selected as describing the participants in their consumption habits. However, beverages including milk were not included in the calculation of the overall food consumption because of their heavy weight compared to regular food stuffs but were also related to the overall food consumption figure.

The proportion on overall consumption was calculated for larger groups of food items. The criterion for this grouping process was either the biological or the food technology similarity. Thus, all sorts of bread were assigned to a food group "bread", all sorts of vegetables to a food group "vegetables". Potatoes were taken as a group of its own. Sugar, together with cake, pudding and chocolate, was assigned to a group "sugar-containing foods".

Finally, the exploratory analysis of food consumption patterns was based on two lists of food groups (table I). The first list comprising 16 food groups had the highest proportion on overall consumption (table I, list 1). This list of food groups was supplemented by another 7 food groups of minor consumption. All

Table I: Food groups and their proportion on overall food consumption (excluding beverages) as used in cluster analysis (EMSIG–participants with food diaries (n=86))

food group	Proportion on total food consumption (excluding beverages) in %
List 1: The 11 most popular food groups	
bread	14.5
vegetable	12.8
fruit	10.5
sour milk products	7.5
meat	7.1
sugar containing food items	6.0
sausage	6.0
potatoes	5.8
eggs	4.0
spreading fat	2.9
cheese	2.6
	sum 79.7
The 5 beverages	
beer	24.0
mineral water	14.9
wine	11.4
soft drinks and juices	10.4
milk and milk drinks	9.4
List 2: The less popular food groups	
poultry	1.8
fish	1.2
cream	1.2
rice and nudels	1.0
flour	0.9
ice cream	0.9
marmelade	0.9
	sum 7.9

18 food groups covered 87.6% of the food consumption.

Cluster analyses were based on several data matrices, a 86x16 matrix (food group list 1), a 86x11 matrix (food group list 1 minus beverages), a 86x23 matrix (food group list 1 and 2), and a 86x18 matrix (food group list 1 and 2 minus beverages).

Cluster analysis as a multivariate statistical tool belongs to the taxonometric or classification methods. Starting point of the analysis is a rectangular N x M data matrix, in which the investigational units N means persons, cases, etc. and M the variables. The purpose of any cluster analysis is to group the investigational units in such a manner as to render the group members as similar as possible to each other according to a distance parameter while the different groups should be as distant as possible from each other (12). In our case it was to be tested whether in the EMSIG sample clusters exist with similar consumption patterns.

The best known measure for similarity or distance, respectively, is the Euclidian distance. Subsequently, the investigational units are grouped on the basis of this similarity matrix which was compiled according to the information containing in the various consumption indices. For the grouping process a large number of cluster procedures is available including hierarchical, divisive and iterative algorithms. Iterative procedures were preferred to hierarchical procedures because they provide generally more efficient solutions. For our calculations the cluster analysis program package CLUSTAN written by Wishart was used (13). From this program package hierarchical and iterative procedures were selected.

Critique was raised against cluster analysis that uncontrollable subjective decisions influence and bias the results. In order to avoid this as far as possible, three counter-strategies were followed in our calculations. First, the stability of the grouping results was controlled by using different clustering procedures. Here, two hierarchical methods "Ward" and "Average Linkage" as well as the iterative method "Relocate" were used (13). Second, the stability of the grouping results was controlled by using different lists of food groups, i.e. including and excluding beverages. Third, the variation in number of clusters constituted a way to test the stability of the grouping results.

Of all solutions the 5-cluster partition obtained by the iterative method "Relocate" on the basis of 16 food groups was given priority and can be described as follows (table II):
Cluster 1: This cluster with 21 participants covers about one quarter of the sample (24% respectively) and is marked by a high meat consumption. Probably meat plays an important role in warm meals above all; cheese and fruit, which are rather to be allocated to cold meals, are consumed to a lesser degree. We assigned this cluster the term "meat pattern".
Cluster 2: The combination of bread, spreading fat and beer is typical for this cluster. Vegetables are consumed to a lesser extent. This may point to the fact that the participants consume cold meals more than the average. One might talk of a "bread pattern". With 11 participants, this cluster covers 13 % of the sample.
Cluster 3: Almost half of the participants joined this cluster (41 participants or 48%, respectively). The consumption of these participants can be characterized by a high vegetable, fruit, milk, and sour milk product consumption. This consumption pattern was called "fruit pattern".
Cluster 4: Most striking for this cluster is the above the average consumption of sausage, cheese and wine. Nine participants (9% respectively) are assigned to this "wine" pattern.
Cluster 5: 4 participants (or 4%, respectively) stand out by an above average

Table II: Average food consumption (in % of overall food consumption excluding beverages) in the 5 cluster groups (Relocate-solution with 16 food groups)

cluster[1)]	"meat" n=21	"bread" n=11	"fruit" n=41	"wine" n=9	"juice" n=4	total population n=86
Bread	12.8	25.8	12.9	12.6	17.0	14.5
Vegetable	13.4	7.0	14.3	12.5	11.3	12.8
Fruit	6.5	7.6	14.6	9.2	5.7	10.5
Sour milk products	4.0	3.8	10.2	5.3	4.5	7.5
Meat	13.1	5.6	4.9	5.3	5.4	7.1
Sugar containing food items	7.2	5.8	5.9	5.9	4.9	6.0
Sausage	5.6	7.9	4.3	9.0	10.5	6.0
Potatoes	7.9	6.8	4.4	8.5	7.0	5.8
Eggs	4.3	4.5	3.2	3.0	7.4	4.0
Spreading fat	3.9	5.8	2.3	2.6	2.0	2.9
Cheese	0.8	3.4	3.1	3.4	1.7	2.6
Beer	39.6	60.3	4.5	24.3	4.7	24.0
Mineral water	12.5	6.8	15.0	9.8	47.5	14.9
Wine	7.4	7.3	6.0	50.0	6.8	11.4
Soft drinks and juices	11.4	5.8	7.5	6.6	51.8	10.4
Milk and milkdrinks	4.5	7.9	10.0	7.8	1.7	9.4

[1)] for description of cluster partitions see text

consumption of non-alcoholic beverages such as lemonade, fruit juices and mineral water. Evidently, eggs and sausage are also consumed to a greater extent than on average. The consumption of spreading fat, milk and fruit is atypical. With 4 participants, this "juice pattern" plays but a subordinate role in the sample.

The just described structure can be more or less clearly recognized in other solutions (data not shown). Clustering with 23 instead of 16 variables produced similar solutions (food group list 1 and 2). The same went for the "Relocate" solutions based on 11 (food group list 1 without beverages) or 18 food groups (food group list 1 and 2 without beverages). However, the omission of beverages led to slightly different results with regard to the formation of some sub-groups.

In all, it was, therefore, possible to allocate participants to particular with several cluster approaches recognizable consumption patterns. The cluster labels were oriented on the typical consumption profiles. In table III, the cluster-specific nutrient intake is presented.

A comparison of the nutrient profiles revealed a well provided cluster compared

Table III: Nutrient intake per day by cluster (Relocate solution with 16 food groups)

cluster[1]	"meat" n=21	"bread" n=11	"fruit" n=41	"wine" n=9	"juice" n=4	total population n=86
Nutrient						
Protein (g)	82	69	84	77	85	81
Fat (g)	112	112	110	108	112	109
Carbohydrates (g)	222	227	245	199	270	232
Crude fiber (g)	6	5	8	6	6	6
Cholesterol (mg)	460	421	401	395	652	431
Sodium (g)	3.1	2.5	3.1	2.3	2.9	2.9
Potassium (g)	2.9	2.1	3.6	4.1	2.9	3.3
Calcium (mg)	454	611	822	884	550	699
Iron (mg)	14	12	14	14	15	13
Vitamin A (mg)	0.81	0.87	0.98	0.81	1.07	0.91
Vitamin B_1 mg)	1.12	0.98	1.18	1.04	1.28	1.14
Vitamin B_2 (mg)	1.33	1.20	1.48	1.25	1.69	1.43
Vitamin C (mg)	87	57	142	84	84	108

[1]for description of cluster partitions see text

to the average, the "fruit" cluster (cluster 3) and a cluster showing a high vitamin intake, the "juice" cluster (cluster 5). They were opposed by two less well provided cluster, the "bread" cluster (cluster 2) and the "wine" cluster (cluster 4). The "meat" cluster corresponded more or less to the average nutrient intake. For a more detailed discussion of the respective nutrient intake of the food consumption patterns it would be necessary to analyse the food consumption more closely.

Food patterns may be imbedded in a specific social settings (14). Therefore, it was investigated inhowfar cluster members show a uniform sociodemographic

Table IV: Age group and sex by cluster (Relocate solution with 16 food groups)

cluster[1] in % / absolute number	"meat" n=21	"bread" n=11	"fruit" n=41	"wine" n=9	"juice" n=4
Age group 15–30	43% / 9	36% / 4	27% / 11	56% / 5	75% / 3
Age group 31–45	33% / 7	36% / 4	39% / 16	22% / 2	25% / 1
Age group 46–65	24% / 5	27% / 3	34% / 14	22% / 2	–
Males	52% / 11	91% / 10	34% / 14	33% / 3	50% / 2
Females	48% / 10	9% / 1	66% / 27	67% / 6	50% / 2

[1]for description of cluster partitions see text

profile and, in this sense, constitute consumption types (table IV). Two thirds of women and one third of men form the "fruit" cluster. In the "wine" cluster the percentage of women is one third, that of men two third, the greater part of them being of younger age.

DISCUSSION

It was aim of this study to identify, based on food diaries covering 11 days on average, the food pattern of 86 participants living in Gießen. Food patterns are understood to be a particular constellation of food intake prevailing in a group of people.

The empirical analysis started with the hypothesis that, while food habits are of individualistic character, they nevertheless show group-specific similarities. Provided this hypothesis is valid, it should, therefore, be possible to classify our participants into groups with similar food consumption indices by means of suitable analytical methods. We decided for cluster analysis. After formation of food groups, five consumption patterns could be recognized. They are named as "meat", "bread", "fruit", " wine", and "juice" cluster (pattern). Each cluster has a specific nutrient intake profile. Some of them may induce health risks in a long run.

Potential critique regarding the validity of the findings might be on the one hand (1) data-oriented, or (2) method-oriented.

Ad (1): The sample is not representative for the Federal Republic but typical for of the middleclass of a medium sized university town. Therefore, the results cannot be considered as generally prevailing. However, it is reminded to the fact that the food intake diaries covered 11 days on average and two seasons. This was thought to warrant that most food items which are consumed during the year were reflected in the data.

Ad (2): Critics about multivariate statistics argue that its results are liable to methodological artefacts (14). However, it has to be pointed out that our cluster-analytical evaluations produced highly comparable results. Even when varying the number of variables, e.g. if beverages were introduced in the cluster analysis by omission or inclusion, the food cluster remained largely stable. In one point, however, this empirical analysis is exposed to a certain danger of artefact. When using the Euclidian distance as a measure of similarity it is known that variables sharing a common variance gain greater weight in clustering. In general, this effect can be avoided by a suitable combination of the variables with common variance before starting the cluster-analytical procedures. Without an intensive evaluation of the data in this direction, e.g. by factor analysis, judgment of whether the grouping of foods into food groups met this criterion is difficult.

In conclusion, there is strong evidence of group-specific interindividual food consumption patterns. However, the patterns identified here cannot claim to be generalizable or representative. Different food pattern may prevail in different populations, different "types" of people, for different times and different situations in life. It has been shown that cluster analytical procedures can help classifying this complex reality.

ACKNOWLEDGEMENT: We would like to thank Mrs Edinger for preparing the manuscript.

REFERENCES

1. Oltersdorf U., Boeing H., Hendrichs A., Bodenstedt A. Strategies for analysing nutritional data for epidemiological purposes: Conceptual framework. (submitted)

2. Good I.J. The philosophy of exploratory data analysis. Philosophy of Science 50, 283 (1983)

3. Schwerin H.S., Stanton J.L., Smith J.L., Riley A.M., Brett B.E. Food, eating habits, and health: A further examination of the relationship between food eating patterns and nutritional health. Am. J. Clin. Nutr. 35, 1319 (1982)

4. Stellman S.D. Chairman's remarks. National Cancer Institute Monograph No 67; 145 (1984)

5. Akin J.S., Guilkey D.K., Popkin B.M., Fanelli M.T. Cluster analysis of food consumption patterns of older Americans. J. Amer. Diet Assoc. 86, 616 (1986)

6. Bodenstedt A.A., Oltersdorf U., Boeing H., Hendrichs A., Behrens U. Erfassung und Deutung des menschlichen Ernährungsverhaltens "Ernährungsmodell-Studie in Gießen" (EMSIG) (Assessment and evaluation of nutrition behaviour "Nutrition Model Study in Gießen" (EMSIG)). Research report. Institute of Rural Sociology, University of Gießen, 1983

7. Bodenstedt A.A., Oltersdorf U., Hendrichs H., Boeing H. M.A.R.S. - Multiple

automatic regulatory system. A comprehensive theoretical approach to empirical studies on nutrition behaviour. In "Measurement and determinants of food habits and food preferences" (Edited by Diehl J.M., Leitzmannn C.), EURO-NUT Report 7, p 294, Ponsen & Looyen Wageningen NL, 1985

8. Sämann U., Kunter M., Bodenstedt A., Boeing H., Hendrichs A., Oltersdorf U. Die Beurteilung des Ernährungs-Zustandes erwachsener Deutscher mittels anthropometrischer Messungen unter besonderer Berücksichtigung der Aspekte Körperzusammensetzung und Körpergestalt. Analyse von Daten der Ernährungsmodelstudie in Gießen (EMSIG). (The assessment of the nutritional status of German adults by anthropometric measurement with special consideration of the Nutrition Model Study in Gießen (EMSIG). Akt. Ernähr. 195 (1984)

9. Boeing, H., Martinez L. Die Ermittlung der Nahrungs- und Nährstoffzufuhr in der epidemiologischen Forschung. (Determination of food and nutrient supply in epidemiological research). Akt. Ernähr. 11, 101 (1986)

10. Boeing H., Bodenstedt A., Hendrichs A., Klein U., Lamberth K., Oltersdorf U., Wankmueller M.. Gießener Liste aller Nahrungsmittel und -zubereitungen (GLANZ) - Ein Programmsystem zur Auswertung von Ernährungserhebungen (Gießen list of foods and recipes (GLANZ) - A program for the analysis of nutrition surveys). Ernährungsumschau 29, 232 (1982)

11. Robinson B.N., Anderson G.D., Cohen E., Gadzik W.F., Karpel L.C., Miller A.H., Stein J.R. SIR, Scientific Information Retrieval Users's Manual, Version 2. Evanston K., 1980

12. Everitt B., Cluster analysis, London 1974

13. Wishart D., CLUSTAN1C. User manual, University College, London 1975

14. Hendrichs A. Ernährung als Gesundheitsrisiko. Eine Fallstudie psychosozialer Bestimmungsgründe des Verzehrs "gesunder" Nahrungsmittel (Nutrition as health risk: A case study of psychosocial determinants of consumption of "healthy" food items). Campus Verlag, Frankfurt New York, 1987

Accepted for publication: April 27, 1989.