UNIVERSITA' DEGLI STUDI DI MILANO

Facoltà di Medicina e Chirurgia

Dipartimento di Medicina del Lavoro "Clinica del Lavoro L. Devoto"

Sezione di Statistica Medica e Biometria "G.A. Maccacaro"

CORSO DI DOTTORATO DI RICERCA IN STATISTICA BIOMEDICA (D4R)

SETTORE SCIENTIFICO DISCIPLINARE MED/01 – XXIII CICLO

TESI DI DOTTORATO DI RICERCA

"Exploratory and Confirmatory Factor Analysis to identify and validate dietary patterns: an application to a case-control study of gastric cancer"

Dottoranda: Paola Bertuccio

Relatori: Prof. Adriano Decarli

Prof. Carlo La Vecchia

Coordinatore del Dottorato: Prof. Silvano Milani

A.A. 2009/2010

Alla mia famiglia di origine
e alla mia nuova famiglia che sta per nascere
che hanno condiviso
e condividono giorno per giorno
il mio percorso di vita lavorativo e quotidiano
e che per questo completano la mia vita

INDEX

1. INTRODUCTION

In nutritional epidemiologic research, the use of dietary pattern methods has increased substantially over the past several years [1-4]. Patterning methods consider multiple foods, beverages, and/or nutrients and therefore create dietary variables that more realistically resemble actual eating behavior. As a result, these methods have grown in popularity as a valuable complement to the traditional approach. In addition, research using the traditional approach, which often studies single nutrients or foods, is limited because of collinearity among nutrients and the inability to detect small effects from single nutrients. Studies using empirically derived patterns are based only on dietary intake data and therefore are better poised to provide an understanding of actual diets.

Use of explorative statistical methods is one way to examine dietary patterns in populations. Of these, exploratory factor analysis (EFA) is a data aggregation procedure used to reduce dietary data into meaningful food or nutrient patterns based on inter-correlations between dietary items. The factors are then named, usually according to those foods or nutrients that most heavily contribute to the pattern, and the patterns can then be used as the primary exposure variables in dietary studies.

The EFA is *a posteriori* approach since is based on empirical data and not on *a priori* hypothesis. It does not require a theoretical basis and uses only the data to derive food/or nutrient patterns empirically.

Several studies have used factor analysis to identify dietary patterns in epidemiological studies [1-4] as an alternative to the analyses on single foods or single nutrients, but few have validated the factors in a larger population. Indeed, there are limited data on the reproducibility of this method [5-10]. Studying the reproducibility of patterns derived by the use of factor analysis is an important step in establishing the validity of this method.

Confirmatory factor analysis (CFA), by contrast, based on *a priori* hypothesis, can be used to assess the reproducibility and validity of dietary patterns identified by an EFA. It may be guided both by results from an EFA and by knowledge of nutritional behavior.

This *a priori* approach is intuitively appealing because it may be based in theory and also reduces some of the subjectivity involved in exploratory procedures. However, few studies have used confirmatory factor analysis in nutritional epidemiology. More research is required to understand how this method is used in nutritional epidemiology, as well as how solutions derived from the use of confirmatory factor analysis differ from those using exploratory factor analysis.

The purpose of my PhD thesis is to further knowledge of factor analysis methods in nutritional epidemiologic research. In particular, I studied the application of the CFA to validate nutrient-dietary patterns derived from EFA. In the chapter 2, I will describe the steps whereby identify different nutrient dietary patterns throughout EFA, and the steps to validate the explored factor solutions by testing

CFA models, in which only the observed variables decided a priori, are included (e.g. by the magnitude of their explored factor loading in previous EFA). In the chapter 3, I will present an application to a case-control study of gastric cancer conducted in northern Italy, by comparing the results from different explored factor solutions, characterized by different numbers of specified factors.

2. STATISTICAL METHODS

2.1 Exploratory factor analysis

This paragraph aims to describe the steps of exploratory factor analysis to identify different nutrient dietary patterns, as *a posteriori* approach.

2.1.1 Factorability of the original matrix

The correlation matrix R of the original data is used to assess its factorability. Variables should not be:

1) too highly correlated ($r \geq 0.80$); this reflects problems of multicollinearity, so that one or more of these variables would be dropped from the analysis;
2) not sufficiently correlated ($r < 0.30$) with one another; this means these variables will not share much of the common variance, thus potentially leading to solution with as many factors as variables.

Then, matrix factorability is evaluated through statistical procedures. Measures of sampling adequacy that compare the simple and partial correlation coefficients may be defined either overall or for single variables. The overall measure, called Kaiser-Meyer-Olkin statistic (*KMO*), is defined as follows [11]:

$$KMO = \frac{\sum_{i \neq j} \sum r_{ij}^2}{\sum_{i \neq j} \sum r_{ij}^2 + \sum_{i \neq j} \sum a_{ij}^2}$$

where $\sum\sum$ are the sum over all variables in the matrix when variable $i \neq$ variable j, r_{ij} is the Pearson correlation coefficient between i and j, and a_{ij} the partial correlation coefficient between i and j. Individual measures of sampling adequacy are computed using only the simple and partial correlation coefficients involving the specific variable under consideration. The overall and individual measures range between 0 and 1. Smaller values indicate that the squared correlation coefficient is small relative to the squared correlation coefficient and therefore a factor analysis may be imprudent. If the sum of the squared partial correlation coefficients is small compared with the sum of the squared correlation coefficients, the measures approach 1.

Bartlett's test of sphericity tests the null hypothesis that the correlation matrix is an identity matrix. It is a chi-square test [11], whose statistic is defined as follows:

$$\chi^2 = -\left[(N-1)-\left(\frac{2k+5}{6}\right)\right]\log|R|$$

where χ^2 is the calculated chi-square value for Bartlett's test, N is sample size, k is the number of variables in the matrix and $|\boldsymbol{R}|$ the determinant of the correlation matrix. The degrees of freedom for this chi-square are $k(k-1)/2$. Larger values of the test suggest that the null hypothesis should be rejected.

Since Bartlett's test is influenced by the sample size, N, for larger samples this test tends to indicate that the correlation matrix is not an identity matrix. For this reason, it should be used only as a minimum standard for assessing the quality of the correlation matrix.

2.1.2 Identification of factors through factor analysis

This approach assumes that the variables included in the analysis can be perfectly calculated by the extracted components or factors. Because each standardized variable has a mean of 0 and variance of 1, the initial estimate of communality (i.e., the explained variance) for each variable is 1.00. This is what will be placed initially on the diagonal of the correlation matrix. The first principal component is a linear combination of the original variables, such that it explains the maximum amount of the variance among the variables. After the first extraction, a residual correlation matrix is created. This matrix contains the variances not explained by the first factor on the diagonal and the partial correlations of the variables with each other after extracting the first factor on the off-diagonal. The second one is extracted from this residual matrix, so it will be uncorrelated to the first one. This process of extracting principal components is repeated on subsequent residual matrices, until the elements in the residual variance-covariance matrix are reduced to random error.

2.1.3 Choosing the number of factors to retain

A crucial aspect of factor analysis is the choice of the number of factors to retain. The choice was based on three main criteria. The first one is to retain those factors with eigenvalues greater than 1.00. The eigenvalues is defined as λ of **R**. An eigenvalue for **R** is a value for which the following polynomial equation, holds:

$p(\lambda) = Det(\mathbf{R}-\lambda\mathbf{I}) = 0$

where *Det* is the determinant of the matrix ($\mathbf{R}-\lambda\mathbf{I}$), **R** is the correlation matrix, λ is the eigenvalue of **R**, and **I** is the identity matrix of the same size as the correlation matrix with 1s on the diagonal and 0s on the off diagonal. The polynomial is called the *characteristic polynomial* of the matrix **R**, and the eigenvalues of **R** are the nontrivial solutions to the *characteristic equation*, $p(\lambda)=0$. With a matrix **R** n x n, there is at most n solution for λ [11].

The second criterion is to add successive factors until the cumulative percentage of variance explained by the retained factors is satisfactory. To terminate the factor extraction process, we considered 75-80% to be a valid threshold for the cumulative variance extracted. The third one, suggested by Cattell [12], is to plot, by the option SCREE in SAS, the extracted factors against their eigenvalues in descending order of magnitude to identify distinct breaks in the slope of the plot, called "scree plot". To determine where the break occurs, a straight line should be drawn with a ruler through the lower values of the plotted eigenvalues. That point where the factors curve above the straight line drawn through the smaller eigenvalues identifies the optimal number of factors to retain.

Finally, to determine the number of factors to retain, a researcher should not be based only on statistical criteria, but also on subjective motivations. In fact, the other criterion to take in account is factor interpretability. In nutritional epidemiology, the identified factors represent potentially uncorrelated dietary habits that, considered altogether, summarize the overall dietary profile of a given population.

2.1.4 Estimating factor scores

Factor scores are estimated for each subject and factor. They indicate the degree to which each subject's diet conforms to one of the identified factors [13], and can be calculated using the weighted least square method, where variables that have lower loadings on the factor are given less weight than those with higher loadings in the calculation of factor scores.

2.1.5 Rotating the identified factors

To improve the interpretation of the generated factors, suggestions have been made to rotate them. If a rotation is not performed, the first unrotated factor is most often a general factor on which most variables load highly in absolute value. The rotation consists in turning the reference axes of the factors about their origin to achieve a simple structure where variables should load highly (in absolute value) on one factor only, and each factor should have high absolute loadings only on some of the variables.

There are two methods of rotation: *orthogonal* and *oblique*. In the first one, pairs of axes are kept at right angles (90°) to one another during rotation, so that they are still uncorrelated after rotation. In the second one, each axis may be rotated independently, so that they are not necessarily perpendicular after rotation.

We preferred forms of *orthogonal* rotation. An important statistical assumption of this method of rotation is that the rotated factors remain statistically uncorrelated. This is an advantage associated with representing a complex set of interrelationships among several correlated variables in terms of a few

uncorrelated indices. This is a crucial aspect in nutritional epidemiology, where one may deal with severe multicollinearity problems. Another property of *orthogonal* rotations is that the amount of the total variance accounted for by the factors under consideration is unaffected by the rotation itself [14].

2.1.6 Naming the identified factors

To name the identified factors, it is suggested to consider only those ones having factor loadings greater or equal to |0.63| on a given factor. The contribution that a factor gives to a nutrient's sample variance is equal to the square of its loading on that factor, so if we choose a |0.63| cut-off, we expect a minimum contribution of the factor on the nutrient's variance of approximately 0.40 [13].

2.2 Confirmatory factor analysis

In contrast to exploratory factor analysis (EFA) there is confirmatory factor analysis (CFA).

CFA is a type of structural equation modeling (SEM) [15] that deals specifically with measurement models, that is, the relationships between *measured variables* and *latent variables*. A measured variable, also called *observed variable* or *indicator*, is a variable that have been directly observed in the study, whereas a latent variable, also called *latent factor* is a hypothetical construct that is not directly measured or observed in the study. In this context, the factors derived from an EFA are examples of latent variables.

A fundamental feature of CFA is its hypothesis-driven nature. Unlike EFA, the researcher must pre-specify all aspects of the CFA model. Thus, the researcher must have a firm *a priori* sense, based on past evidence and theory, of the number of factors that exist in the data, of which indicators are related to which factors, and so forth.

The CFA is a statistical technique that allows the researcher to test and verify a particular model or factor structure that they believe underlies the variables measured in the study. The researcher can use knowledge of the theory, empirical research, or both, to assume that a relationship between observed variables and their underlying latent construct exists, and then test the hypothesis statistically.

In the present thesis, CFA was applied to test the validity of the underlying dimensions of a construct identified through previous EFA [11, 16-19].

In addition to its greater emphasis on theory and hypothesis testing, the CFA framework provides many other analytic possibilities that are not available in EFA. These possibilities include the evaluation of method effects and the examination of the stability or invariance of the factor model over time or informants.

CFA has become one of the most commonly used statistical procedures in applied research.

2.2.1 The structure of Confirmatory Factor Analysis

The CFA models are represented by "flow" diagrams. By tradition, causal models have three kinds of elements: text inside geometric shapes, lines with arrows

pointing to or away from these shapes, and coefficients on the lines. Measured variables are represented by squares or rectangles; latent variables are represented as circles or ovals; regression paths or, in the case of factor analysis, pattern coefficients, are represented by one-headed arrows. These lines are drawn *from* latent factors *to* measured variables to reflect the hypothesis that the latent variables are in fact an underlying influence on the manifestation of the factors in the form of scores on the measured variables. Correlations (and covariances) are represented as two-headed arrows drawn to connect either measured or latent variables in pairs for which correlations or covariances are freed to be nonzero and are estimated in the analysis. In the following paragraphs, details on the CFA model will be described.

Parameters of a CFA Model

All CFA models contain *factor loadings*, *unique variances*, and *factor variances* [17]. *Factor loadings* are the regression slopes for predicting the indicators from the latent factor. *Unique variance* for each measured variable is the variance in the indicator that is not accounted for by the latent factors and is typically presumed to be measurement error. *Factor variance*, in an unstandardized solution, expresses the sample variability or dispersion of the factor.

A CFA may include error covariances (referred to as "correlated residuals," or "correlated errors"), which suggest that two indicators covary for reasons other than the shared influence of the latent factor.

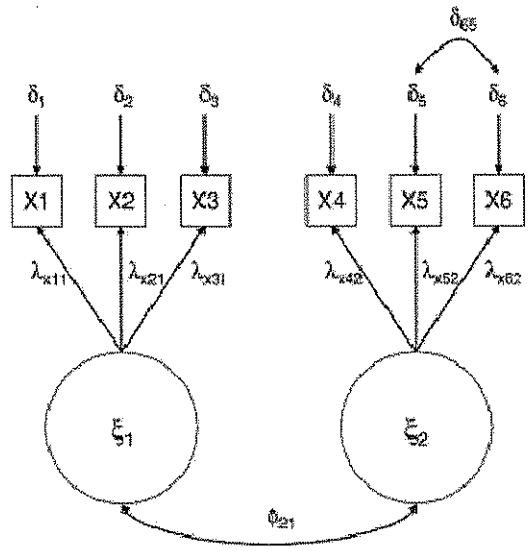
When the CFA solution consists of two or more factors, a factor covariance (a "factor correlation" being the completely standardized counterpart) is usually

specified to estimate the relationship between the latent dimensions. However, one may fix factor covariances to zero, like an orthogonal EFA solution.

CFA is often confined to the analysis of variance–covariance structures. In this instance, the above-mentioned parameters (factor loadings, error variances and covariances, factor variances and covariances) are estimated to reproduce the input variance–covariance matrix.

Latent variables in CFA may be either *exogenous* or *endogenous*. An *exogenous* variable is a variable that is not caused by other variables in the solution. Conversely, an *endogenous* variable is caused by one or more variables in the model. Thus, exogenous variables can be viewed as synonymous to X, independent, or predictor (causal) variables. Similarly, endogenous variables are equivalent to Y, dependent, or criterion (outcome) variables. However, in the case of structural models, an endogenous variable may be the cause of another endogenous variable.

As an example, a two-factor CFA model is represented in the following "flow" diagram:



Name	Parameter	Matrix	Type	Description
Lambda-X	λ_x	Λ_x	Regression	Factor loadings
Theta delta	δ	Θ_δ	Variance–covariance	Error variances and covariances
Phi	ϕ	Φ	Variance–covariance	Factor variances and covariances
Xi (Ksi)	ξ		Vector	Names of exogenous variables

Figure 1. A two-factor CFA model.

Latent factors are a weighted composite of all the measured variables. Each factor is represented by a different set of weights, called also factor loadings, symbolized by lambdas (λ) with x subscripts in the case of exogenous latent variables. The unidirectional arrows ($\rightarrow$) from the factors (e.g., ξ_1) to the indicators (e.g., X1) depict direct effects (regressions) of the latent dimensions onto the observed measures; the specific regression coefficients are the lambdas (λ). Theta-delta (Θ_δ) represent matrices of indicator error variances and covariances; for notational ease, the symbol δ is often used in place of θ_δ in reference to elements of Θ_δ. Θ_δ is symmetric variance–covariance matrices consisting of error variances on the diagonal, and error covariances, if any, in the off-diagonal.

Factor variances and covariances are notated by phi (φ). Curved, bidirectional arrows are used to symbolize covariances (correlations). In the Figure 1, curved

arrows indicate the covariance between the factors (φ_{21}) and the error covariance of the X5 and X6 indicators (δ_{65}).

The parameters in Figures 1 also possess numerical subscripts to indicate the specific elements of the relevant matrices. For example, λ_{x11} indicates that the X1 measure loads on the first exogenous factor (ξ_1), and λ_{x21} indicates that X2 also loads on $\xi 1$, and so on. This numeric notation assumes that the indicators were ordered X1, X2, X3, X4, X5, and X6 in the input variance–covariance matrix. If the input matrix was arranged in this way, the lambda X matrix (Λ_x) would be as follows:

	ξ_1	ξ_2
X1	λ_{x11}	0
X2	λ_{x21}	0
X3	λ_{x31}	0
X4	0	λ_{x42}
X5	0	λ_{x52}
X6	0	λ_{x62}

where the first numerical subscript refers to the row of Λ_x (i.e., the positional order of the X indicator) and the second numerical subscript refers to the column of Λ_x (i.e., the positional order of the exogenous factors, ξ); e.g., λ_{x52} conveys that the fifth indicator in the input matrix (X5) loads on the second latent X factor (ξ_2). Thus, Λ_x is full matrices whose dimensions are defined by *p* rows (number of indicators) and *m* columns (number of factors). The zero elements of Λ_x (e.g., λ_{x12}, λ_{x41}) indicate the absence of cross-loadings (e.g., the relationship between X1 and ξ_2 is fixed to zero). This is also showed in Figures 1 by the absence of directional arrows between certain indicators and factors (e.g., no arrow connecting ξ_2 to X1).

A similar system is used for variances and covariances among factors (φ) and indicator errors (δ). However, because these aspects of the CFA solution reflect variances and covariances, they are represented by $m \times m$ symmetric matrices with variances on the diagonal and covariances in the off-diagonal.

For example, the phi matrix (Φ) in Figure 1 would look as follows:

	ξ_1	ξ_2
ξ_1	ϕ_{11}	
ξ_2	ϕ_{21}	ϕ_{22}

where φ_{11} and φ_{22} are the factor variances, and φ_{21} is the factor covariance.

Similarly, the theta-delta matrix (Θ_δ) is the following $p \times p$ symmetric matrix:

	X1	X2	X3	X4	X5	X6
X1	δ_{11}					
X2	0	δ_{22}				
X3	0	0	δ_{33}			
X4	0	0	0	δ_{44}		
X5	0	0	0	0	δ_{55}	
X6	0	0	0	0	δ_{65}	δ_{66}

where δ_{11} through δ_{66} are the indicator errors and δ_{65} is the covariance of the measurement errors of indicators X5 and X6. The zero elements of Θ_δ (e.g., δ_{21}) indicate the absence of error covariances (i.e., these relationships are fixed to zero).

Unlike in EFA, in which all parameters implicit in a factor model must be estimated, in CFA, the researcher can "constrain" or "fix" certain parameters to mathematically "permissible" values (e.g., a variance may be constrained to equal any positive number; a correlation *r* may be constrained to equal -1, +1, or any number in between), while at the same time "freeing" the use of the input data to derive estimates of other model parameters (e.g., factor pattern coefficients, factor

variances). Moreover, a researcher must declare as input into the analysis one or more specific models, each containing some "fixed" and some "freed" parameters.

Fundamental equations of a CFA model

CFA aims to reproduce the sample variance–covariance matrix by the parameter estimates of the measurement solution (e.g., factor loadings, factor covariances, etc.). Considering the Figure 1, the first set of measures (X1, X2, X3) are indicators of one latent factor (ξ_1), whereas the second set of measures (X4, X5, X6) are indicators of another latent factor (ξ_2). It is said, for example, that indicators X4, X5, and X6 are *congeneric* [17] because they share a common factor (ξ_2). An indicator would not be considered congeneric if it loaded on more than one factor. In the case of congeneric factor loadings, the variance of an indicator is reproduced by multiplying its squared factor loading by the variance of the factor, and then summing this product with the indicator's error variance. The predicted covariance of two indicators that load on the same factor is computed as the product of their factor loadings times the variance of the factor. The model-implied covariance of two indicators that load on separate factors is estimated as the product of their factor loadings times the factor covariance. For example, based on the parameter estimates in the solution presented in Figure 1, the variance of X2 would be reproduced by the following equation:

$$\text{VAR}(X2) = \sigma_{22} = \lambda_{x21}^2 \varphi_{11} + \delta_2$$

In the case of completely standardized solutions, one can reproduce the variance of an indicator by simply squaring its factor loading and adding its error, because

the factor variance will always equal 1.00. The factor variance must be included in this calculation when dealing with unstandardized solutions.

The squared factor loading represents the proportion of variance in the indicator that is explained by the latent factor (often referred to as a *communality*). For example, the communality of X2 is:

$$\xi_{22}^2 = \lambda_{x21}^2$$

Similarly, in the completely standardized solution, the errors represent the proportion of variance in the indicators that is not explained by the latent factor. These errors (residual variances) can be readily calculated as 1 minus the squared factor loading. Using the X2 indicator, the computation would be:

$$\delta_2 = 1 - \lambda_{x21}^2$$

The predicted covariance (correlation) between X2 and X3 would be estimated as follows:

$$\text{COV}(X2, X3) = \sigma_{3,2} = \lambda_{x21}\, \varphi_{11}\, \lambda_{x31}$$

In the case of completely standardized solutions the factor variance will always equal 1.00, so the predicted correlation between two congeneric indicators can be calculated by the product of their factor loadings.

In Figure 1, the covariation between the indicators is not accounted for fully by the latent factor (ξ_2); that is, X5 and X6 share additional variance due to influences other than the latent construct. Thus, the equation to calculate the predicted correlation of X5 and X6 includes the correlated error:

$$COV(X5, X6) = \sigma_{6,5} = (\lambda_{x52}\varphi_{22}\lambda_{x62}) + \delta_{65}$$

2.2.2 CFA Model Identification

In order to estimate the parameters in CFA, the measurement model must be *identified.* A model is identified if, on the basis of known information (i.e., the variances and covariances in the sample input matrix), it is possible to obtain a unique set of parameter estimates for each parameter in the model whose values are unknown (e.g. factor loadings, factor correlations, etc.). Model identification pertains in part to the difference between the number of freely estimated model parameters and the number of pieces of information in the input variance-covariance matrix.

The parameters in a CFA model are the pattern or structure coefficients relating the independent to the dependent variables, correlation coefficients relating the independent variables to each other, and the variance of the independent variables. These parameters can be estimated only if the number of freely estimated parameters does not exceed the number of pieces of information in the input variance–covariance matrix.

1) A model is *under-identified* when the number of unknown (freely estimated) parameters exceeds the number of pieces of known information (i.e., elements of the input variance–covariance matrix). An under-identified model cannot be solved because there are an infinite number of parameter estimates that result in perfect model fit.

2) A model is *just-identified* when the number of known parameters would equal the number of those unknowns. To obtain a just-identified model, for instance, the researcher could add the restriction of constraining the factor loadings to equality. Because the number of knowns equals the number of unknown parameters, in just-identified models, there exists a single set of parameter estimates that perfectly fit the data.
3) A model is *over-identified* when the number of known parameters (i.e., number of variances and covariances in the input matrix) exceeds the number of freely estimated model parameters.

The difference in the number of knowns and the number of unknowns (i.e., freely estimated parameters) constitutes the model's degrees of freedom (*df*). Over-identified solutions have positive *df*, just-identified models have 0 *df* (because the number of knowns equals the number of unknowns), and under-identified models have negative *df* (cannot be solved or fit to the data).

If we subtract the number of unknown parameters from the number of known or nonredundant elements, we obtain the degrees of freedom for the analysis:

$$df = \text{n. of nonredundant elements} - \text{n. of unknown parameters}$$

The number of known parameters is equal to the number of unique or nonredundant entries in a matrix that represents the covariances or correlations of the indicator variables:

$$\text{N. of nonredundant elements} = p\,(p+1) / 2$$

where p is the number of the measured variables in the study

Model identification also requires that the measurement scale (i.e., the variance or the standard deviation) of each latent variable is specified or constrained. This is because latent variables, by definition, have no intrinsic scaling, and so there are infinitely many plausible scales for these variables, each suggesting a corresponding plausible set of estimates for the other model parameters.

In effect, if the researcher wants to estimate scores of a latent variable, he must first declare a metric for the estimate. It is usually irrelevant what this metric is, but some metric must be selected. There are two common ways to identify CFA models.

First, any factor pattern coefficient on each factor can be fixed to any number. The number "1" is a common choice. But it could be used any number. In effect, this means that the researcher wants to scale the scores on the latent variable as some multiple of the selected measured variable.

The decision of which measured variable's pattern coefficient on a factor has to be selected to fix to some number (usually "1") makes no real difference. However, some researchers stylistically prefer for model identification purposes to pick the measured variable thought to most reflect the factor or to have scores that are the most reliable from a measurement point of view.

Second, in case of a first-order factor model, it is possible constrain the factor variances to be any mathematically plausible number (i.e., positive). When this strategy is selected, it is useful to use the same number to constrain all factor variances, and usually the number "1" is used for these purposes. One advantage of doing so is that the covariances of the factors become factor correlation

coefficients, because when computing the factor correlations as r=COV/(SDxSD) all pairs of computations will then have denominators of one.

The selection of scaling for the latent variables will not affect the model fit statistics. The scaling decisions are basically arbitrary and can reasonably be made as a function of the researcher's stylistic preferences.

2.2.3 Estimation of CFA Model Parameters

The objective of CFA is to obtain estimates for each parameter of the measurement model (i.e., factor loadings, factor variances and covariances, indicator error variances and possibly error covariances) that produce a predicted variance–covariance matrix (symbolized as Σ) that resembles the sample variance–covariance matrix (symbolized as S) as closely as possible.

To minimize the difference between Σ and S, the fitting function most widely used in applied CFA research (and SEM, in general) is *maximum likelihood* (ML). the fitting function that is minimized in ML is:

$$F_{ML} = \ln|S| - \ln|\Sigma| + \text{trace}[(S)(\Sigma^{-1})] - p$$

where $|S|$ is the determinant of the input variance–covariance matrix, $|\Sigma|$ is the determinant of the predicted variance–covariance matrix, p is the order of the input matrix (i.e., the number of input indicators).

The determinant and trace summarize important information about matrices such as S and Σ. The determinant is a single number (i.e., a scalar) that reflects a generalized measure of variance for the entire set of variables contained in the matrix. The trace of a matrix is the sum of values on the diagonal (e.g., in a

variance–covariance matrix, the trace is the sum of variances). The objective of ML is to minimize the differences between these matrix summaries (i.e., the determinant and trace) for S and Σ.

The underlying principle of ML estimation in CFA is to find the parameter values that make the observed data most likely (or conversely, maximize the likelihood of the parameters given the data). Finding the parameter estimates for an over-identified CFA model is an iterative procedure. That is, the computer program begins with an initial set of parameter estimates (referred to as starting values or initial estimates, which can be automatically generated by the software or specified by the user) and repeatedly refines these estimates in an effort to reduce the value of F_{ML} (i.e., minimize the difference between Σ and S). Each refinement of the parameter estimates to minimize F_{ML} is an iteration. The program conducts internal checks to evaluate its progress in obtaining parameter estimates that best reproduce S (i.e., that result in the lowest F_{ML} value). Convergence of the model is reached when the program arrives at a set of parameter estimates that cannot be improved upon to further reduce the difference between Σ and S.

Occasionally, a latent variable solution will fail to converge. Convergence is often related to the quality and complexity of the specified model (e.g., the number of restrictions imposed on the solution) and the adequacy of the starting values. In the case of complex models, convergence may not be reached because the program has stopped at the maximum number of iterations, which is set by either the program's default or a number specified by the user. This problem may be rectified by simply increasing the maximum number of iterations or possibly using the preliminary parameter estimates as starting values. However, a program

may also cease before the maximum number of iterations has been reached because its internal checks indicate that progress is not being made in obtaining a solution that minimizes F_{ML}.

2.2.4 Descriptive Goodness-of-fit Indices

CFA relies on several statistical tests to determine the adequacy of model fit to the data.

The classic goodness-of-fit index is the Chi-square (χ^2). The χ^2 test indicates the amount of difference between expected and observed covariance matrices. The null hypothesis is that the model fits the data. If the model provides a good fit, the χ^2 value will be relatively small (close to 0), indicating little difference between the expected and observed covariance matrices. In addition, the probability level will be relatively large: above 0.05 and preferably closer to 1.00 when χ^2 is close to 0 [20].

Under typical ML model estimation, χ^2 is calculated as:

$$\chi^2 = F_{ML}(N-1)$$

Because this model is associated with 1 *df*, the critical χ^2 value ($\alpha = 0.05$) is 3.84 (i.e., $\chi^2 = z^2 = 1.962 = 3.8416$). Thus, a statistically significant χ^2 supports the alternate hypothesis that $S \neq \Sigma$, meaning that the model estimates do not sufficiently reproduce the sample variances and covariances (i.e., the model does not fit the data well).

Although χ^2 is steeped in the traditions of ML and SEM (e.g., it was the first fit index to be developed), it is rarely used in applied research as a sole index of

model fit. Indeed, important criticisms of χ^2 include the following: (1) in many instances (e.g., small N, non-normal data) its underlying distribution is not χ^2 distributed (compromising the statistical significance tests of the model χ^2); (2) it is inflated by sample size, and thus large N solutions are routinely rejected on the basis of χ^2 even when differences between S and Σ are negligible; and (3) it is based on the very stringent hypothesis that S = Σ. Many alternative fit indices are based on less stringent standards such as "reasonable" fit and fit relative to an independence model. Nevertheless, χ^2 is used for other purposes, such as nested model comparisons and the calculation of other fit indices. While χ^2 is routinely reported in CFA research, other fit indices are usually relied on more heavily in the evaluation of model fit.

Although a host of fit indices are available, only a handful is described here. These fit indices were selected on the basis of their popularity in the applied literature.

Fit indices can be broadly characterized as falling under three categories: absolute fit, fit adjusting for model parsimony, and comparative or incremental fit.

Absolute Fit

Absolute fit indices assess model fit at an absolute level; in various ways, they evaluate the reasonability of the hypothesis that $S = \Sigma$ without taking into account other aspects such as fit in relation to more restricted solutions. Thus, χ^2 is an example of an absolute fit index. Another index that falls in this category is the standardized root mean square residual (SRMR). Conceptually, the SRMR can be viewed as the average discrepancy between the *correlations* observed in the input

matrix and the correlations predicted by the model (though in actuality, the SRMR is a positive square root average). Accordingly, it is derived from a residual correlation matrix. A similarly named index, the root mean square residual (RMR), reflects the average discrepancy between observed and predicted *covariances*. However, the RMR can be difficult to interpret because its value is affected by the metric of the input variables; thus, the SRMR is generally preferred. In most instances (e.g., models involving a single input matrix), the SRMR can be calculated by (1) summing the squared elements of the residual correlation matrix and dividing this sum by the number of nonredundant elements in this matrix (on and below the diagonal), that is, $b = p(p + 1) / 2$, and (2) taking the square root (SQRT) of this result.

$$\text{SRMR} = \text{SQRT}[(sum(\text{element of the residual correlation matrix})^2)/b]$$

The SRMR can take a range of values between 0.0 and 1.0, with 0.0 indicating a perfect fit (i.e., the smaller the SRMR, the better the model fit).

Parsimony Correction

A widely used and recommended index from this category is the root mean square error of approximation (RMSEA). The RMSEA is a population-based index that relies on the *non-central χ^2 distribution*, which is the distribution of the fitting function (e.g., F_{ML}) when the fit of the model is not perfect. The non-central χ^2 distribution includes a *non-centrality parameter* (NCP), which expresses the degree of model misspecification.

The NCP is estimated as $\chi^2 - df$ (if the result is a negative number, NCP = 0). When the fit of a model is perfect, NCP = 0 and a central χ^2 distribution holds.

When the fit of the model is not perfect, the NCP is greater than 0 and shifts the expected value of the distribution to the right of that of the corresponding central χ^2. The RMSEA is an "error of approximation" index because it assesses the extent to which a model fits *reasonably* well in the population (as opposed to testing whether the model holds exactly in the population). To foster the conceptual basis of the calculation of RMSEA, the NCP is rescaled to the quantity $d = (\chi^2 - df) / (N - 1)$.

The RMSEA is then computed:

$$\text{RMSEA} = \text{SQRT}[d / df]$$

where *df* is the model *df*. The RMSEA compensates for the effect of model complexity by conveying discrepancy in fit (*d*) per each *df* in the model. Thus, it is sensitive to the number of model parameters; being a population-based index, the RMSEA is relatively insensitive to sample size.

Although its upper range is unbounded, it is rare to see the RMSEA exceed 1.0. As with the SRMR, RMSEA values of 0 indicate perfect fit (and values very close to 0 suggest good model fit).

The non-central χ^2 distribution can be used to obtain confidence intervals for RMSEA (a 90% interval is typically used). The confidence interval indicates the precision of the RMSEA point estimate. Methodologists recommend including this confidence interval when reporting the RMSEA. However, researchers should be aware that the width of this interval is affected by sample size and the number of freely estimated parameters in the model (e.g., unless *N* is very large, complex models are usually associated with wide RMSEA confidence intervals).

Comparative Fit

Comparative fit indices evaluate the fit of a user-specified solution in relation to a more restricted, nested baseline model [21]. Typically, this baseline model is a "null" or "independence" model in which the covariances among all input indicators are fixed to zero, although no such constraints are placed on the indicator variances. As you might expect, given the relatively liberal criterion of evaluating model fit against a solution positing no relationships among the variables, comparative fit indices often look more favorable (i.e., more suggestive of acceptable model fit) than indices from the preceding categories. Nevertheless, some indices from this category have been found to be among the best behaved of the host of indices that have been introduced in the literature.

One of these indices, the Bentler's comparative fit index (CFI), is computed as follows:

$$CFI = 1 - \max[(\chi_T^2 - df_T), 0] / \max[(\chi_T^2 - df_T), (\chi_B^2 - df_B), 0]$$

where χ_T^2 is the χ^2 value of the target model (i.e., the model under evaluation), df_T is the *df* of the target model, χ_B^2 is the χ^2 value of the baseline model (i.e., the "null" model), and df_B is the *df* of the baseline model; max indicates to use the largest value – for example, for the numerator, use ($\chi_T^2 - df_T$) or 0, whichever is larger. The χ_B^2 and df_B of the null model are included as default output in most software programs. If the user wishes to obtain these values in programs that do provide this information, χ_B^2 and df_B can be calculated by fixing all relationships to 0 (but freely estimating the indicator variances). The CFI has a range of possible values of 0.0 to 1.0, with values closer to 1.0 implying good model fit. Like the RMSEA, the CFI is based on the non-centrality parameter (i.e., λ

$= \chi_T^2 - df_T$), meaning that it uses information from expected values of χ_T^2 or χ_B^2 (or both, in the case of the CFI) under the non-central χ^2 distribution associated with $S \neq \Sigma$ (e.g., central χ^2 is a special case of the non-central χ^2 distribution when λ=0).

Bentler and Bonnett's normed-fit index (NFI) has also been proposed. Values on this index may range from 0 to 1, with values over 0.9 indicative of an acceptable fit of the model to the data. This index may be viewed as the percentage of observed-measure covariation explained by a given measurement or structural model. Although, the NFI has the advantage of being easily interpreted, it has the disadvantage of sometimes underestimating goodness of fit in small samples.

A variation on the NFI is the non-normed fit index (NNFI). The NNFI has been shown to better reflect model fit at all sample sizes. NNFI values over 0.90 are also viewed as desirable, although, unike the NFI, the NNFI may assume values below 0 and above 1.

Finally, Bentler's CFI is similar to the NNFI in that it provides an accurate assessment of fit regardless of sample size. In addition, the CFI tends to be more precise that the NNFI in describing comparative model fit. Values of the CFI will always lie between 0 and 1, with values over 0.9 indicating a relatively good fit.

2.2.5 Modification Indices

Often a CFA model will need to be revised. The most common reason for respecification is to improve the fit of the model. In this case, the results of an initial CFA indicate that one or more of the three major criteria used to evaluate

the acceptability of the model are not satisfied: that is, the model (1) does not fit well on the whole, (2) does not reproduce some indicator relationships well, or (3) does not produce uniformly interpretable parameter estimates. Based on fit diagnostic information (e.g., *modification indices*) and substantive justification, the model is revised and fit to the data again in the hope of improving its goodness of fit.

Modification indices can be computed for each fixed parameter (e.g., parameters that are fixed to zero such as indicator cross-loadings and error covariances) and constrained parameter in the model. The modification index reflects an approximation of how much the overall model χ^2 would decrease if the fixed or constrained parameter was freely estimated. Indeed, if the parameter is freely estimated in a subsequent analysis, the actual decrease in model χ^2 may be somewhat smaller or larger than the value of the modification index. In other words, the modification index is roughly equivalent to the difference in the overall χ^2 between two models, where in one model the parameter is fixed or constrained and in the other model the parameter is freely estimated.

In general, a good-fitting model should also produce modification indices that are small in magnitude. Because the modification index can be conceptualized as a χ^2 statistic with 1 *df*, indices of 3.84 or greater (which reflects the critical value of χ^2 at $p<0.05$, 1 *df*) suggest that the overall fit of the model could be significantly improved ($p<0.05$) if the fixed or constrained parameter was freely estimated.

Like overall model χ^2 and standardized residuals, modification indices are sensitive to sample size. For instance, when N is very large, a large modification

index may suggest the need to add a given parameter despite the fact that the magnitude of the parameter in question, if freely estimated, is rather trivial.

It is recommended that model modification begin by freely estimating the fixed or constrained parameter with the largest modification index if this parameter can be interpreted substantively. If there does not exist a substantive basis for relaxing the parameter with the largest modification index, consider the parameter associated with the second largest modification index, and so on.

Revisions of a model should always focus exclusively on parameters justified by prior evidence or theory. Re-specified models should be interpreted with caution, especially in instances where substantial changes have been made to the initial model, modified solutions should be replicated in independent samples.

Models can be revised by eliminating statistically non-significant parameters. The presence of unnecessary parameters may be reflected by large, negative standardized residuals that indicate that the model is overestimating the observed relationship between a pair of indicator variables.

Two modification indices are to consider in this stage of the CFA:

1) **Wald test**: identifies parameters that should possibly be dropped from the model. It provides an estimate of how much the overall model χ^2 would increase if a freely estimated parameter were fixed to zero. A non-significant Wald test value (e.g., < 3.84 in the case of a single parameter) would indicate that removing the freely estimated parameter (e.g., fixing it to zero) would not result in a significant decrease in model fit.
2) **Lagrange multiplier test**: identifies parameters that should possibly be added. It estimates the reduction in model χ^2 that would result from freeing

a fixed parameter and allowing it to be estimated. In other words, for a CFA the Lagrange multiplier estimates the degree to which χ^2 would improve if a new factor loading or covariance were added to the model. Two matrices of Lagrange multiplier tests are provided. First, the *phi matrix* that contains indices for every possible combination of latent factors and residual terms. The second matrix is the *gamma matrix* that indicates whether it should be added a new path from some latent factor to some indicator variable.

3. APPLICATION TO A CASE-CONTROL STUDY OF GASTRIC CANCER

3.1 Design and participants

Data were derived from a case-control study of stomach cancer conducted between 1997 and 2007 in the Greater Milan area, Italy [22]. Briefly, cases were 230 patients (143 men and 87 women; median age 63 years, range 22-80 years), admitted to major teaching and general hospitals in the study area with incident, histologically confirmed stomach cancer (ICD IX, 151.0-151.9), diagnosed no longer than 1 year before the interview, and with no previous diagnosis of cancer. The control group included 547 patients (286 men and 261 women; median age 63 years, range 22-80 years) frequency matched to cases by age and sex (with a ratio of 2:1 for men, and 3:1 for women), admitted to the same hospitals as cases for a wide spectrum of acute, non neoplastic conditions, unrelated to known or potential risk factors for stomach cancer and long term diet modification.

For both cases and controls, data were collected during their hospital stay by centrally trained interviewers. The questionnaire included information on socio-demographic characteristics, anthropometric measures, selected lifestyle habits, a personal medical history and a family history of cancer. A satisfactorily reproducible [23] and valid [24] food frequency questionnaire (FFQ) was used to assess the patients' usual diet in the two years preceding diagnosis (for cases) or hospital admission (for controls). The FFQ included questions on 78 foods and beverages, including a range of the most common recipes in Italian diet. Subjects were asked to indicate the average weekly raw frequency and corresponding

portion size (small, medium, large) of consumption for each dietary variable. To estimate micro- and macro-nutrients, an Italian food composition database was used, integrated with other sources, when needed [25, 26].

3.2 Statistical Analysis

I applied EFA analyses to derive nutrient-dietary patterns, based on a set of 28 selected micro- and macro-nutrients, and tested their validity throughout CFA analyses. Statistical details of both procedures are described in Chapter 2.

Briefly:

Exploratory factor analysis

EFA was performed according to the PROC FACTOR procedure in SAS (version 9.1, SAS Institute), which uses principal components; uncorrelated factors were derived using orthogonal rotation. Varimax rotation that consists in rotating the axes to orientations that maximize variances of the loadings within the factors, while maximizing differences between the high and low loadings on a particular factor. Varimax orthogonal rotation provided a relatively clear information about which items correlated most strongly with a given factor. Factor scores generated for each individual were also more interpretable because explained variances among the factors do not overlap and are therefore independent of each other [11]. The factors are weighted combinations of nutrients, which best explain the variance in the nutrient intake (the correlation matrix). Factor loadings are the correlations between nutrients and the factor, and the individual factor scores are estimates of factor relationship to the individual's nutrient intake, and hence, the factor scores reflect the values of each of the nutrients that identify the factors.

Solutions from 2 to 6 factors were derived and rotated, and the scree plots and the factors themselves were observed to see which solution was most meaningful.

Confirmatory factor analysis

To decide how many factors to extract, I compared and verified the different CFA models that tested structures from 2 to 6 latent factors derived from EFA, in which I included nutrients decided on a priori, on the basis of the magnitude of their loadings in the previous EFA [17, 20]. Therefore, in contrast to EFA, I decided to retain only nutrients with factor loadings above a defined cut-point, in order to determine if a nutrient-dietary pattern could be represented and interpreted only by a set of nutrient variables, i.e., those nutrients highly correlated with the pattern. Therefore, each of the explored factors was tested in separate confirmatory models, considering only those nutrients having factor loadings greater or equal to |0.63| on a given factor. The contribution that a factor gives to a nutrient's sample variance is equal to the square of its loading on that factor. If we choose a |0.63| cut-off, we expected a minimum contribution of the factor on the nutrient's variance of approximately 0.40 [13]. Thus, in CFA models, the included nutrient items were allowed to load on only one factor, and loadings were fixed at zero for the other factor. Since, through the exploratory model with two factors, considering 0.63 as cut-off, there were nutrients having large loadings with more than one factor, a cut-off of 0.70 was also considered. Moreover, since the latent factors in CFA models were derived from orthogonal EFA solution, in a first step I fixed to zero the factors' covariance. Then, to improve the parsimony and interpretability of a CFA solution, I carried out

revised models, i.e. factors' covariance was specified to estimate the relationship between the latent dimensions.

In CFA, the observed correlation matrix, which takes into account the correlations, means, and variances of all nutrients, is used to calculate confirmed factor loadings. The maximum-likelihood parameter estimation method was used to estimate the variances of residual terms for nutrient item variables, covariance between factors, if any, and the estimated factor loadings.

The goodness of fit for a CFA is determined by using the comparative fit index (CFI), the normed and non-normed fit indices (NFI and NNFI). By convention, a CFI $\geq$0.90 and NNFI $\geq$0.90 indicate an acceptable fit [20]. The fit of the model is also judged by the root mean square residual (RMR) and by the root mean square error of approximation (RMSEA). By convention, RMR and RMSEA values close to 0 indicates a good fit [20]. To assess the fit of a CFA model, the chi-square test was also used. This test has as null hypothesis that the model fits the data. If the model provides a good fit, the chi-square value will be relatively small, and the corresponding *p*-value will be relatively large (above 0.05 and preferably closer to 1.00). However, with large samples and real-world data, the chi-square statistic is very frequently significant even if the model provides a good fit. This is particularly true with CFA models, which tend to be more complex than simple path analysis models. For these reasons, it is frequently appropriate to conclude that a CFA model fits the data even if *p* is significant [20].

CFA was performed according to the PROC CALIS procedure in SAS software (version 9.1, SAS Institute).

3.3 Results and conclusions

Using the EFA, the cumulative percentages of variance explained by six-, five-, four-, three-, and two-factor solutions approximately were equal to 84%, 80%, 75%, 69% and 63%, respectively (**Appendix1-Appendix5**). I excluded from CFA models the six-factor solution, since it showed a pattern based only on a single nutrient.

Diagram 1 shows the CFA model according the five-factor solution, in which only those nutrients with factor loading ≥0.63 are included.

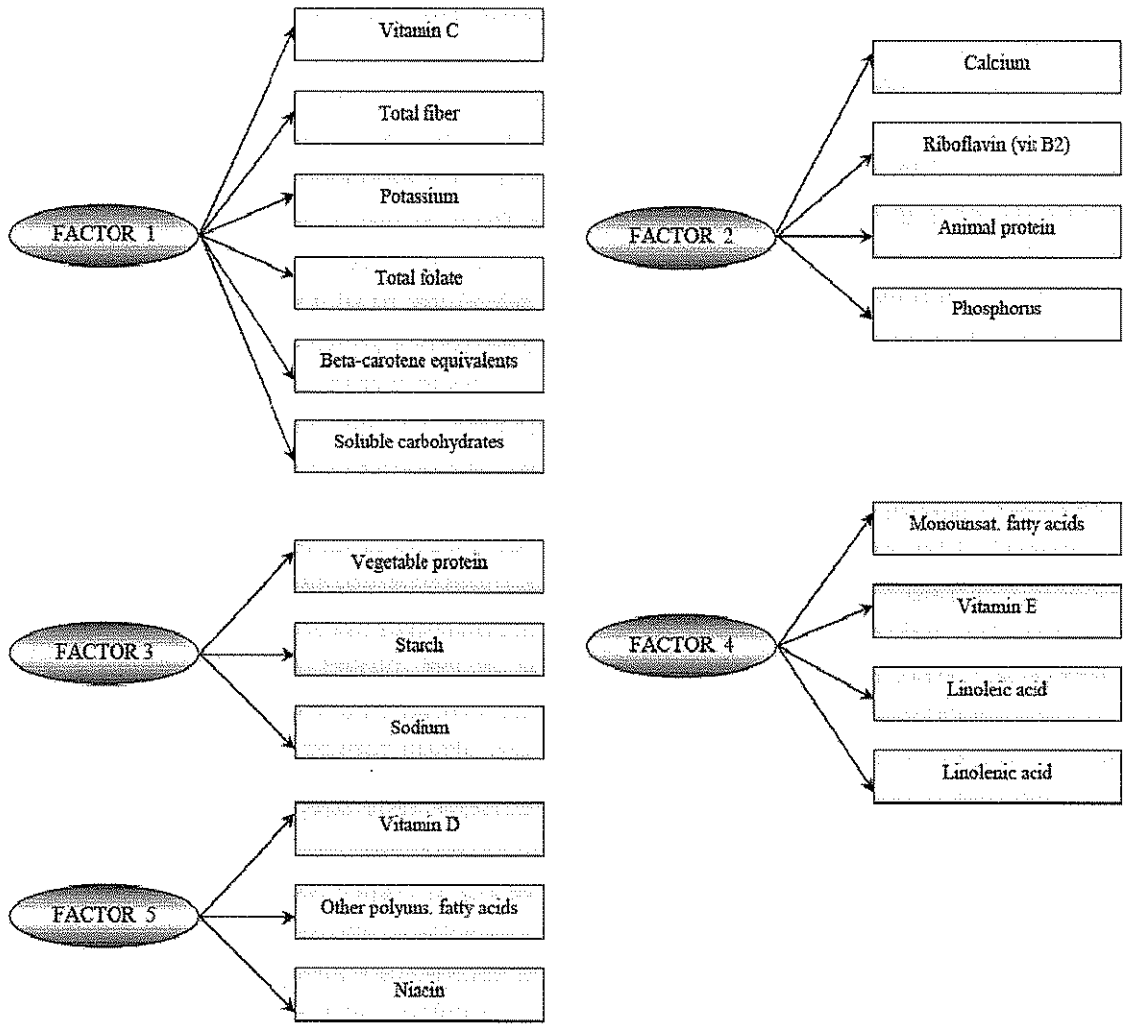


Diagram 1. Five-factors CFA model: nutrients with factor loading ≥ 0.63.

Diagram 2 shows the CFA model according the four-factor solution, in which only those nutrients with factor loading ≥0.63 are included.

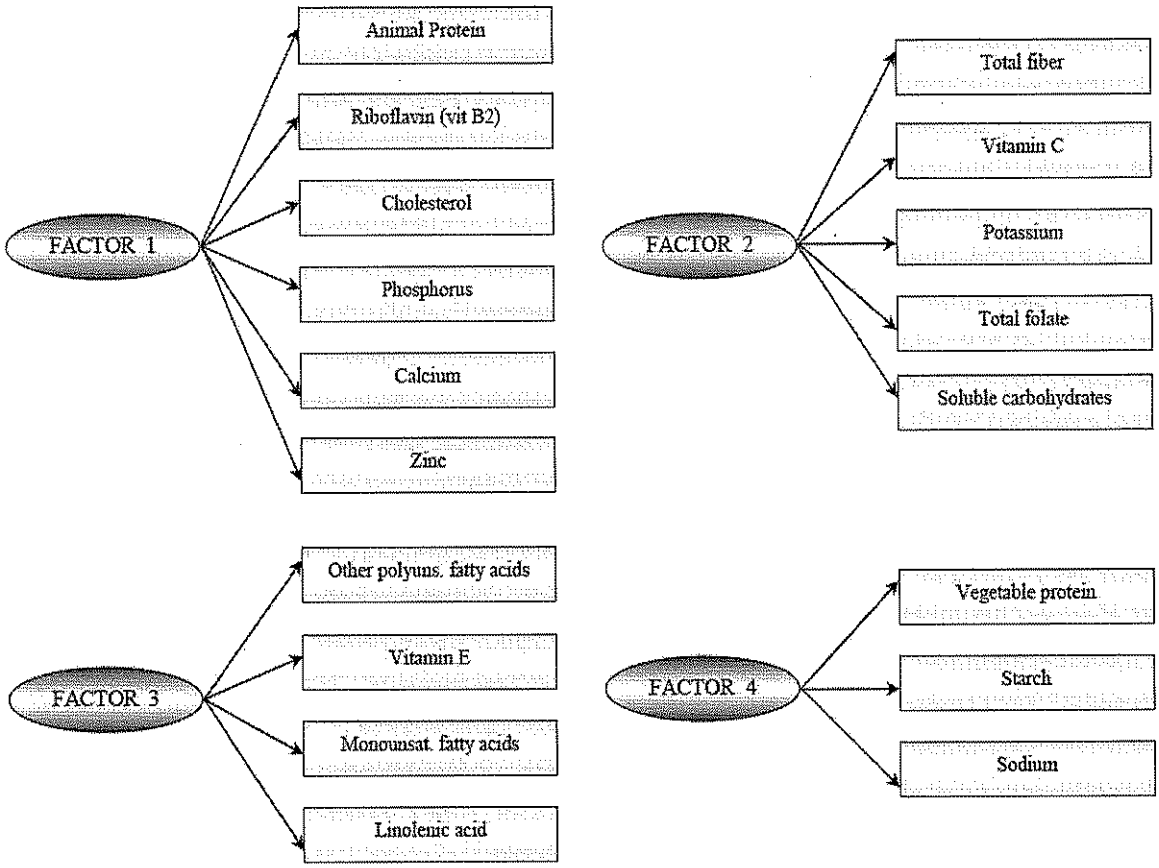


Diagram 2. Four-factors CFA model: nutrients with factor loading ≥ 0.63.

Diagram 3 shows the CFA model according the three-factor solution, in which only those nutrients with factor loading ≥0.63 are included.

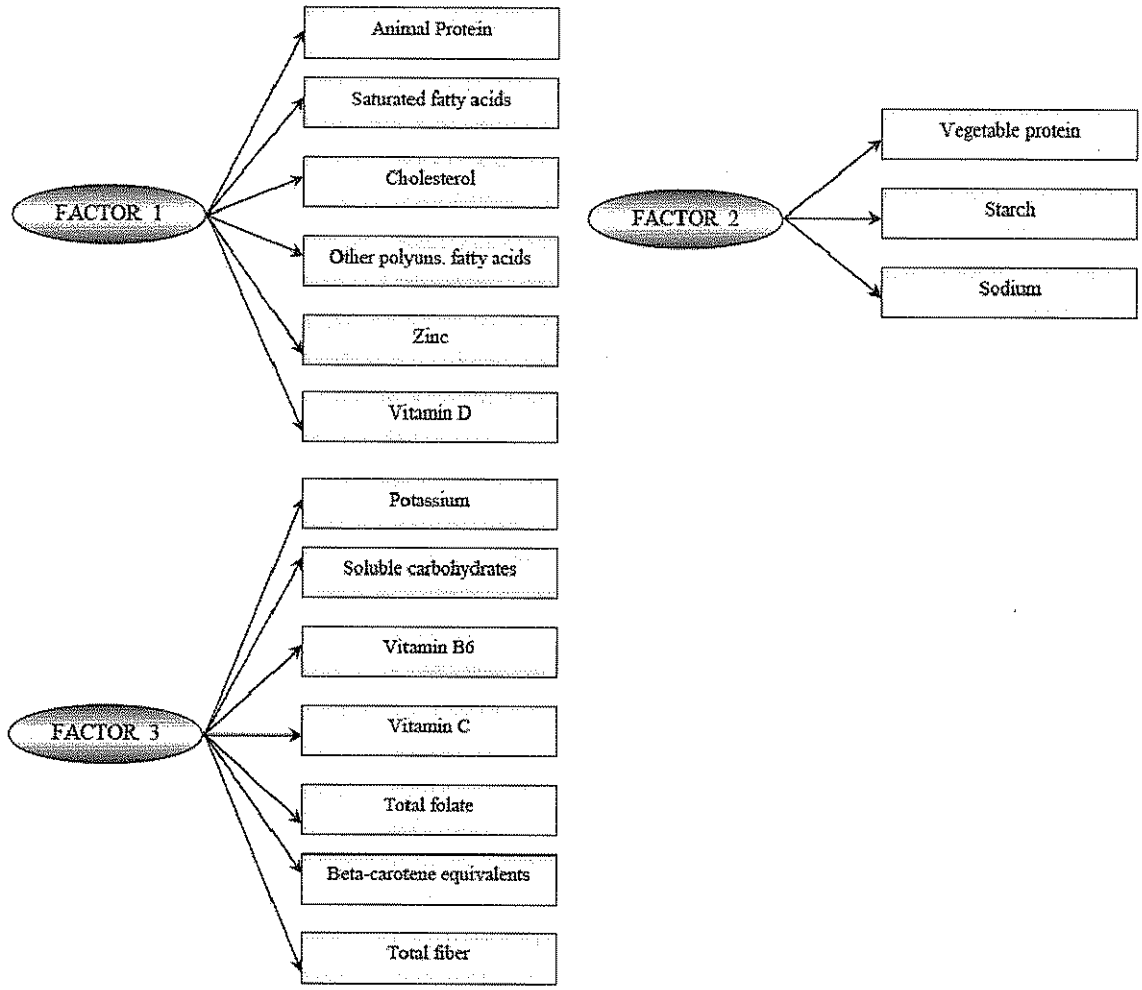


Diagram 3. Three-factors CFA model: nutrients with factor loading ≥ 0.63.

Diagram 4.1 shows the CFA model according the two-factor solution, in which only those nutrients with factor loading ≥0.63 are included.

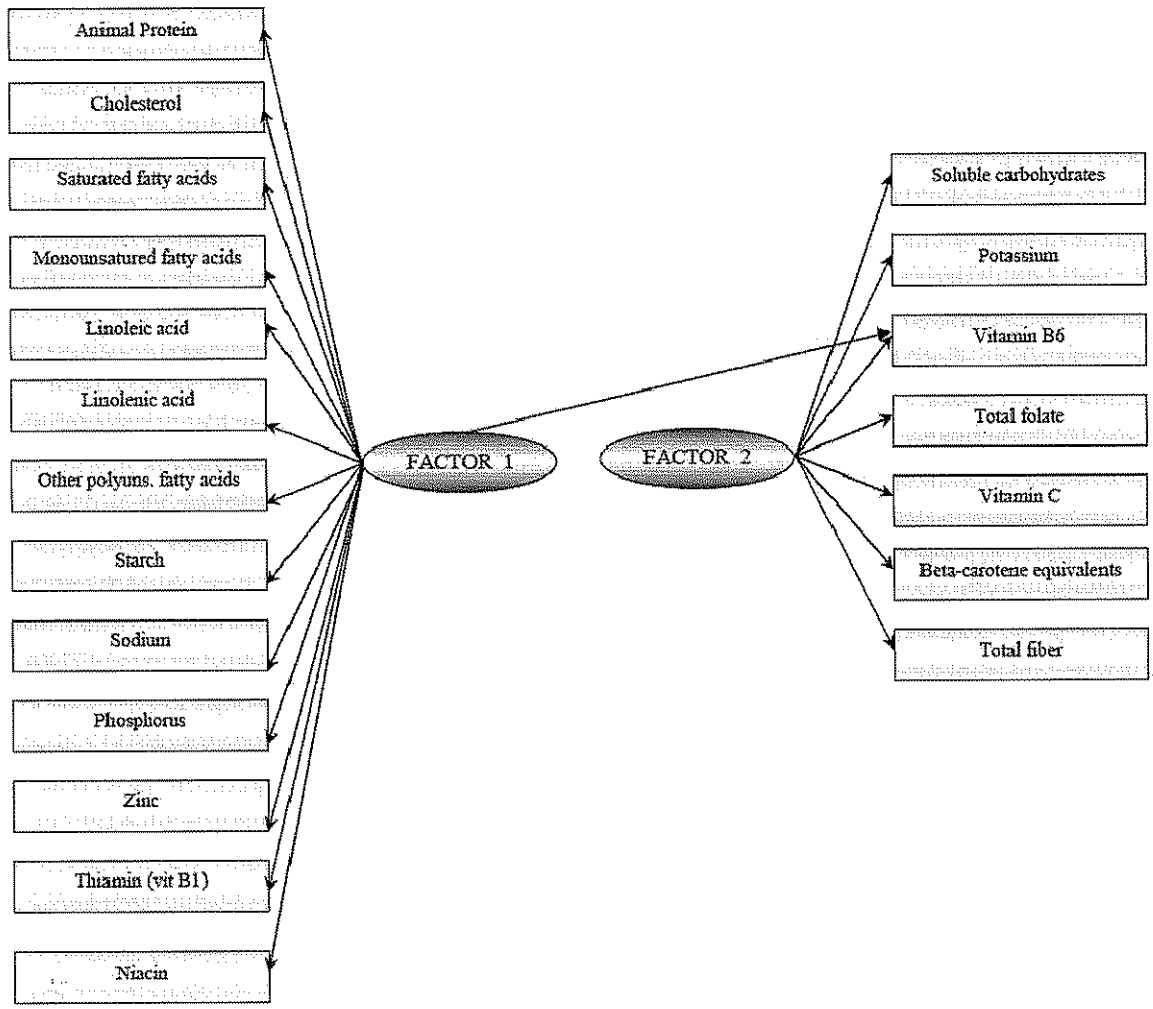


Diagram 4.1. Two-factors CFA model: nutrients with factor loading ≥ 0.63.

Since, through the two-factors solution considering 0.63 as cut-off, there was a nutrient loading on both the factors, a cut-off of 0.70 was also considered. Therefore, Diagram 4.2 shows the CFA model, in which only those nutrients with factor loading ≥0.70 are included.

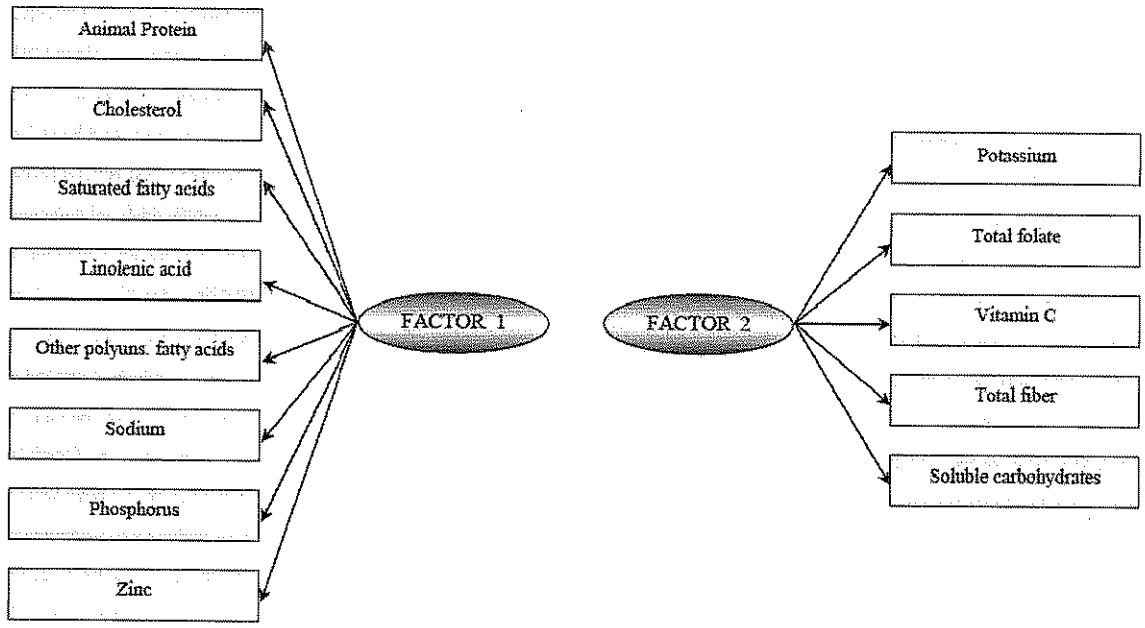


Diagram 4.2. Two-factors CFA model: nutrients with factor loading ≥ 0.70.

In the following table, the goodness of fit indices from the different CFA models considered, were reassumed.

Table 1. Goodness of fit indices for different confirmatory factor analysis (CFA) including nutrients with explored factor loadings ≥0.63.

CFA Model	F_{ML}	RMR	Chi-Square (*p*-value)	RMSEA	CFI	NNFI	NFI
Five-factor solution							
Factor covariance fixed to zero	11.95	0.41	9270.39 (<.0001)	0.26	0.57	0.52	0.57
Factor covariance free to estimate	9.28	0.11	7197.98 (<.0001)	0.24	0.67	0.61	0.67
Four-factor solution							
Factor covariance fixed to zero	12.05	0.41	9351.10 (<.0001)	0.26	0.59	0.55	0.59
Factor covariance free to estimate	9.98	0.09	7741.16 (<.0001)	0.24	0.67	0.61	0.66
Three-factor solution							
Factor covariance fixed to zero	7.17	0.35	5566.51 (<.0001)	0.26	0.65	0.60	0.65
Factor covariance free to estimate	6.03	0.11	4680.70 (<.0001)	0.24	0.71	0.65	0.70
Two-factor solution							
Factor covariance fixed to zero	8.96	0.29	6950.33 (<.0001)	0.23	0.66	0.62	0.66
Factor covariance free to estimate	8.18	0.09	6346.79 (<.0001)	0.22	0.69	0.65	0.69
Two-factor solution *							
Factor covariance fixed to zero	3.59	0.31	2788.30 (<.0001)	0.23	0.76	0.71	0.75
Factor covariance free to estimate	2.94	0.09	2279.07 (<.0001)	0.21	0.80	0.76	0.80

*Nutrients with explored factor loadings ≥0.70 were also included.
F_{ML} = Maximum-likelihood estimation of Fit Function.
RMR = Root mean square residual; RMSEA = Root mean square error of approximation.
CFI = Comparative fit index; NNFI = Non-normed fit index; NFI = Normed fit index.

The maximum-likelihood estimation of fit functions decreased from five- to two-factor solutions, to reach a small value of 2.94 in the CFA model including two factors with covariance among factors free to estimate. The chi-square test gives *p*-

values highly significant for each CFA model, that lead to reject that the models fit the data. However, because of the problems with this significance test, these findings by itself did not cause to reject the models.

Throughout the different CFA models with covariance among factors free to estimate, the RMR values were around the 0.1 threshold for an acceptable fit. The RMSEA values were around 0.2, somewhat higher than the threshold for an acceptable fit.

Considering the CFA models with covariance among factors fixed to zero, the CFI values were 0.57 for the five-factor model, 0.59 for the four-factor one, 0.65 for the three-factor one, and 0.66 for the two-factor model including nutrients with loadings ≥ 0.63 and 0.76 for that including nutrients with loadings ≥ 0.70. The CFI values for the CFA models with covariance among factors free to estimate, were higher compared to those with covariance among factors fixed to zero, to reach 0.80 for the two-factor model including nutrients with loadings ≥ 0.70, close to the 0.90 threshold for an acceptable fit. The NFI values were very similar than those of the CFI, while the NNFI values were lower.

Finally, all confirmed standardized coefficients, i.e. factor loadings, from CFA models, ranged from 0.5 to 1. The associated t tests (greater than 3.291 with $p<0.001$) indicated that the loading of each nutrient was significantly different from zero (data not shown).

It could conclude that the two-factors model was confirmed with moderately satisfactory values of goodness of fit indices (around 0.8, quite close to the 0.90 threshold for an acceptable fit). From this solution, it could emerge that the diet of the

study population could be characterized by two major profiles. The first confirmed dietary pattern, named "*Animal protein*", that had the greatest loadings on animal protein, cholesterol, saturated fatty acids, linolenic acids, other polyunsaturated fatty acids, sodium, phosphorus and zinc. The second one, named "*Vitamins and fiber*", had the greatest loadings on potassium, total folate, vitamin C, total fiber and soluble carbohydrates. These two dietary patterns are consistent with those of previous studies on gastric cancer [27-32], but also on cancers of other sites [33-36].

Nevertheless, results from all CFA models are not satisfactory. For this reason, to better understand the behavior of the goodness-of-fit indices, results from CFA applications on simulated data will be shown in the next paragraph.

3.4 CFA application on simulated data

In this paragraph, I will present results from CFA models applied on simulated datasets, in order to better understand the behavior of the goodness-of-fit indices computed by this statistical technique.

I simulated a new dataset characterized by a structure generated "*ad hoc*": I defined 24 variables (X1-X24), such as each variable was highly correlated only to one factor and a normally distributed random-error component. I defined four orthogonal factors in this way: the first factor was correlated to the variables from X1 to X9, the second one to those from X10 to X16, the third one to those from X17 to X21, the fourth one to those from X22 to X24. For this dataset, I generated 5,000 and 10,000 random samples with sample sizes of 100, 500, 750 and 1,000, respectively, in order to better

understand the behavior of the goodness-of-fit indices when the sample size increases. I tested confirmatory four-factor models in each of the 8 datasets. For each CFA model, a set of goodness-of-fit indices was computed. Moreover, the proportion of significant parameters was calculated by each factor, among the 8 datasets. I performed these simulations (**Appendix 6**) with the SAS software (version 9.1, SAS Institute).

Table 2. Effect of sample size on selected goodness-of-fit indices.

	5.000 samples				10.000 samples			
	100 %	500 %	750 %	1.000 %	100 %	500 %	750 %	1.000 %
***p*-value (Chi-Square)**								
< 0.01	11.5	1.9	1.3	1.5	11.9	1.6	1.4	1.3
0.01-<0.05	18.8	5.9	5.3	4.2	18.7	6.0	5.0	4.6
0.05-<0.10	12.2	6.8	6.2	5.8	12.9	6.6	6.3	5.7
0.10-<0.20	17.4	12.0	11.6	11.2	16.6	12.1	11.6	11.0
≥0.20	40.1	73.5	75.6	77.3	39.9	73.7	75.7	77.4
Chi-Square/DF								
< 2	100	100	100	100	100	100	100	100
≥ 2	0	0	0	0	0	0	0	0
RMR								
≤ 0.1	98.6	100	100	100	98.6	100	100	100
> 0.1	1.4	0	0	0	1.4	0	0	0
RMSEA								
≤ 0.1	100	100	100	100	100	100	100	100
> 0.1	0	0	0	0	0	0	0	0
GFI								
< 0.90	100	0	0	0	100	0	0	0
0.90-0.95	0	0.5	0	0	0	0.5	0	0
> 0.95	0	99.5	100	100	0	99.5	100	100
AGFI								
< 0.90	100	0	0	0	100	0	0	0
0.90-0.95	0	30.7	0	0	0	30.7	0	0
> 0.95	0	69.3	100	100	0	69.3	100	100
Probability of Close Fit								
< 0.90	44.9	0	0	0	45.8	0	0	0
0.90-0.95	15.1	0	0	0	14.4	0	0	0
> 0.95	40.0	100	100	100	39.8	100	100	100
Bentler's CFI								
< 0.90	73.5	3.7	0.4	0	73.6	3.6	0.3	0
0.90-0.95	6.1	17.7	8.5	3.2	6.1	17.8	8.4	3.3
> 0.95	20.4	78.6	91.2	96.8	20.3	78.6	91.3	96.7
NNFI								
< 0.90	74.7	5.1	0.5	0.1	74.7	5.0	0.6	0.1
0.90-0.95	5.3	18.4	10.9	4.6	5.4	18.6	10.7	4.5
> 0.95	20.0	76.5	88.6	95.3	19.9	76.4	88.7	95.4
NFI								
< 0.90	100	100	100	100	100	100	100	100
0.90-0.95	0	0	0	0	0	0	0	0
> 0.95	0	0	0	0	0	0	0	0

RMR = Root mean square residual; RMSEA = Root mean square error of approximation; GFI=Goodness-of-fit index; AGFI=Adjusted GFI; CFI = Comparative fit index; NNFI = Non-normed fit index; NFI = Normed fit index.

Table 2 shows the effect of sample size on selected goodness-of-fit indices, throughout the 8 simulations. The Chi-square test gave *p*-values ≥0.05, that lead to do not reject that the models fit the data, in 69.7% of the 5,000 samples with 100 observations, 92.3% with 500, 93.4% with 750, to reach a proportion of 94.3% with

1,000 observations. The 100% of the 5,000 samples, from 500 to 1,000 observations, showed values of RMR ≤0.1. Also the GFI, AGFI, the Probability of Close Fit, CFI, NNFI were close to 1, leading to confirm that the tested models fit the data, in particular when the sample size increases from 500 to 1,000. Similar results were obtained when 10,000 samples were considered.

Table 3. Effect of sample size on the proportion of significant parameters by each factor.

	5.000 samples				10.000 samples			
	100 obs.	**500 obs.**	**750 obs.**	**1.000 obs.**	**100 obs.**	**500 obs.**	**750 obs.**	**1.000 obs.**
	%	%	%	%	%	%	%	%
Latent Factor 1								
0	3.4	-	-	-	3.4	-	-	-
1	1.2	-	-	-	1.3	-	-	-
2	3.6	-	-	-	3.7	-	-	-
3	10.1	-	-	-	9.9	-	-	-
4	19.1	-	-	-	19.3	-	-	-
5	22.7	0.02	-	-	23.1	0.01	-	-
6	21.1	0.2	-	-	20.7	0.2	-	-
7	13.1	2.7	0.2	0.02	12.9	2.6	0.2	0.01
8	4.5	21.6	4.8	0.96	4.8	21.6	4.6	0.90
9	1.1	75.5	94.9	99.02	0.9	75.6	95.2	99.09
Latent Factor 2								
0	0.8	-	-	-	0.9	-	-	-
1	0.3	-	-	-	0.3	-	-	-
2	1.1	-	-	-	1.1	-	-	-
3	5.7	-	-	-	6.0	-	-	-
4	16.9	-	-	-	17.3	-	-	-
5	32.6	-	-	-	32.1	-	-	-
6	30.1	2.6	0.3	-	29.9	2.5	0.2	-
7	12.5	97.4	99.7	100	12.4	97.5	99.8	100
Latent Factor 3								
0	18.8	0.04	-	-	18.9	0.02	-	-
1	8.7	-	-	-	8.5	-	-	-
2	14.6	-	-	-	14.8	0.02	-	-
3	26.1	0.2	-	-	25.7	0.2	-	-
4	22.4	4.3	0.3	0.02	22.3	4.2	0.3	0.01
5	9.5	95.5	99.7	99.98	9.9	95.6	99.7	99.99
Latent Factor 4								
0	18.3	-	-	-	18.4	0.01	-	-
1	3.8	-	-	-	3.7	-	-	-
2	7.5	-	-	-	7.3	-	-	-
3	70.4	100	100	100	70.5	99.99	100	100

Table 3 shows the effect of sample size on the significance of the estimated confirmed factor loadings, in terms of proportion of significant factor loadings by each factor,

among the 8 simulations. Also the proportion of significant parameters increased systematically as sample size increased from 500 to 1,000.

In summary, the simulations showed that the goodness-of-fit indices improve systematically as sample size increases.

4. DISCUSSION

In my PhD thesis, I studied factor analyses methods in nutritional epidemiology. Initially, I applied EFA on a set of 28 nutrients in a case-control study of gastric cancer, conducted in northern Italy, to reduce a set of factors that summarizes and describes the structural interrelationships among the items in a concise and understandable way.

The choice of number of factors to retain is influenced by three major statistical criteria [11]: factor eigenvalue greater or equal to 1, scree plot construction and factor interpretability. The first mentioned criterion is to be considered with due caution, because the researcher may over or under-estimate the correct number of factors. For example, if there are large numbers of items in the data set, there will also be large numbers of eigenvalues that satisfy this criterion. Nevertheless, there is no precise solution to determining the number of factor to extract. Given the same data set, a team of researchers might arrive at very different solutions.

CFA has among its objectives overcome the ambiguity because it is designed to test a hypothesis about the relation of certain hypothetical common factor variables, whose number and interpretation are given in advance, to the observed variables [17]. To decide how many valid factors to retain, I compared and verified the different CFA models that tested structures from 2 to 6 latent factors derived from EFA, in which I included nutrients decided on a priori, on the basis of the magnitude of their loadings in the previous EFA. I excluded from this analysis the six-factor solution, since it

showed a pattern based only on a single nutrient. Throughout the comparison among the different solutions from 2 to 5 latent factors, with the help of the application of CFA, the two-factors model was confirmed with moderately satisfactory values of goodness of fit indices around 0.8, close to the 0.90 threshold for an acceptable fit. Nevertheless, results from CFA models are not satisfactory.

Subsequently, in order to better understand the performance of this statistical technique, I tested and compared results from CFA applied on simulated datasets characterized by a structure "*ad hoc*", (such as each variable was highly correlated only to one factor, for a total of four orthogonal factors). In this case, I verified that CFA technique provides satisfactory results, in particular when the sample size is at least of 500, although limitations regarding some goodness-of-fit indices remain [37].

In general, CFA, allowing the researcher to test the hypothesis that a relationship between the observed variables and their underlying latent construct exists, has the fundamental advantage over EFA in that it allows to control every aspect of the model specification (e.g., generate an unstandardized solution, specify correlated errors, place various constraints on the solution, such as fixing cross-loadings to zero or holding model parameters to equality). This method is more rigorous than EFA because the researcher is able to create a pattern using prior theory and therefore requires fewer subjective decisions. CFA is an intuitively appealing method because it can be based in theory and also reduces some of the subjectivity involved in explorative procedures [18]. Unfortunately, CFA cannot confirm that this is the best

fit from the infinity of possible models that might have been tried [17] and it cannot test the degree of model agreement. On the other hand, there are several **limitations**. As emerged by the application described in the previous chapter, goodness-of-fit indices did not demonstrate satisfactory values. Several causes may have reduced these values, among these: 1) a relatively large number of indicator variables, i.e., nutrients, 2) large residuals and 3) a part of not explained variance. The large number of indicator variables often results in large χ^2 values that make it difficult to fit the model with data [20]. Another reason for the less-than-perfect fit of the model is measurement error, which remains a problem in all dietary studies [38] even when a validated dietary instrument is used [24].

In conclusion, using confirmatory factory analysis together with exploratory factor analysis overcomes major methodological problems and subjective aspects, in determining a valid latent factor structure under a set of observed variables. Moreover, a different use of the CFA could be particularly useful. For example if the confirmed factors were tested in a different study as true *a priori* factors: the factors identified in one group could be applied in a different group using CFA based on the same nutrients to compute scores. Hence, the factor scores could be acceptable and robust as markers of nutrient intake pattern on group levels and may prove useful in studies of diet–disease relationships. Nevertheless, until factor analysis gains more experience in nutrition, it will be difficult to define valid criteria for a good fit in this discipline and methodologies for improving fit.

REFERENCES

1. Kant AK. Dietary patterns and health outcomes. J Am Diet Assoc 2004; 104: 615-635.

2. Moeller SM, Reedy J, Millen AE et al. Dietary patterns: challenges and opportunities in dietary patterns research an Experimental Biology workshop, April 1, 2006. J Am Diet Assoc 2007; 107: 1233-1239.

3. Newby PK, Tucker KL. Empirically derived eating patterns using factor or cluster analysis: a review. Nutr Rev 2004; 62: 177-203.

4. Hu FB. Dietary pattern analysis: a new direction in nutritional epidemiology. Curr Opin Lipidol 2002; 13: 3-9.

5. Togo P, Heitmann BL, Sorensen TI, Osler M. Consistency of food intake factors by different dietary assessment methods and population groups. Br J Nutr 2003; 90: 667-678.

6. Maskarinec G, Novotny R, Tasaki K. Dietary patterns are associated with body mass index in multiethnic women. J Nutr 2000; 130: 3068-3072.

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

8. Park SY, Murphy SP, Wilkens LR et al. Dietary patterns using the Food Guide Pyramid groups are associated with sociodemographic and lifestyle factors: the multiethnic cohort study. J Nutr 2005; 135: 843-849.

9. Pierce BL, Austin MA, Crane PK et al. Measuring dietary acculturation in Japanese Americans with the use of confirmatory factor analysis of food-frequency data. Am J Clin Nutr 2007; 86: 496-503.

10. Lau C, Glumer C, Toft U et al. Identification and reproducibility of dietary patterns in a Danish cohort: the Inter99 study. Br J Nutr 2008; 99: 1089-1098.

11. Pett MA, Lackey NR, Sullivan JJ. Making sense of factor analysis: the use of factor analysis for instrument development in health care research. Thousand Oaks, CA: Sage Publications; 2003.

12. Cattell R. The scree test for the number of factors. Multivariate Behavioral Research 1966; 1: 245:276.

13. Comrey A, Lee H. A first course in factor analysis. 2nd edition. New Kersey: Lawrence Erlbaum Associates, Publishers. 1992.

14. Kleinbaum D, Kupper L, Muller K, Nizam A. Applied Regression Analysis and Other Multivariate Methods. Duxury Press. 1998.

15. Brown T. Confirmatory factor analysis for applied research. New York: Guilford. 2006.

16. Tatsuoka M. Multivariate Analysis. New York: John Wiley & Sons, Inc. 1971.

17. Brown TA. Confirmatory Factor Analysis for Applied Research. New York: The Guilford Press; 2006.

18. Thompson B. Exploratory and Confirmatory factor analysis. Understanding Concepts and Applications. Washington, DC: American Psychological Association; 2004.

19. Meyers LS, Gamst G, Guarino AJ. Applied Multivariate Research: Design and Interpretation. Thousand Oaks, CA: Sage Publications; 2006.

20. Hatcher L. A step-by-step approach to using the SAS System for factor analysis and structural equation modeling. Cary, NC: SAS Institute Inc. 1994.

21. Bentler PM. Comparative fit indices in structural models. Psychol Bull 1990; 107: 238-246.

22. Bertuccio P, Edefonti V, Bravi F et al. Nutrient dietary patterns and gastric cancer risk in Italy. Cancer Epidemiol Biomarkers Prev 2009; 18: 2882-2886.

23. Franceschi S, Barbone F, Negri E et al. Reproducibility of an Italian food frequency questionnaire for cancer studies. Results for specific nutrients. Ann Epidemiol 1995; 5: 69-75.

24. Decarli A, Franceschi S, Ferraroni M et al. Validation of a food-frequency questionnaire to assess dietary intakes in cancer studies in Italy. Results for specific nutrients. Ann Epidemiol 1996; 6: 110-118.

25. Gnagnarella P, Parpinel M, Salvini S et al. The update of the Italian food composition database. J Food Comp Analysis 2004; 17: 509-522.

26. Salvini S, Parpinel M, Gnagnarella P et al. Banca di composizione degli alimenti per studi epidemiologici in Italia. Milano, Italia: Istituto Europeo di Oncologia; 1998.

27. Campbell PT, Sloan M, Kreiger N. Dietary patterns and risk of incident gastric adenocarcinoma. Am J Epidemiol 2008; 167: 295-304.

28. Palli D, Russo A, Decarli A. Dietary patterns, nutrient intake and gastric cancer in a high-risk area of Italy. Cancer Causes Control 2001; 12: 163-172.

29. De Stefani E, Correa P, Boffetta P et al. Dietary patterns and risk of gastric cancer: a case-control study in Uruguay. Gastric Cancer 2004; 7: 211-220.

30. Bastos J, Lunet N, Peleteiro B et al. Dietary patterns and gastric cancer in a Portuguese urban population. Int J Cancer 127: 433-441.

31. Kim MK, Sasaki S, Sasazuki S, Tsugane S. Prospective study of three major dietary patterns and risk of gastric cancer in Japan. Int J Cancer 2004; 110: 435-442.

32. Masaki M, Sugimori H, Nakamura K, Tadera M. Dietary patterns and stomach cancer among middle-aged male workers in Tokyo. Asian Pac J Cancer Prev 2003; 4: 61-66.

33. De Stefani E, Deneo-Pellegrini H, Boffetta P et al. Dietary patterns and risk of cancer: a factor analysis in Uruguay. Int J Cancer 2009; 124: 1391-1397.

34. Randi G, Edefonti V, Ferraroni M et al. Dietary patterns and the risk of colorectal cancer and adenomas. Nutr Rev 68: 389-408.

35. Edefonti V, Randi G, La Vecchia C et al. Dietary patterns and breast cancer: a review with focus on methodological issues. Nutr Rev 2009; 67: 297-314.

36. Edefonti V, Bravi F, La Vecchia C et al. Nutrient-based dietary patterns and the risk of oral and pharyngeal cancer. Oral Oncol 46: 343-348.

37. Mulaik SA, James, L.R., van Alstie, J., Bennett, N., Lind, S., Stilwell, C. D. Evaluation of goodness-of-fit indices for structural equation models. Psychol Bull 1989; 105: 430-445.

38. Willett W. Nutritional Epidemiology. 2nd edition. New York: Oxford University Press. 1998.

APPENDIXES

Appendix 1. Factor loading matrix[1], explained variances from principal component factor analysis: six-factor solution.

Nutrient	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Animal protein	0.10502	**0.68759**	0.21256	0.25475	0.54771	0.09762
Vegetable protein	0.35223	0.16516	**0.85339**	0.24491	0.08340	0.06107
Cholesterol	0.05586	0.57604	0.28965	0.30705	0.45487	0.24816
Saturated fatty acids	0.11270	**0.63723**	0.27658	0.57198	0.23493	0.05116
Monounsaturated fatty acids	0.25643	0.19302	0.23244	**0.75703**	0.24841	0.06773
Linoleic acid	0.11959	0.21927	0.24680	**0.78915**	0.18685	0.05964
Linolenic acid	0.23023	0.36882	0.23557	**0.77327**	0.20624	0.08377
Other polyunsaturated fatty acids	0.01851	0.13793	0.12848	0.36882	**0.82649**	0.09173
Soluble carbohydrates	**0.65693**	0.45358	0.13892	0.02088	0.10479	-0.03732
Starch	0.06963	0.22725	**0.90008**	0.22872	0.06051	0.06245
Sodium	0.01334	0.50491	**0.74018**	0.19190	0.05780	0.06974
Calcium	0.31012	0.84878	0.07725	0.19184	0.03133	-0.04107
Potassium	**0.75503**	0.35648	0.33151	0.18836	0.29519	0.03955
Phosphorus	0.35475	**0.67289**	0.41055	0.25878	0.33575	0.10975
Iron	0.47551	0.22762	0.47141	0.25158	0.36043	0.29297
Zinc	0.28008	0.50350	0.50990	0.31176	0.44959	0.18003
Thiamin	0.48918	0.50743	0.44647	0.23853	0.27459	0.08700
Riboflavin	0.43996	**0.68654**	0.18426	0.16570	0.18227	0.40464
Vitamin B6	0.58447	0.37243	0.37435	0.23912	0.44720	0.14667
Total folate	**0.68997**	0.27961	0.30211	0.23785	0.13045	0.38734
Niacin	0.38952	0.22753	0.38729	0.16020	**0.63080**	0.24256
Vitamin C	**0.85489**	0.09385	-0.07058	0.10850	0.10136	0.00076
Retinol	0.05651	0.08796	0.05077	0.06168	0.11281	**0.94883**
Beta-carotene equivalents	**0.65635**	0.04631	0.01306	0.27385	-0.00899	0.07978
Lycopene	0.27442	-0.17580	0.48346	0.25387	0.35562	-0.11105
Vitamin D	0.10213	0.18948	-0.03353	0.10434	**0.85491**	0.00929
Vitamin E	0.49876	0.08735	0.17857	**0.78877**	0.19205	0.02085
Total fiber	**0.83742**	0.09885	0.34835	0.14993	0.01773	-0.03179
Proportion of variance explained (%)	19.36	16.77	15.27	13.93	12.80	5.60
Cumulative variance explained (%)	19.36	36.13	51.40	65.33	78.13	83.73

[1]Loadings greater or equal to 0.63 (in absolute value) were shown in bold typeface.

Appendix 2. Factor loading matrix[1], explained variances from principal component factor analysis: five-factor solution.

Nutrient	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Animal protein	0.10635	**0.63208**	0.20554	0.32300	0.56646
Vegetable protein	0.36358	0.14535	**0.85319**	0.23999	0.10550
Cholesterol	0.06828	0.55369	0.29738	0.31680	0.52948
Saturated fatty acids	0.12674	0.62206	0.28398	0.58269	0.23882
Monounsaturated fatty acids	0.27381	0.19260	0.23784	**0.74784**	0.22140
Linoleic acid	0.13914	0.22684	0.25638	**0.77135**	0.16077
Linolenic acid	0.25031	0.37324	0.24619	**0.75494**	0.19726
Other polyunsaturated fatty acids	0.01419	0.06754	0.10684	0.47131	**0.78228**
Soluble carbohydrates	**0.65392**	0.41843	0.12942	0.05219	0.10995
Starch	0.08210	0.21092	**0.90401**	0.22359	0.08899
Sodium	0.02536	0.48851	**0.74805**	0.18956	0.10468
Calcium	0.31414	**0.82757**	0.08114	0.21103	0.04994
Potassium	**0.75615**	0.31276	0.31956	0.22353	0.30140
Phosphorus	0.36216	**0.63481**	0.41003	0.28755	0.37805
Iron	0.49022	0.21541	0.47703	0.22972	0.44749
Zinc	0.28988	0.46560	0.50920	0.33772	0.49794
Thiamin	0.49585	0.47239	0.44343	0.26059	0.30613
Riboflavin	0.46203	**0.70572**	0.21004	0.09342	0.35277
Vitamin B6	0.58928	0.32909	0.36624	0.27160	0.47977
Total folate	**0.71318**	0.30570	0.32301	0.15054	0.27385
Niacin	0.39304	0.17807	0.37747	0.20204	**0.68180**
Vitamin C	**0.85358**	0.07636	-0.08054	0.11568	0.09404
Retinol	0.10320	0.20188	0.11537	-0.16788	0.46614
Beta-carotene equivalents	**0.66620**	0.05954	0.01748	0.23456	0.00978
Lycopene	0.26847	-0.23272	0.45826	0.33118	0.26671
Vitamin D	0.08576	0.09641	-0.06646	0.24513	**0.79487**
Vitamin E	0.51485	0.08817	0.18021	**0.77677**	0.14403
Total fiber	**0.83963**	0.07759	0.33896	0.15357	0.00744
Proportion of variance explained (%)	19.93	15.50	15.32	14.61	14.46
Cumulative variance explained (%)	19.93	35.43	50.75	65.36	79.82

[1] Loadings greater or equal to 0.63 (in absolute value) were in bold typeface.

Appendix 3. Factor loading matrix[1], explained variances from principal component factor analysis: four-factor solution.

Nutrient	Factor 1	Factor 2	Factor 3	Factor 4
Animal protein	**0.80333**	0.10154	0.40815	0.22550
Vegetable protein	0.14804	0.38682	0.29293	0.80321
Cholesterol	**0.71851**	0.06580	0.41052	0.30168
Saturated fatty acids	0.56207	0.15062	0.49742	0.41443
Monounsaturated fatty acids	0.20189	0.28964	**0.72480**	0.28199
Linoleic acid	0.18528	0.16115	**0.71117**	0.33357
Linolenic acid	0.32586	0.27308	**0.68115**	0.34333
Other polyunsaturated fatty acids	0.48012	-0.01790	**0.74702**	-0.04280
Soluble carbohydrates	0.39883	**0.66117**	0.01810	0.16741
Starch	0.18438	0.10982	0.25846	**0.88278**
Sodium	0.41425	0.05526	0.16226	**0.80298**
Calcium	**0.65450**	0.34035	0.03274	0.27707
Potassium	0.41506	**0.75731**	0.29003	0.27915
Phosphorus	**0.69606**	0.37244	0.31218	0.44890
Iron	0.42188	0.48436	0.39104	0.37379
Zinc	**0.63032**	0.29195	0.45272	0.47479
Thiamin	0.53095	0.50584	0.29657	0.44893
Riboflavin	**0.76249**	0.46568	0.10243	0.26141
Vitamin B6	0.52849	0.58265	0.40886	0.29149
Total folate	0.40016	**0.71418**	0.21604	0.28151
Niacin	0.54013	0.36873	0.46978	0.21047
Vitamin C	0.12224	**0.84836**	0.12649	-0.10909
Retinol	0.46556	0.07790	0.02563	0.00432
Beta-carotene equivalents	0.03823	**0.67176**	0.20207	0.02329
Lycopene	-0.05420	0.26433	0.49406	0.31707
Vitamin D	0.54019	0.04342	0.53723	-0.22896
Vitamin E	0.07509	0.53126	**0.73636**	0.21555
Total fiber	0.06030	0.85101	0.15368	0.30607
Proportion of variance explained (%)	21.67	20.30	18.02	15.10
Cumulative variance explained (%)	21.67	41.97	59.99	75.09

[1]Loadings greater or equal to 0.63 (in absolute value) were shown in bold typeface.

Appendix 4. Factor loading matrix[1], explained variances from principal component factor analysis: three-factor solution.

Nutrient	Factor 1	Factor 2	Factor 3
Animal protein	**0.82616**	0.36030	0.18082
Vegetable protein	0.08790	0.86283	0.38626
Cholesterol	**0.74592**	0.43109	0.13520
Saturated fatty acids	**0.63200**	0.56346	0.20196
Monounsaturated fatty acids	0.46268	0.50407	0.30724
Linoleic acid	0.44390	0.54655	0.17671
Linolenic acid	0.53303	0.55080	0.30201
Other polyunsaturated fatty acids	**0.79433**	0.20421	0.03595
Soluble carbohydrates	0.24853	0.18202	**0.69520**
Starch	0.11062	**0.92428**	0.11278
Sodium	0.27173	**0.82217**	0.08293
Calcium	0.47706	0.29133	0.40018
Potassium	0.37274	0.37539	**0.79182**
Phosphorus	0.62074	0.54368	0.43459
Iron	0.43777	0.49327	0.51980
Zinc	**0.64129**	0.61004	0.34811
Thiamin	0.46340	0.53732	0.55037
Riboflavin	0.59422	0.30240	0.53651
Vitamin B6	0.54197	0.42436	**0.63008**
Total folate	0.32563	0.35341	**0.74697**
Niacin	0.61791	0.36388	0.42035
Vitamin C	0.10980	-0.04914	**0.85897**
Retinol	0.39182	0.02275	0.12499
Beta-carotene equivalents	0.07038	0.09605	**0.67267**
Lycopene	0.12656	0.45910	0.25406
Vitamin D	**0.76518**	-0.03640	0.10544
Vitamin E	0.35380	0.44552	0.53594
Total fiber	-0.00587	0.35209	**0.84763**
Proportion of variance explained (%)	24.00	22.64	22.64
Cumulative variance explained (%)	24.00	46.64	69.28

[1]Loadings greater or equal to 0.63 (in absolute value) were shown in bold typeface.

Appendix 5. Factor loading matrix[1], explained variances from principal component factor analysis: two-factor solution.

Nutrient	Factor 1	Factor 2
Animal protein	**0.86456**	0.19663
Vegetable protein	0.55794	0.53408
Cholesterol	**0.84528**	0.16966
Saturated fatty acids	**0.82825**	0.26630
Monounsaturated fatty acids	**0.65035**	0.36831
Linoleic acid	**0.66907**	0.24924
Linolenic acid	**0.73485**	0.36780
Other polyunsaturated fatty acids	**0.75576**	0.02737
Soluble carbohydrates	0.26194	**0.70139**
Starch	**0.63054**	0.27592
Sodium	**0.70097**	0.21839
Calcium	0.52939	0.41913
Potassium	0.47010	**0.82493**
Phosphorus	**0.79226**	0.49147
Iron	0.61006	0.57625
Zinc	**0.85383**	0.41777
Thiamin	**0.65479**	0.61293
Riboflavin	0.62097	0.54815
Vitamin B6	**0.64543**	**0.66560**
Total folate	0.42221	**0.77956**
Niacin	**0.68408**	0.44426
Vitamin C	0.00266	**0.82711**
Retinol	0.31944	0.10422
Beta-carotene equivalents	0.06951	**0.67361**
Lycopene	0.35766	0.32718
Vitamin D	0.58487	0.05253
Vitamin E	0.51332	0.58805
Total fiber	0.14910	**0.89718**
Proportion of variance explained (%)	37.04	26.11
Cumulative variance explained (%)	37.04	63.15

[1]Loadings greater or equal to 0.63 (in absolute value) were shown in bold typeface.

Appendix 6

SAS Program for the simulations

```
%let s=12101951;

DATA CASUALI;
DO CAMP=1 TO 5000;
DO CAS=1 TO 100;


  x1=2.464*rannor(&s);
  x2=2.387*rannor(&s);
  x3=2.245*rannor(&s);
  x4=2.057*rannor(&s);

  y11=x1+1.97*2.464*rannor(&s);
  y12=x1+2.38*2.464*rannor(&s);
  y13=x1+2.82*2.464*rannor(&s);
  y14=x1+3.04*2.464*rannor(&s);
  y15=x1+3.45*2.464*rannor(&s);
  y16=x1+3.70*2.464*rannor(&s);
  y17=x1+4.38*2.464*rannor(&s);
  y18=x1+4.60*2.464*rannor(&s);
  y19=x1+4.60*2.464*rannor(&s);

  y21=x2+1.49*2.387*rannor(&s);
  y22=x2+1.52*2.387*rannor(&s);
  y23=x2+2.43*2.387*rannor(&s);
  y24=x2+2.86*2.387*rannor(&s);
  y25=x2+3.28*2.387*rannor(&s);
  y26=x2+3.39*2.387*rannor(&s);
  y27=x2+4.17*2.387*rannor(&s);

  y31=x3+2.53*2.245*rannor(&s);
  y32=x3+2.54*2.245*rannor(&s);
  y33=x3+2.75*2.245*rannor(&s);
  y34=x3+2.89*2.245*rannor(&s);
  y35=x3+3.29*2.245*rannor(&s);

  y41=x4+1.27*2.057*rannor(&s);
  y42=x4+1.97*2.057*rannor(&s);
  y43=x4+1.97*2.057*rannor(&s);

OUTPUT;
END; END;
RUN;


proc sort data=casuali; by camp; run;
```

```
proc calis data=casuali corr residual pall modification
outram=provaram;
by camp;
lineqs
y11 =  p1 f1 + e1,
y12 =  p2 f1 + e2,
y13 =  p3 f1 + e3,
y14 =  p4 f1 + e4,
y15 =  p5 f1 + e5,
y16 =  p6 f1 + e6,
y17 =  p7 f1 + e7,
y18 =  p8 f1 + e8,
y19 =  p9 f1 + e9,

y21 =  p10 f2 + e10,
y22 =  p11 f2 + e11,
y23 =  p12 f2 + e12,
y24 =  p13 f2 + e13,
y25 =  p14 f2 + e14,
y26 =  p15 f2 + e15,
y27 =  p16 f2 + e16,

y31 =  p17 f3 + e17,
y32 =  p18 f3 + e18,
y33 =  p19 f3 + e19,
y34 =  p20 f3 + e20,
y35 =  p21 f3 + e21,

y41 =  p22 f4 + e22,
y42 =  p23 f4 + e23,
y43 =  p24 f4 + e24;

std
e1-e24=var1-var24,
f1=1, f2=1, f3=1, f4=1;
cov
f1 f2 = 0,
f1 f3 = 0,
f1 f4 = 0,
f2 f3 = 0,
f2 f4 = 0,
f3 f4 = 0;

var
y11-y19 y21-y27 y31-y35 y41-y43;

run;
```

RINGRAZIAMENTI

- ✓ Università degli Studi di Milano, Facoltà di Medicina e Chirurgia, Dipartimento di Medicina del Lavoro "Clinica del Lavoro L. Devoto" Sezione di Statistica Medica e Biometria "G.A. Maccacaro".

- ✓ Istituto di Ricerche Farmacologiche "Mario Negri", Dipartimento di Epidemiologia.

- ✓ Fondazione Italiana per la Ricerca sul Cancro (FIRC).