UNIVERSITÀ DEGLI STUDI DI MILANO

Dipartimento di Scienze Cliniche e di Comunità

Laboratorio di Statistica Medica, Biometria, Epidemiologia "G. A. Maccacaro"

DOTTORATO DI RICERCA IN EPIDEMIOLOGIA, AMBIENTE E SANITÀ PUBBLICA

XXXI CICLO

SETTORE SCIENTIFICO DISCIPLINARE MED/01 STATISTICA MEDICA

A LATENT VARIABLE APPROACH TO DIETARY PATTERNS RESEARCH

DOTTORANDO:

Michela Dalmartello

Matricola: R11430

TUTOR:

Prof.ssa Monica Ferraroni

CO-TUTOR:

Prof. Jeroen Vermunt

Prof. Adriano Decarli

COORDINATORE DEL DOTTORATO:

Prof. Carlo La Vecchia

A.A. 2017-2018

Index

ABSTRACT vii
ACRONYMS x
1. INTRODUCTION 11
1.1. Background 11
1.2. Problem statement 11
1.3. Research Purpose 12
1.4. Outline of the thesis 13
2. DIETARY PATTERN ANALYSIS THROUGH LCA 15
2.1. Introduction 15
2.2. Identifying dietary patterns: a comparison of methodologies 15
2.2.1. Dietary pattern definition and interpretation 16
2.2.2. Latent variables 16
2.2.3. Statistical model 17
2.2.4. Indicators 17
2.2.5. Inclusion of external variables 18
2.2.6. Classification of subjects 18
2.3. Dietary patterns and the risk of oral/pharyngeal and esophageal cancer 19
2.3.1. The network of case-control studies 19
2.3.2. Dietary pattern and the risk of oral/pharyngeal and esophageal cancer: a comparison of the works on the multicentric case-control studies, Italy 20
3. METHODS 23
3.1. Introduction 23
3.2. Latent Class Analysis 23
3.2.1. Basic LCA 23
3.2.2. Extensions of traditional LCA 24
3.2.3. Fitting LCA 24
3.3. Three-Step analysis 25
3.4. Latent Class Trees 26
3.5. Estimation 27
4. DIETARY PATTERNS AND THE RISK OF ORAL AND PHARYNGEAL CANCER USING LATENT CLASS ANALYSIS 29
4.1. Introduction 29
4.2. Materials and methods 29

4.2.1. Study population 29
4.2.2. Dietary intake assessment 30
4.2.3. Statistical methods 30
4.3. Results 31
4.4. Discussion 33
4.5. Tables 36
5. ENERGY INTAKE ADJUSTMENT IN DIETARY PATTERN ANALYSIS THROUGH LATENT CLASS MODELS 39
5.1. Introduction 39
5.2. Materials and methods 40
5.2.1. Study population 40
5.2.2. Dietary intake assessment 40
5.2.3. Statistical methods 40
5.3. Results 42
5.4. Discussion 43
5.5. Tables 46
6. DIETARY PATTERNS INSPECTION THROUGH LATENT CLASS TREE: AN APPLICATION TO MULTICENTRIC CASE-CONTROL STUDIES ON SELECTED DIGESTIVE TRACT CANCER 51
6.1. Introduction 51
6.1.1. Materials and methods 51
6.1.2. Study population 51
6.1.3. Dietary intake assessment 52
6.1.4. Covariate adjustment and local dependency inspection 52
6.1.5. Latent Class Analysis solution inspection 53
6.1.6. Latent Class Tree model 53
6.1.7. Choice of the number of starting classes in LCT 53
6.1.8. Model interpretation and fit statistics 54
6.1.9. Assessment of the association between dietary patterns and the risk of selected digestive tract cancer: a 3 Step analysis 54
6.2. Results 54
6.2.1. Covariate adjustment and local dependency inspection 54
6.2.2. Latent Class Analysis solution inspection 55
6.2.3. Choice of the number of starting classes in LCT 55
6.2.4. Model interpretation and fit statistics 55
6.3. Discussion 57
6.4. Tables and Figures 60
7. CONCLUSIONS 71

7.1. General conclusions 71
7.1.1. Dietary patterns and the risk of oral and pharyngeal cancer 71
7.1.2. Energy intake adjustment in dietary pattern research using Latent Class Analysis 71
7.1.3. Dietary inspection through Latent Class Tree 72
7.2. Future works 72
REFERENCES 75
SUPPLEMENTARY MATERIALS 83

ABSTRACT

INTRODUCTION

The dietary pattern approach is useful to study the effect of the overall diet on health outcomes, through considering the network of complex interactions between foods or nutrients. The main methods traditionally used to identify dietary patterns are principal components analysis, factor analysis, principal components factor analysis and cluster analysis.

Latent class analysis (LCA) is a latent variable approach, that has some advantages in comparison to the previous methods. Unlike principal component, factor and principal component factor analysis, it can be used to classify individuals into mutually exclusive groups conceived as dietary patterns and differently from cluster analysis, which has the same aim of grouping subjects, it permits quantification of the uncertainty of class membership, and assessment of goodness of fit. Moreover, it allows for adjustment for covariates directly in the pattern identification.

OBJECTIVES

As latent class analysis has rarely been applied in dietary pattern studies, the aim of this research is to apply the recent developments of the techniques to this area of research. We aimed to address the issue of dietary pattern identification in the case-control setting using latent class analysis and latent class trees. We provided estimation of pattern sizes and their characterization, taking into account correlations between dietary variables (local dependencies), and covariate adjustment. We also evaluated the robustness of the identified dietary patterns to total non-alcoholic energy intake adjustment, for different types of correction. Finally, we illustrated the method's properties in the assessment of the relation between the identified dietary patterns and selected health outcomes, given the all the above.

DIETARY PATTERNS AND THE RISK OF ORAL AND PHARYNGEAL CANCER

We analyzed data from an Italian multicentric case-control study on oral and pharyngeal cancer (OPC) carried out between 1992 and 2009, including 946 cases and 2492 hospital controls. Information on diet was collected through a food frequency questionnaire (FFQ). Using LCA, we found 4 dietary patterns, conceived as mutually exclusive groups of people who shared a common dietary behaviour within groups. The first pattern, labelled 'Prudent pattern', showed higher probability of consuming more leafy and fruiting vegetables, citrus fruit and all other kinds of fruits, tea while showing lower probability of consuming red meat. The second pattern, that we named 'Western pattern', reported higher consumption of red meat and lower consumption of fruits, cruciferous and fruiting vegetables. We termed the third pattern 'Lower consumers-combination pattern' as people in it were less likely to eat fruits, leafy and fruiting vegetables, pulses, potatoes, fish, white and red meat, bread and tea/decaffeinated coffee. The last pattern had higher probability to eating fruiting, leafy and other vegetables, white and red meat and bread, while showed a lower probability to consume coffee, tea, processed meat, cheese, fish, sugary drinks and desserts. We called this last pattern 'Higher consumers-combination pattern'. Dietary patterns were adjusted for total non-alcoholic energy intake and correlation between certain foods item (sugar-coffee, soups-pulses) was allowed during classes identification. Compared to the Prudent pattern, the Western and the Lower consumers-combination ones were positively related to the risk of OPC (OR=2.56, 95% CI: 1.90 –

3.45 and OR=2.23, 95% CI: 1.64 – 3.02). Higher consumers-combination pattern didn't differ significantly from the Prudent pattern (OR=1.28, 95% CI: 0.92 – 1.77).

ENERGY INTAKE ADJUSTMENT IN DIETARY PATTERN RESEARCH USING LATENT CLASS ANALYSIS

Using data from the same multicentric case-control study on OPC (Italy, 1992-2009), we identified and compared dietary patterns adjusting or not for total non-alcoholic energy intake in the classes identification phase of the analysis. Three possible ways to correct for total energy intake in class identification were presented, corresponding to different hypothesis on the effect of this variable. In general unadjusted and adjusted solutions were comparable. The main difference was related to the patterns that showed highest/lowest non-alcoholic energy intake, that resulted in a variation of number of classes (4/5/7 patterns for the different adjusted solutions and 5 patterns for the unadjusted one).

Then, to determine the effect of adjustment in predicting an health outcome, we compared the effect of unadjusted dietary patterns, unadjusted dietary patterns with non-alcoholic energy intake variable also included in the model as a confounder, and adjusted dietary patterns on the risk of OPC. Differences in the estimations for the distinct solutions were found when ORs were not corrected for known/potential risk factors. In general, adjustments for non-alcoholic energy intake results in a mitigation of the effects, thus remaining in the same order. When adjusting for known/potential risk factors, estimations of ORs and related CIs remained consistent in all the models we fitted.

In the end, specific suggestions on how to perform energy correction in dietary patterns research using LCA were delivered, basing on the results of the current analysis.

DIETARY PATTERNS INSPECTION THROUGH LATENT CLASS TREE

We analyzed data from two Italian case–control studies, the first included 946 cases with OPC and 2492 hospital controls, and the second included 304 cases with squamous cell carcinoma of the esophagus (ESCC) and 743 hospital controls. In our application of latent class analysis on the combined dataset of the two studies (Italy, 1992-2009), we found the best fit for a solution that was difficult to interpret and included minor differences between clusters. To address these issues, the Latent Class Tree method was proposed. Three fit statistics (AIC, AIC3, BIC) were used for their different level of penalty that resulted in different lengths of the tree. For the first split we allowed for a 4-class solution which identified a pattern characterized by high intake of leafy and fruiting vegetable and fruits ('Prudent pattern'), a pattern with a high intake of red meat and low intake of certain fruits and vegetables ('Western pattern') and two patterns which showed a combination-type of diet. The first 'combination' pattern showed a low intake of the majority of foods ('Lower consumers-combination pattern'), and the other one high intake of various foods ('Higher consumers-combination pattern'). Compared to the Prudent pattern, the Western one was positively related to OPC (OR=1.91, 95% CI: 1.41-2.58) and to ESCC (OR=3.22, 95% CI: 1.78 – 5.82). The Lower consumers-combination pattern was positively associated to OPC (OR=2.14, 95% CI: 1.58-2.91) and to ESCC (OR=2.85, 95% CI: 1.47-5.55). No significant association was found between the Higher consumers-combination pattern and OPC (1.04, 95% CI: 0.74-1.46) and ESCC (OR=0.89, 95% CI: 0.39-1.99). In the 'Prudent pattern' branch of the tree, at the third level, we found two classes that differed in the risk of both cancer types. These two classes differed mainly for the intake of citrus fruit, showing respectively, OR=1.85, 95% CI:1.07-3.19 for OPC and OR=5.37, 95% CI: 1.48-19.44 for ESCC for the class that reported low intake of citrus fruit with respect to the class which exhibit a high intake of citrus fruit. No other significant differences were found between the other pairs of classes at any other level of the tree.

CONCLUSION

We presented latent class methods as powerful tools to determine dietary patterns conceived as mutually exclusive homogeneous groups of subjects which shared common dietary habits. These methods exhibit some advantages, with respect to classical approaches, that can address important issues in dietary pattern research. For example, it is possible to obtain estimation for pattern prevalence in the population, and to perform energy intake adjustment in the pattern identification phase of the analysis. Moreover, class formation inspection, comparison between different solutions and the analysis of subgroups that may be relevant for the research at hand are features offered by the newly developed latent class tree approach.

ACRONYMS

CA	Cluster Analysis
CI	Confidence interval
FA	Factor Analysis
LC	Latent Class
LCA	Latent Class Analysis
LCT	Latent Class Tree
PCA	Principal Components Analysis
PCFA	Principal Components Factor Analysis
ESCC	esophageal squamous cell cancer
OPC	oral/pharyngeal cancer
OR	Odds Ratio
NAE	non-alcoholic energy intake

1. INTRODUCTION

1.1. Background

The study of the association between dietary habits and disease has usually been addressed focusing on single foods or nutrients as exposures. Nevertheless, people's diet consists in a variety of foods eaten together, which provide a complex mixture of nutrients that are likely to have additive or interactive effect on health.

In the last years, defining dietary patterns to represent the combined effect of all foods/nutrients consumed has become increasingly important. Aiming to catch the whole diet effect, dietary pattern analysis can describe the ways in which dietary variables are combined in actual diets and can account for the complex interactions among foods or nutrients. Being a more realistic picture of what individuals eat, they may be more powerful in predicting disease risk. Dietary patterns are also useful in summarizing confounding by diet [5]. Finally, patterns of diet intake are also more easy for the public to interpret and to translate into guidelines, and they can be helpful in evaluating the effect of dietary practices and adherence to dietary guidelines.

Three general approaches have been used to define dietary patterns: *a posteriori* empirical methods, *a priori* hypothesis-oriented methods and approaches which combine characteristics of the two previous methods.

A posteriori dietary patterns have been commonly derived using principal components(PCA), factor(FA) or cluster analysis(CA). However, these methods take alternative approaches to addressing the issue.

FA examines the correlation matrix of dietary variables and search for underlying traits (factors) that explain most of the variation in the data. Commonly, in FA the emerging factors are modified by using an axis rotation. In PCA, a large number of correlated dietary variables are reduced to a smaller set of uncorrelated variables that are called components and capture the major dietary traits in the studies population. For each factor/component, scores are obtained that define the position of each individual along a gradient.

CA aims to uncovering or discovering groups or clusters of observations that are homogeneous and separated from other groups [108]. These techniques have the goal of grouping similar observations into a number of clusters based on the observed values of several dietary variables collected for each individual.

These methods can use either foods or nutrients as input variables. The data collection on consumption of foods is often reduced by combining foods into 20-40 nutritionally similar groups. Moreover, dietary variables may be transformed to obtain a normal distribution or adjusted for energy intake. At the present, there is no standard or clear advantage among these various approaches [1,5].

1.2. Problem statement

One of the main objectives of dietary pattern identification is to find dietary habits that may be related to specific diseases. One possible way to target this goal regards the classification of the population in mutually exclusive eating groups, characterized by similar diet, and evaluate and compare their association with specific health outcomes.

Both PCA/FA and CA can be used to target this approach, however standard techniques have some limits.

Methods like PCA/FA do not group subjects, but dietary variables (foods/nutrients). Then, individuals get a score for each diet component/factor. When the aim is to estimate patterns' prevalence or risk of disease for one group of subjects compared with another group, an additional step of cross-classification of the dimensions is necessary. This requires stronger subjective decisions as the number of dimensions get larger. Moreover, while FA estimation can rely on a parametric approach, this is generally not true for PCA.

CA, instead, aims to classify individuals in mutually exclusive dietary patterns such that within the same groups, individuals share a similar food intake. The major limit of this approach, is that it mainly relies on non-parametric techniques. Another limitation is that classification uncertainty is assumed to be 0.

Another approach to identify mutually exclusive dietary groups is to apply consequently the above mentioned methods: first PCA/FA helps explain which foods/nutrients are eaten in combination, then CA helps classify individuals. Despite this approach gives interesting insights to dietary patterns, presenting a double perspective, it has the disadvantage of the application of two methods carrying with them their respective limitation.

Finally, as Fahey [2] pointed out, the research regarding dietary patterns has taken little effort in adapting statistical methods for pattern identification, so all the traditional methods like those mentioned above lack in some extensions and generalizations that are now available to address important issues in the study of the association between dietary habits and disease.

1.3. Research Purpose

In the last decades, Latent Class Analysis (LCA) has become popular in social and behavioural research. LCA allows to identify unobserved homogeneous groups in a population based on subject's responses on a set of observed, often categorical variables. The basic assumption of the traditional formulation is that the latent categorical variable identifies K latent classes/groups in the population and the set of observed categorical variables are its indicators. The second traditional assumption is that of local independence, implying that indicators are statistically independent given the latent variable.

The increasing popularity that LCA gained, especially in the last decades, led to many extensions of the traditional model and to software availability.

LCA has not been used in dietary pattern studies as the previous traditional methods, but it has some advantages in comparison to them. Unlike PCA/PCFA/FA, it can be used to classify individuals into mutually exclusive groups/dietary patterns and differently from CA, it permits the quantification of the uncertainty of class membership, and the assessment of goodness of fit. Moreover, it allows for adjustment for covariates and for correlation between food items directly in the pattern identification. All these features can be applied with important implications for dietary patterning, addressing issues that are relevant in this field.

As LCA has rarely been applied in dietary pattern research, most of the new developments of techniques, as well as new implementations in statistical software, have seldom been applied to this area of research.

Therefore, in summary, the purpose of this research is to investigate a latent class solution to the following issues:

1. identification of dietary patterns in the case-control setting using food groups as dietary indicators. Estimation of pattern sizes and characterization of the patterns, taking to account for correlations between food items (local dependencies), and covariate adjustment;
2. evaluation of the robustness of the identified dietary patterns to total non-alcoholic energy intake adjustment, for different types of correction;
3. application of a new LC approach aimed to help the interpretation of classes in complex situations, on a database from multiple case control studies;
4. assessment of the relation between the identified dietary patterns and selected health outcomes, given all of the above.

1.4. Outline of the thesis

The following part of this thesis is structured as follows.

Chapter 2 gives an overview on empirically or *a posteriori* dietary patterns methods and research. Comparisons between LCA and classical methods are made. Differences between the current research and the previous released publications on dietary patterns using the same data from the multiple case-controls studies [3-4] are also highlighted.

Chapter 3 describes the methods behind this study. LCA in the basic formulation and with the extensions used in this thesis is presented. An illustration on how to relate dietary patterns identified through LCA and a specific health outcome using a procedure in three steps (3 Step analysis) is given. Latent class tree, a recent development of LCA, is also defined and presented.

Chapter 4 targets dietary pattern identification with LCA and presents the analysis of their association with the chosen health outcome, using data from an Italian multicentric case-control study.

In Chapter 5 a contribution on the issue of energy adjustment in dietary patterning with LCA is made. Different types of correction are compared and robustness of dietary patterns to total non-alcoholic energy intake adjustment is assessed.

Chapter 6, develops the Latent Class Tree solution for class identification and inspection in dietary patterning, using data from two Italian multicentric case-control studies. It is shown how to inspect class formation and compare different LCA solutions, and how to allow for different granularity in the analysis of the association between dietary patterns and the risk of the selected health outcomes.

In Chapter 7 final conclusions, remarks and future possible developments are presented.

2. DIETARY PATTERN ANALYSIS THROUGH LCA

2.1. Introduction

Latent Class Analysis is a methodology developed in the framework of social and behavioural sciences to detect unobservable homogeneous subgroups in a population. Among the methods used to empirically derive dietary patterns (*a posteriori* methods), LCA has rarely been applied.

In the majorities of the studies on dietary patterning through LCA, the basic traditional method has been applied [6-21]. The two basic assumptions of the traditional LCA are that the population consists of K mutually exclusives latent classes/groups and the observed categorical variables, that are indicators of the latent one, are mutually independent conditional on the latent variable.

Many of these studies were descriptive and didn't relate the identified dietary patterns with any health outcomes. Whereas this kind of association was assessed, it was mostly done through cross-tabulation or regression without (at least explicitly) correcting for bias (Bolck, Croon and Hagenaars) [55-56,110].

Most of the studies applied directly LCA on foods/nutrients items, while few of them applied LCA on the factor scores derived by *a posteriori* FA[22-23] or on subjects' scores on the adherence to certain dietary habits, often defined by an Index [24-26].

In the last years, more attention has been given to some extensions of LC models. Some studies on dietary patterns, for example, applied latent class trajectory or transition analysis [27-30]. The most important development in LCA was it contextualization in the framework of finite mixture models. This principally allowed the analysis of indicators of different scales and permitted different assumptions on their distributions. Recent publications on dietary patterns used this extended methodology [2,31-36].

2.2. Identifying dietary patterns: a comparison of methodologies

In recent years, epidemiologists addressing dietary patterns research have adopted different multivariate techniques able to cope with the simultaneous analysis of various dietary variables. Among exploratory (or *a posteriori*) methods, which empirically derive dietary patterns from the data, the most used techniques are Principal Component Analysis(PCA), Factor Analysis(FA) and Cluster Analysis(CA) [108-109].

Explaining these techniques in depth is beyond the purpose of this thesis. There are, in fact, a lot of different model specifications especially in the framework of FA and CA [57,108-109]. We focused here only on the principal applications that have been done in dietary patterns research, to highlight differences that are relevant in this field.

Principal Component Analysis is a data reduction technique with the aim of reducing the dimensionality of a multivariate data set while accounting for as much of the original variation as possible present in it [109]. This is done by transforming the original dietary variables into a new set of variables, the principal components, that are linear combinations of them, uncorrelated and ordered so that the first few of them account for most of the variation in all the original data. An individual score on each principal component is then derived and it can be used to assess the relationship with health outcomes of interest.

Factor analysis is a multivariate method which aims to identify underlying latent dimensions (factors), of food/nutrient consumption, by aggregating dietary items on the bases of the degree to which they correlate with each other in the dataset. Like in PCA, an individual score on each factor can be used to assess the relationship with any health outcomes [57,108-109].

The term Cluster analysis covers a wide range of techniques with the aim of discovering groups or clusters, that are relatively homogeneous in terms of dietary habits and separated from other groups. The most famous clustering techniques are the k-means and the hierarchical agglomerative technique [57,108-109]. The classification obtained can then be used to compare groups in terms of the association with a health outcome.

In the following paragraphs the major differences between these methods and Latent Class Analysis (LCA) will be presented.

2.2.1. Dietary pattern definition and interpretation

The first important difference is how these approaches define the dietary patterns identified. These definitions follow directly from how the methods cope with reduction and grouping.

Conceptually, LCA and CA are subject-centred techniques that focus on similarities and differences among subjects on the basis of responses to items and try to identify homogeneous groups of subjects characterized by similar dietary behaviour that differentiate from the other subgroups.

On the other hand, FA and PCA are feature-centred, concerned with the structure of variables (food or nutrient items). PCA attempts to 'group' dietary variables in combinations that are representative of the original features of the dataset. In FA, instead, the emphasis is on a transformation from the underlying factors to the observed data. Therefore, the two techniques do not have the aim to identify clusters or groups of people.

This has a immediate consequence on the meaning of 'dietary patterns' identified by the methods. Dietary patterns identified with FA and PCA are dimensions based on combinations of dietary variables, while dietary patterns identified with LCA or CA are groups of individuals which share a common dietary behaviour.

Regarding interpretation, dietary patterns identified through CA are described and labelled through the distribution of dietary variables within clusters. Higher values of intake of certain dietary variables define a positive attitude of the cluster for them while lower values define an avoidance of those foods/nutrients.

In contrast, in FA/PCA dietary patterns correspond to factors/components, and the interpretation is done thorough factor/principal component loadings. A factor/principal component loading of 0 represents no relation between the dietary item and the latent factor/principal component, whereas factor/principal component loadings closer to -1 and 1 represent stronger relations.

In LCA the description of classes (or groups) is done according to the conditional distribution of foods/nutrients intake giving the latent classes (class-specific response probabilities). That is, a very high or low probability indicates almost all or almost none of the class members giving a certain response.

2.2.2. Latent variables

Another trait that discriminates among the mentioned techniques is the postulation of the existence of latent variables. Latent variables are variables non directly measurable, but indirectly identifiable by using

observed variables as indicators. A latent variable and its observed indicators make up a measurement model.

In LCA the measurement model is composed by a categorical latent variables and its indicators, while in FA the measurement model is composed by continuous latent variables and their indicators.

These issues, along with the fact that both approaches are based on the covariance structure of the data, often led to the consideration of LCA as the 'categorical' counterpart of FA.

On the contrary, CA and PCA do not posit the existence of a latent variable that accounts for any association between observed indicators.

2.2.3. Statistical model

An important feature of LCA is that it consist in a model based approach, hence a statistical model is postulated for the population where the sample belongs. Maximum likelihood estimation is used to obtain the parameters in LCA.

FA provides different methods for parameters estimation that can be parametric (maximum likelihood factor analysis) or non-parametric (principal factor analysis or principal axis factoring) [57, 108-109]. Both types of estimation are currently used in dietary pattern research.

In dietary patterning, mostly commonly used CA techniques mainly rely on non-parametric methods, such as the k-means clustering, or the agglomerative hierarchical techniques [109].

Although there are inferential methods for using the sample principal components derived from a random sample of individuals from some population to test hypotheses about population principal components, they are very rarely seen in the literature in general [109]. In summary, PCA is a data reduction technique based on linear transformation of the original variables aiming to help to understand the observed data set whether or not this is actually a 'sample' in any real sense [109].

An advantage of the techniques using a statistical model is that the choice of the clustering criterion is less arbitrary and formal tests can be used to assess parameters and goodness of fit [111].

2.2.4. Indicators

A CA can be performed on categorical, ordinal or continuous indicators but with some remarks and limitations. The available cluster analysis algorithms all depend on the concept of measuring the distance (or some other measure of similarity) between the different observations we're trying to cluster. If one of the variables is measured on a different or much larger scale than the other variables, then whatever measure we use will be overly influenced by that variable. Hierarchical agglomerative methods can deal with categorical, ordinal and continuous variables but a crucial caution to be used regards the choice of the appropriate measure of distance in accordance to the scale of the observed variables. This aspect has a strong impact on the results of the analysis and working with different scaled indicators can be troublesome. The k-means method is based on Euclidean distance that is proper for at least indicators at interval level of the measurement and it's not a scale-invariant method. Other model specifications can deal with also other types of indicators. K-medoids method, for example, uses a measure of dissimilarity instead of Euclidean distance, but the scale of variables is anyway an issue in cluster analysis.

FA and PCA require that variables are linear in nature, as the methods provide a linear function of the variables. Indicators must be at least at interval level of measurement, as nominal and ordinal items are not a appropriate for a FA/PCA.

LCA can be performed with observed indicators of different scale types (nominal, ordinal or continuous) or a combination of these. The combination of differently scaled variables leads to the finite mixture models, that are the generalization of LCA. Instead, if the variables are all categorical we obtain the traditional LC model. Another important feature of LCA is that it's scale invariant.

FA and the basic formulation of LCA assume local independence. This assumption means that indicators are independent after controlling for the latent variable. In FA the violation of this assumption may lead to additional and spurious factors needed to obtain a good fit. Similarly, in LCA it may lead to additional classes. An important strength of LCA is a further development which allows for correlated errors between dietary variables.

2.2.5. Inclusion of external variables

When researches want to build classes/dimensions that are independent from certain variables, the only way to deal with this issues for the majorities of the mentioned methods is the inclusion of the variable in classes/dimensions identification together with the dietary variables. With the traditional methods, if an external variable (e.g. a confounder) is included in the model with the dietary indicators, it will influence the formation of the classes and would, in essence, become an indicator itself. Therefore, from the theoretical point of view, this approach does not completely fulfill the objective.

In LCA the definition of the probability structure, which describes the relevant set of dependence assumptions among the variables in the model, allows to specify the relation with external variables and to distinguish between covariates and distal outcomes. Covariates are conceived as external variables influencing the classes, while distal outcomes are external variables affected by the classes.

LCA model can be defined by a measurement part and a structural part. The measurement part establishes the relation between a block of manifest indicators and its latent variable. The structural part defines the relation between external variables and the latent variable. When dealing with covariates, it is possible to distinguish between proper indicators and the external variable and at the same time permitting both the measurement part and the structural part of the model to be performed simultaneously using a single ML estimation algorithm. Therefore, when dealing with confounders, differently from the traditional methods, LCA can be easily extended to include exogenous variables that affect latent classes as covariates.

The issue of relating with external variables considered not covariates but distal outcomes will be presented in Chapter 3, Par.3.3.

2.2.6. Classification of subjects

The three methods differ substantively in how they classify subjects. In CA, people are assigned to classes directly as part of the pattern identification process. For example, in a hierarchical agglomerative approach subjects are linked looking at all possible pairs of cases and linking those in the pair with the smallest distance, continuing the process until all cases belong in one big cluster. When homogeneity measures exhibit a large drop in value the classes are defined. Hence, a disadvantage of CA is that subjects are assigned to one pattern with a probability of 1 and to all others with a probability of 0. Therefore it assume no classification uncertainty.

As we've seen previously, FA and PCA do not provide a general classification of subjects, factor/principal component scores in fact are derived for each factor/principal component separately. If the researcher is interested in classifying the individuals basing on factor/principal component scores, subjective decisions have to be taken. In fact, when there are only 2 factors/principal components a cross-tabulation of the factor/principal component scores' quantiles is an easy way to proceed, but when they are more than two it could be difficult to collapse into mutually exclusive groups without making strong decisions.

LCA does not automatically assign subjects to clusters like CA, using a probability based classification instead. LCA classifies subjects into clusters using model based posterior membership probabilities. This approach yields ML estimates for misclassification rates. Moreover, this approach avoids bias in estimating cluster specific-means as individuals contribute to the means of clusters with a weight equal to the posterior membership probability for each clusters. Popular options for probability based classification are the proportional or the modal assignments (see Chapter 3, par.3.3).

2.3. Dietary patterns and the risk of oral/pharyngeal and esophageal cancer

To our knowledge, no study relating dietary patterns derived through LCA and oral/pharyngeal or esophageal cancer has ever been performed. With regards to traditional *a posteriori* methods, the majority of the studies which assessed the relation between dietary patterns and the risk of these two types of cancer performed PCA, FA or Principal Component Factor Analysis(PCFA, see par. 2.3.2) [37-54].

The data we analyzed in this work comes from a network of case-controls studies on different neoplasms conducted in Italy. Previous studies on dietary patterns and the risk of certain types of cancer have been already performed on these data, but using different approaches. In particular, two previous works regarding dietary patterns and the risk of cancer of the oral/pharyngeal cancer study (using a subset of the data collected between 1992 and 2005) and on the data on esophageal cancer study [3-4] were performed.

In the following paragraphs we aimed to compare their approach with our proposal, showing differences and explaining what our research adds to the results obtained in those studies.

2.3.1. The network of case-control studies

Between 1991 and 2009, a series of hospital-bases case-control studies on different neoplasms were carried out in various areas of northern (the greater Milan area, the provinces of Pordenone, Padua, Udine, Gorizia and Forlì, and the urban area of Genoa), central (the provinces of Rome and Latina), and southern (the urban area of Naples) Italy, and the Canton of Vaud, Switzerland.

All studies included incident cancer cases (diagnosed within 1 year before inclusion in the study), admitted to major hospitals in the study areas. Controls were subjects admitted to the same hospital networks in the same period for acute, non-neoplastic conditions, unrelated to known and potential risk factors for the concerned cancer site.

The first database we analyzed in this work belongs to the case control study on oral/pharyngeal cancer conducted in Italy, between 1992 and 2009 and included 946 and 2492 controls. The second database belongs to the case-control study on esophageal cancer, conduced in Italy from 1992 to 1997, included 304 cases and 743 controls.

2.3.2. Dietary pattern and the risk of oral/pharyngeal and esophageal cancer: a comparison of the works on the multicentric case-control studies, Italy

Two previous works were performed regarding dietary patterns and the risk of cancer on a subset of the data of the oral/pharyngeal cancer study (data collected between 1992 and 2005) and on the data on esophageal cancer study [3-4].These two studies, and all the studies related to dietary patterns and the risk of specific cancers conducted on the multicentric case-control studies mentioned above, followed the approach recommended by R. Johnson and D. Wichern [57] with regards to FA.

Johnson and Wichern [57] introduced Principal Component Factor Analysis (PCFA) as a proper FA, which uses PCA for parameters estimation. Therefore, even though the dimension extraction is done through PCA, authors framed this method in the FA approach, with its aims and assumption.

Exploratory PCFA was performed on the correlation matrix of selected macro and micro nutrients to identify a few unobservable factors conceived as dietary patterns. Prior to the analysis, original nutrients intakes were standardized. Number of factors to be included in the analysis was chosen according to factor eigenvalues>1, scree-plot construction and factor interpretability. Varimax rotation was performed to the factor loading matrix and nutrients with an absolute rotated factor loading higher or equal to a certain threshold on a given factor were used to label the dietary patterns. Regarding risk estimates, participants were grouped into categories according to quantiles of factor scores among the controls, for each factor. Odds ratios and related 95% confidence intervals for each quartile category were estimated using multiple logistic regression models.

Despite formally aiming both at identifying dietary patterns and relating them to cancer risk, those studies and the analyses that we presented here are theoretically and structurally different.

The first important difference regards the type of the indicators on which the dietary pattern identification process is carried out. This choice has a direct effect on the final aim of the analysis.

The previous analyses were performed using nutrients as indicators. The primary advantage of this approach is that information can be directly related to the fundamental knowledge of biology. In epidemiologic studies, the use of nutrient intake can be powerful in hypothesis testing, especially when single foods alone contribute moderately to that nutrient intake. In summary this approach can be conceived as clinical-biological oriented.

The analysis presented in this work used food groups as indicators instead. This kind of analysis is generally most directly related to dietary recommendations, because subjects can modify their nutrient intake primarily by their choice of foods. Moreover, as foods are complex combinations of nutrients that together may compete, antagonize or interact, it is not possible to predict the effect of a certain food based on the content of a specific nutrient. Finally, dietary recommendations on food consumption can be made also without knowing its beneficial/harmful effect. For example, the positive effect of certain vegetables on the reduction of some diseases has been observed, yet without knowing which combination of nutrients is important. Then, as Mertz [107] pointed out, foods are not fully represented by their nutrients composition. This approach can be conceived more as public health oriented.

The second main difference comes from the method applied to identify dietary patterns. This choice results in a different definition of dietary patterns. The previous studies applied PCFA, that consists in FA where the parameters estimation is done through principal components, according to Johnson and Wichern [57]. A FA with principal axis factoring estimation was also performed to assess the previous solution. In

summary, the whole approach chosen in the previous studies belongs to the FA framework. Specific differences between this method and LCA were described in the previous paragraph, and the principal consequence of this choice is a substantive difference in dietary pattern definition. Dietary patterns derived through FA describe combination of foods/nutrients that are eaten together, while dietary patterns identified through LCA describe groups of people with similar dietary habits. These two approaches, in the end, can provide two different perspectives for understanding and describing dietary habits.

As a consequence of the methodological choice and its implications, the third important difference regards the assessment of the relation between the dietary patterns derived and health outcomes. In the previous studies, participants were grouped into categories according to quantiles of factor scores among the controls, and risk estimation was done for each quantile category of the factors. Then the new variables were entered into the model separately and all together. This aspect results in a different question of research. With this approach, one wants to estimate the association between estimated dietary patterns and the disease, comparing low vs high adherence to a specific pattern. On the opposite, with LCA, dietary patterns are not food/nutrients combinations, but mutually exclusive group of people, and while assessing the effect of these patterns on cancer risk, the question of research regards the estimation of the risk for a group with a specific diet, compared to a reference one.

Finally, the studies previously delivered on nutrient dietary patterns and the risk of oral/pharyngeal and esophageal cancer addressed the issue of total energy intake adjustment in different ways. In the study on OPC [3], total non-alcoholic energy intake was taken into account by the inclusion of the related variable in dimension identification, as one dietary indicator. On the contrary, LCA can be easily extended to include confounders as covariates, keeping them separated from proper indicators while permitting both to be estimated using a single estimation algorithm. The study on ESCC [4], instead, was performed without correction for total non-alcoholic energy intake. In Chapter 5 we evaluated the robustness of the dietary patterns identified through LCA to energy adjustment, by comparing unadjusted dietary patterns and different types of corrections allowed by this method.

For all these reasons, the two approaches cannot be seen just a replication of a research with different methodologies, but results in two different perspectives that combined can give a broader insight on the effect of the diet on the risk of cancer.

3. METHODS

3.1. Introduction

In the last decades, Latent Class Analysis has become a popular method among social and behavioural researchers aiming to cluster subjects basing on their answers on a set of observed variables. The resulted classes represents unobserved homogeneous groups, that can be interpreted substantively basing on the conditional response probabilities within a class.

Lazarsfeld [58] introduced it in 1950 as a clustering method for dichotomous survey items. In 1974, Goodman [59] formalized the methodology and extended it to nominal variables, solving some identification issues and developing an algorithm to obtain maximum likelihood estimates, that is still the dominant approach used for parameters estimation and it is known as EM algorithm. In 1979 Haberman [60] showed how the model can be specified as a log-linear model for the contingency table derived from the cross-tabulation of the latent and observed variables.

Many important extensions of the classical LC model have been proposed since then, such as the inclusion of covariates, local dependencies, ordinal/continuous indicators, several latent variables, and repeated measures. A general framework for categorical data analysis with discrete latent variables was proposed by Hagenaars [61] and extended by Vermunt [62].

In the following paragraphs we introduced the LC model specification and applications that are relevant in this thesis.

3.2. Latent Class Analysis

3.2.1. Basic LCA

In the Latent Class models, we have T observable response variables or indicators, denoted by y_{it} ($i = 1 \dots n, t = 1..T$) and a single categorical latent variable x, with K categories.

The general mixture model probability structure that defines the relationships between the latent variable and the indicators is the following:

$$f(y_i)=\sum_{x=1}^{K} P(x) \prod_{t=1}^{T} f(y_{it}|x) \quad (1)$$

Depending on the scales of the indicators, a particular distribution is assumed for y_{it}. In case of categorical indicators, a multinomial distribution is assumed for y_{it} with M_t entry.

Therefore, the distribution for each y_{it} is of the form:

$$P(y_{it} = m|x) = \pi_{m|t,x} = \frac{\exp(\eta^t_{m|x})}{\sum_{m'} \exp(\eta^t_{m'|x})}$$

Here, $\pi_{m|t,x}$ is the probability of giving response m, given latent class membership as indicated by x, and $\eta^t_{m|x}$ is the linear term that can be further restricted by a regression model, yielding a multinomial logistic regression:

$$\eta_{m|x}^{t} = \beta_{m0}^{t} + \beta_{mx0}^{t}$$

3.2.2. Extensions of traditional LCA

An important extension of the classical LC model is the possibility of including covariates. Therefore the (1), with categorical indicators, can be extended in the following ways, depending on the assumptions about the effect of the covariate z_i:

$$P(y_i|z_i)=\sum_{x=1}^{K} P(x|z_i) \prod_{h=1}^{T} P(y_{it}|x)$$

when the covariate affects the latent variable but have no direct effects on the indicators;

$$P(y_i|z_i)=\sum_{x=1}^{K} P(x) \prod_{h=1}^{T} P(y_{it}|x, z_i)$$

when the covariate is assumed to affect only the indicators;

$$P(y_i|z_i)=\sum_{x=1}^{K} P(x|z_i) \prod_{h=1}^{T} P(y_{it}|x, z_i)$$

when the covariate affects both the latent variable and the indicators.

In presence of a covariate affecting the indicators, the single indicator distribution becomes:

$$P(y_{it} = m|x, z_i) = \pi_{m|t,x} = \frac{\exp\,(\eta_{m|x,z_i}^{t})}{\sum_{m'} \exp\,(\eta_{m'|x,z_i}^{t})}$$

Like categorical indicators, the values of the latent variable are assumed to come from a multinomial distribution.In presence of a covariate affecting the latent variable, the multinomial probability $P(x|z_i)$ is parameterized as follows:

$$P(x|z_i) = \pi_{x|z_i} = \frac{\exp\,(\eta_{x|z_i})}{\sum_{x'}^{K} \exp\,(\eta_{x'|z_i})}$$

Multinomial logit models for the latent classes and the single indicators are therefore modified with the inclusion is the term z_i and related parameter.

Local independence is the basic assumption of the standard LC model, that implies that indicators are mutually independent given the latent class. The standard model can be extended to relax this string assumption, sometimes unrealistic in practical application, that can results in lack of fit in presence of its violation.

We can define H subset of the T indicators. We use the symbol y_{ih} to denote one of the H subset of y_{it}. The y 's belonging to the same set h may be correlated within latent classes. In presence of local dependencies, the (1) becomes:

$$f(y_i)=\sum_{x=1}^{K} P(x) \prod_{h=1}^{H} f(y_{ih}|x)$$

3.2.3. Fitting LCA

In LCA the number of classes is determined by fitting first the trivial 1-class model, where all the individuals belong to the same class and then increasing number of classes as long as some fit measures improves. Some fit statistics which aim to balance model fit with parsimony [105-106], are defined as follow:

$$BIC = -2\log L + \log(n)P$$

$$AIC3 = -2\log L + 3P$$

$$AIC = -2\log L + 2P$$

with P equal to the number of parameters in the model and n equal to the sample size.

The identified classes are characterized by their class proportions and their response probabilities for all the observed indicators. Labelling and interpretation of the classes are done by inspecting these conditional class-response probabilities.

3.3. Three-Step analysis

The identification of homogeneous sub-groups in a population is usually the first step in Latent Class Analysis, as researchers are often interested in how the groups affect one or more outcome of interest (distal outcomes).

With standard LCA, the relation between classes and distal outcomes of interest can be assessed by two different procedure.

The one-step procedure consists in including the external variable in the model, performing simultaneous estimation of the measurement part of the model with a logistic regression in which the latent classes are related to it (structural part of the model).

The second option is a three-step approach which consists in the following steps:

STEP 1. a LC model is build for a set of response variables.

STEP 2. subjects are assigned to LCs based on their posterior class probabilities that can be obtained from their observed response and the estimated parameters of the step 1 LC model. Possible classification methods are modal or proportional. Where modal assignment classifies respondents with a probability of 1 to the class with the highest posterior probability (i.e. the class someone most likely belongs to is the one they are classified into), proportional assignment uses the posterior probabilities as weights, whereby a person is classified into all classes with the respective probability of belonging to that class.

STEP 3. a standard regression is estimated using the step 2 class assignment and the observed external variable of interest.

Differently from covariates control (see Chapter 2, par.2.2.5), when the interest is assessing the effect of the latent variable on a specific outcome (distal outcome), this second approach is usually preferred for several reasons. First, in this case is preferable to separate the measurement part from the structural one. As the causal mechanism is specified from the latent variable to the distal outcome (opposite to what happen with covariates control), the external variable would become an indicator itself. Second, in the one-step approach the external variables are used in the formation of latent classes, while the goal is relating latent classes previously defined to an external outcome. Then, the three step approach is usually less affected by assumptions on the class-specific conditional distribution of the external variables.

The main disadvantage of the traditional three step approach was that it underestimated the relation between the external variables and the latent class membership. Recently, methods have been developed to adjust for this bias by Bolck, Croon and Hagenaars [55] and Vermunt [56] [110].

3.4. Latent Class Trees

In 2018, van den Bergh proposed a LC extension to help the interpretation of models when it is troublesome [66]. For example, when datasets are large (in terms of respondents or variables) the fit of the model usually improves until it contains a large number of classes, as many dependencies has to be taken into account. Moreover, the choice of the criterion (e.g. BIC or AIC) can lead to totally different solutions that are very hard to compare.

Therefore he proposed the Latent Class Tree as an alternative way to perform LCA. It consists in imposing a hierarchical structure on latent classes.

The procedure starts with the estimation of standard 1 class and 2 class LC models on the total sample (root node of the tree). If the 2 class model is preferred according to a certain fit measure (e.g. AIC or BIC), subjects are assigned to the two 'child' classes having the total sample as the 'parental' class. Subsequently, child nodes are treated as parental nodes. For each node, 1 and 2 class model are estimated and if a 2 class model is preferred, subjects are assigned to the new child classes. The same procedure is repeated until only 1 class models are preferred. The probability structure at each node, can be formulated as following:

$$P(y_i|x_{parent})=\sum_{k=1}^{K} P(x_{child}=k|x_{parent}) \prod_{h=1}^{H} P(y_{ih}|x_{child}=k, x_{parent})$$

where x_{parent} represents the specific parent class and x_{child} represents one of the Kchild classes (in general $K=2$).

$P(x_{child}=k|x_{parent})$ represents the class proportions and $P(y_{ih}|x_{child}=k, x_{parent})$ the class specific response probabilities for the class k at the node concerned.

When a split is accepted, the assignment of subjects to the new classes is done based on their posterior class membership probabilities, that are obtained as follows:

$$P\left(x_{child}=k|y_i, x_{parent}\right)=\frac{P(x_{child}=k|x_{parent}) \prod_{h=1}^{H} P(y_{ih}|x_{child}=k, x_{parent})}{P(y_i|x_{parent})}$$

Estimation of the LC model at a specific parental node, involves maximizing the following weighted log-likelihood function:

$$logL\left(\theta, y, x_{parent}\right)=\sum_{i=1}^{N} \log w_{i,x_{parent}} P(y_i|x_{parent})$$

where $w_{i,x_{parent}}$ is the weight for the person i at the parental class, that is equal to the posterior probability of being in that class. If this class is further split in two, the weights for the two child classes are obtained as follows:

$$w_{i,x_{child=1}}=w_{i,x_{parent}} P(x_{child}=1|y_i, x_{parent})$$

$$w_{i,x_{child=2}}=w_{i,x_{parent}} P(x_{child}=2|y_i, x_{parent})$$

Therefore, a weight at a particular node equals the weight at the parent node times the posterior probability of belonging to the child node concerned conditional on belonging to the parental node.

Inclusion of covariates and local dependence definition can be done as in standard LCA.

Three Step analysis can be performed as in standard LCA at each split of the tree, using the proper posterior class probabilities for each node.

3.5. Estimation

LC models are typically estimated by maximum likelihood(ML), which involves maximizing the following log-likelihood function:

$$logL(\theta, y) = \sum_{i=1}^{n} \log f(y_i, \theta)$$

when n denotes the total sample size and $f(y_i)$ takes the form defined in (1) for the general case.

It is possible to allow the utilization of prior distributions for the parameters of the model to prevent boundary solutions, resulting in a Bayesian procedure called posterior mode(PM) or maximum a posterior estimation(MAP) [67,112-113]. Given that we used categorical indicators, boundary problems, in our case, derive from multinomial probabilities that become 0. This problem can be circumvent by using Dirichlet priors for the latent variable and the conditional response probabilities.

Denoting the assumed priors for θ by $p(\theta)$ and the posterior by P, MAP estimation involves maximizing the following log-posterior function:

$$logP = \sum_{i=1}^{n} \log f(y_i, \theta) + \log p(\theta)$$

MAP estimation can be considered a form of penalized ML estimation, in which the term $p(\theta)$ penalizes solutions that are too close to boundary of the parameter space.

The use of a Dirichlet prior for the latent variable can be interpreted as adding pseudo-elements equally distributed among the classes (and the covariate patterns). The same prior is used for the categorical items and can be interpreted as adding pseudo-elements to the latent classes with preservation of the observed marginal item distribution in the models for indicators. We maintained the default value of Latent Gold program that prevent boundary estimates coming from cells exactly equal to 0. The default value can be interpret as adding 1/K pseudo-cases to the cells where K is equal to the number of latent classes. This choice with just a moderate sample size has a negligible effect on parameters estimation. Maximum likelihood and posterior mode estimation were compared in all the models fitted. In all the analyses presented in Chapter 4 and 5 and in the classification part of the analyses presented in Chapter 6 we found few differences and only related to the third/fourth decimal places (results not shown). In the 3 Step models presented in Chapter 6 boundary solutions were an issue, especially related to the smaller size of the ESSC case-control study database with respect to the OPC one in the combined analysis (results for ML estimation presented in Supplementary Materials). Performing a stronger penalization in this last analysis is also possible, resulting in increasing the weight allocated to the Dirichlet prior. Anyways, we chose to maintain the above defined penalization in all the analyses as a more conservative approach, preventing boundary estimates where present and not affecting estimation when the problem did not occur with respect to the classical maximum likelihood approach.

Maximization is typically done by means of the EM algorithm, alone or combined with Newton-Raphson algorithm. In this dissertation the combined algorithm present in Latent GOLD statistical software (Vermunt & Magidson,2016) was used [67].

4. DIETARY PATTERNS AND THE RISK OF ORAL AND PHARYNGEAL CANCER USING LATENT CLASS ANALYSIS

4.1. Introduction

Cancer of the oral cavity and pharynx (OPC hereafter) collectively ranks seventh for incidence and eighth for cancer mortality[68]. Tobacco smoking and excessive alcohol drinking are recognized as the two major risk factors for oral and pharynx cancer. Among other factors, diet has been suggested to play an important role. In particular, an inverse association between high intake of vegetable and fruits and a possible positive association between meat and OPC risk were found[3,69-84]. Most of the evidence came from studies focusing on single foods while the relationship between diet and oral and pharyngeal cancer has been less frequently addressed considering dietary patterns.

The dietary pattern approach is useful to study the effect of the overall diet on health outcome, through considering the network of complex interactions between foods or nutrients. The main methods traditionally used to identify dietary patterns are principal components analysis (PCA), factor analysis (FA), principal components factor analysis (PCFA) and cluster analysis (CA). With regard to a posteriori dietary patterns, association between diet and OPC has been traditionally assessed by PCA and PCFA [3,37-40,44,79,85-86].

Latent Class Analysis (LCA) is a latent variable model, which has some advantages in comparison to the previous methods. Unlike PCA/PCFA/FA, it can be used to classify individuals into mutually exclusive groups/dietary patterns and differently from CA, which has the same aim of grouping subjects, it permits quantification of the uncertainty of class membership, and assessment of goodness of fit. Moreover, it allows for adjustment for covariates directly in the pattern identification.

The aim of this study is to identify dietary patterns through LCA to add a new perspective on the evidence about the association between dietary habits and OPC in Italy.

4.2. Materials and methods

4.2.1. Study population

We use data from a multicentric case-control study on OPC carried out between 1992 and 2009, in the greater Milan area (northern Italy), the provinces of Pordenone (North-East Italy), Rome and Latina (Central Italy). The study included 946 patients (756 men, and 190 women; median age 58 years, range 22–79 years) admitted to major hospitals in the study areas with incident, histologically confirmed OPC diagnosed within 1 year prior to the interview. Controls were 2492 subjects (1497 men and 995 women; median age 58 years, range 19–82 years) admitted to the same hospital networks in the same period for acute, non-neoplastic conditions, unrelated to alcohol drinking, tobacco smoking or long term dietary modifications. Of the controls, 24% were admitted for traumas, 27% for other orthopedic causes, 22% for surgical conditions, 9% for eye diseases, and 19% for miscellaneous other illnesses. Fewer than 5% of potential cases and controls contacted refused to participate. Centrally trained interviewers used the same structured questionnaire and coding material in all centers. Apart from the dietary habits, the questionnaire collected information on socio-demographic characteristics such as education and occupation, tobacco and alcohol consumption, physical activity, anthropometric measures, personal medical history and family history of cancer. The study protocol was approved by the local ethical committees and all participants gave informed consent to participate.

4.2.2. Dietary intake assessment

Dietary intake was assessed through a structured validated[87] and reproducible[88-89] food frequency questionnaire(FFQ) including weekly consumption of 78 food items or recipes and five alcoholic beverages. Intake frequencies lower than once in a week, but at least once per month were coded as 0.5. Italian food composition tables were used to calculate energy intake and nutrients [90].

Food items and recipes were grouped into 25 food groups according to similar nutritional characteristics. Daily intake (g/d) was calculated for the food groups (Table 1) using standard portion sizes. The major part of food groups' distributions were skewed with a huge spike at zero (nonconsumers). We decided for categorization instead of transformation as we wanted to treat zeros differently from non-zeros. Especially with FFQ[2], they are expected to represent habitual non-consumption, therefore, they are likely to correspond to interesting population subgroups, e.g. vegetarians. Moreover, original variables were not continuous in nature. Categorization was done as follows. Indicators with a percentage of nonconsumers less than 10% (n=16) were categorized in a 2-level variable: below or above the median. Indicators with a proportion of non consumers between 10-50% (n=6) were categorized in a 3-level variable: nonconsumers and below or above the median among consumers (g/d>0). Indicators with a proportion of nonconsumers (n=3) equal or higher than 50% were dichotomized in consumers and nonconsumers. Categories were considered to be nominal, rather than ordinal due to a higher classification performance.

4.2.3. Statistical methods

We defined dietary patterns as unobserved classes in a population having different food consumption probability distributions. LCA was used to identify a set of mutually exclusive clusters of individuals, based on their responses to the set of observed food groups (indicators).

Total non-alcoholic energy intake influence was evaluated in the pattern identification using Wald test on the regression parameters related to its association with single food groups and the latent pattern variable. The correction for energy intake permits to obtain dietary patterns controlled for the overall energy intake.

Given the assumption of conditional independence, any residual association between two indicators after including the latent variable indicates a violation. These can be quantified and tested using the bivariate residuals (BVR) statistic. When the BVR becomes too high, and it is theoretically warranted, the indicators can be allowed to covary to locally relax the assumption. Therefore, we evaluated the within-class residual correlations (local dependencies) among food groups intake checking the BVR between pairs of food groups and allowed for correlated errors between food groups that showed high values of the statistic.

Class parity was determined as follows. The trivial 1-class model, where all individuals belong to the same class, was first fitted. The number of classes was successively increased by 1 in each subsequent model until the value of the BIC ceased to monotonically decrease or until the number of classes reached 10. This parity was chosen as the maximum to ensure substantial reduction in dimension from 25 food groups.

Names of the clusters were chosen according to the conditional distribution of food groups intake giving the latent classes (class-specific response probabilities).

Subjects were assigned to latent classes based on their posterior class membership probabilities. These were obtained from the estimated parameters of the LC model and their observed responses. Proportional allocation was chosen to permit a 'soft' classification, assigning subjects to each class with a weight equal to posterior membership probability for that class.

We examined the distribution of the identified clusters according to the selected nutrients used as dietary variables in the previous publication[3], performing a comparison between the previous nutrient based dietary patterns and the ones from the current analysis. As the previous publication regarded a subsample

of the current database (data collected between 1992 and 2005), the same analysis of the previous study was repeated and the robustness of the solution was checked and guaranteed (data not shown).

We also assessed the characterization of the clusters in terms of selected demographic/antrophometric characteristics and the known main risk factors for OPC, tobacco and alcohol consumption.

Odds ratios and related 95% confidence intervals for OPC risk were derived through a multiple logistic regression model using the class assignment to evaluate the effect of dietary patterns on the risk of OPC including terms for age, sex, education, body mass index (BMI), tobacco and alcohol consumption as confounders. Bolck, Croon, and Hagenaars [55] demonstrated that the classical three-step approach, which first identifies patterns, then assigns subject to each cluster and finally builds the prediction model, underestimates the associations between covariates and class membership. They proposed resolving this problem by means of a specific correction method. Vermunt [56] proposed a new maximum likelihood (ML) based correction method which is more efficient [110]. In this study, this ML correction is used which incorporates uncertainty about classification in the estimation procedure. As classification errors exist even in proportional assignment, this source of error or uncertainty must be taken into account when estimating effects between the latent variable and outcome variables.

LCA was performed on both cases and controls. Analysis on controls only was also carried out to check the robustness of the previous solution. As dietary patterns identified on controls were consistent (number and characteristics of the patterns) with the ones obtained on the overall sample (data not shown) we based all our analysis on the overall sample.

Statistical analysis were performed using SAS 9.4 (SAS Institute, Cary, NC, USA) and Latent GOLD 5.1 (Vermunt & Magidson,2016) statistical software.

4.3. Results

When fitting the LC model, we chose the solution with 4 classes according to the BIC criterion (Supplementary Table 4.1).

Cluster prevalence and food groups consumption were conditioned on total non-alcoholic intake in the final models as there were significant associations according to Wald tests on the related regression parameters (see Supplementary Table 5.1).

The BVR statistics showed high correlated errors between sugar and coffee food groups and between pulses and soups food groups. As the FFQ questions on sugar were related to hot beverages and in the construction of food groups variables pulses and soup shared an item, we specified correlated errors between coffee and sugar groups and between soup and pulses groups in the final model.

Table 2 reports the conditional distribution of food groups intake giving the latent classes for the food groups more relevant in discriminating and labeling the clusters. The complete table is given in Supplementary Table 4.2. Cluster 1 labeled ‘Prudent pattern’, showed higher probability to consume more leafy and fruiting vegetables, citrus fruit and all other kinds of fruits, tea and lower probability to consume red meat. Subjects in Cluster 2, that we named ‘Western pattern’, reported higher consumption of red meat and lower consumption of fruits, cruciferous and fruiting vegetables. Clusters 3 and 4 were related with similar food groups, but with a difference in the amount of intake. We termed Cluster 3 ‘Lower consumers-combination pattern’ as people in it were less likely to eat fruits, leafy and fruiting vegetables, pulses, potatoes, fish, white and red meat, bread and tea/decaffeinated coffee. Cluster 4 had higher probability to eating fruiting, leafy and other vegetables, white and red meat and bread, while showed a lower probability to consume coffee, tea, processed meat, cheese, fish, sugary drinks and desserts. We

called this cluster 'Higher consumers-combination pattern'. Estimated cluster's sizes were 36.8% of the population (n=1265) for the 'Prudent pattern', 27.0% (n=929) for the 'Western pattern', 21.1% (n=725) for the 'Lower consumers-combination pattern' and 15.1% (n=519) for the 'Higher consumers-combination pattern'.

Descriptions of the clusters for selected variables are given in Table 3. With regard to demographic characteristics, the 'Western pattern' showed the highest proportions of subjects less than 50 years old (25.3%), while the 'Higher consumers-combination pattern' the lowest one (19.4%). The 'Lower consumers-combination pattern' had the highest proportion of people more than 60 years old (16.5%), with respect to other clusters. The 'Prudent' pattern was populated by comparable proportions of men and women (53.1% and 46.9% respectively), while in other clusters men were predominant (65.5% to 81.1%). Subjects in the 'Prudent pattern' tended to be highly educated (23.3%), while the 'Higher consumers-combination pattern' showed the highest proportion of subjects with less than 7 years of education (64.7%). Regarding the two main risk factors for OPC, the 'Prudent pattern' was characterized by a lower consumption of alcohol and tobacco (respectively, 24.5% and 44.8% the proportions of non consumers) with respect to other clusters. The 'Higher consumers-combination pattern' had the highest proportion of heavy drinkers (64.4%), followed by the 'Western pattern' (52.5%). These two pattern showed a similar characterization in terms of tobacco consumption, with the smallest proportion of non smokers (29.0% and 27.1%, respectively).

Table 4 reports Cluster's characteristics in terms of non-alcoholic energy intake and nutrients intake: the 'Higher consumers-combination' pattern showed the highest energy intake, followed by the 'Western', the 'Prudent' and the 'Lower consumers-combination'.

The 'Prudent' pattern's diet was characterized by high intake of all the nutrients associated to the 'Vitamin and fiber' pattern found in the previous analysis [3]. Those nutrients were soluble carbohydrates, vitamin C, beta-carotene equivalents, total fiber. This pattern also exhibits highest intake of calcium. People in the 'Western' pattern reported a diet rich in those nutrients related to the previous 'Animal products' nutrient based pattern (animal protein, animal fat, cholesterol, saturated fatty acids, phosphorus and vitamin B2), with the exception of calcium. This pattern was also characterized by the highest intake of retinol which was related to the previous 'Retinol and Niacin'. The 'Higher consumers-combination' pattern, exhibited high consumption of many nutrients, manifesting characteristics in common with all the different previous nutrients based patterns. People in this group reported high intake of animal protein, animal fat, cholesterol, saturated fatty acids, phosphorus and vitamin B2 (previous 'Animal products' pattern), beta-carotene equivalent and total fiber (previous 'Vitamin and fiber' pattern), vegetable protein, starch, sodium (previous 'Starch' pattern), vegetable fat, vitamin E, monounsaturated and polyunsaturated fatty acids (previous 'Unsaturated fats' pattern) and niacin (previous 'Retinol and niacin' pattern). This pattern was also characterized by the highest intake of potassium, total folate and lycopene. The 'Lower consumers-combination' pattern reported a diet with the lowest intake of every nutrient.

Table 5 reports the ORs and corresponding CIs for OPC by the classification in the four dietary pattern from the composite model including the relevant confounding and risk variables. Interactions between dietary patterns and alcohol drinking or smoking habits were not significant. Hence, the composite model did not include interaction terms.

Compared to the Prudent pattern, the Western and the Lower consumers-combination ones were positively related to the risk of OPC (OR=2.56, 95% CI: 1.90 – 3.45 and OR=2.23, 95% CI: 1.64 – 3.02). Higher consumers-combination pattern didn't differ significantly from the Prudent pattern (OR=1.28, 95% CI: 0.92 – 1.77).

4.4. Discussion

Empirical *a posteriori* dietary patterns are derived predominantly using principal components, exploratory factor analysis, confirmatory factor analysis or cluster analysis. Despite the same label, dietary patterns derived from different methods are conceptually different [91]. Principal components analysis groups food variables in combinations that are representative of the original features of the dietary dataset. These combinations are the identified dietary patterns. Factor analysis group food items into dimensions with the assumption that if those items correlate highly, they might measure aspects of a common underlying dimension that represents a dietary pattern. Therefore, these techniques help to understand which foods are eaten in combination and to study the effect of these food groups dimensions/combinations on health outcomes. A disadvantage of these methods is that they do not give rise to mutually exclusive groups. Thus, when the interest is to compare groups of people, an additional step of cross-classification of the dimensions/combinations is needed. While FA and PCA group foods/nutrients items, cluster analysis groups individuals into relatively homogeneous classes. Therefore, CA define dietary patterns as classes of people where subjects share similar dietary habits and they are useful to study how these groups differs in terms of risk of an health outcome. However, some disadvantage of this method are that it assumes classification uncertainty to be 0, it mostly relies on non-parametric approaches which lack in assessment of goodness of fit. Moreover, all the above mentioned techniques do not take into account external covariates (e.g. confounders).

Our main objective was to identify dietary patterns conceived as mutually exclusive groups of people characterized by similar food intake and to compare the resulting patterns in terms of OPC risk. LCA can provide interesting insight into dietary patterning allowing to identify prevalent types of eating behavior in a population and to compare risk for people with different types of diet. The application of a LC model to the Italian case-control study on OPC has shown to overcome the above mentioned problems of the traditional methods and gives further advantages in dietary patterning, such as covariate adjustment, pattern prevalence estimation, and a probability based classification under a general parametric approach.

A previous publication regarding dietary patterns and the risk of OCP [3] using data from this multicentric case-control study was performed in 2010. The data used in that analysis were collected between 1992 and 2005, while the current database was updated including further 142 cases and 412 controls. The previous study aimed to identify dietary patterns conceived as ‘combination of dietary components intended to summarize key aspect of the diet for a given population’ [3] by performing principal component factor analysis on selected 28 nutrients and total non-alcoholic energy.

Hence, the aim of the study was slightly different from the current analysis, thus remaining in the general framework of the assessment of the effect of diet on OPC cancer. Moreover, techniques like FA/PCA/PCFA can be fairly applied on continuous variables (like nutrients), but when dealing with categorical ones (like foods/food groups) they may result in biased estimation. LCA, instead, can properly deal with categorical indicators and also be extended to ordinal or continuous variables in the framework of finite mixture models.

When the method applied is devoted to classifying individuals according to their food/nutrients intake and the data came from areas characterized by homogeneous diet, it is likely that the clustering method indentifies groups characterized by similar diet but different amount of food intake and consequently, energy intake. Not taking into account energy intake could lead to results that reflect the effect of the energy intake on disease and not the real effect of food itself. Direct covariate adjustment is not possible with other standard methods. In the previous publication[3], total non-alcoholic energy intake was taken into account by the inclusion of the related variable in dimensions identification together with the dietary indicators. Even though this approach has been commonly applied, it implies that the energy variable will influence the formation of dimensions and would, in essence become an ‘indicator’ of the dietary patterns,

treated in the same way as the proper dietary variables. Instead, theoretically it would be preferable to separate proper indicators from confounding variables. LCA, in contrast to the other methods, can be easily extended to include confounders keeping the measurement part of the model (which defines the relations with external variables) separate from the structural one (which defines the relation between the latent variables and its indicators) and permitting both to be estimated using a single ML estimation algorithm. In the current study, with LCA, we could correct dietary patterns including the energy variable as an external covariate, keeping it separated from the proper dietary indicators (the single food group item) to work with items that represented relative (adjusted) rather than absolute food intake. Then as the patterns prevalence also resulted depending on energy intake, it was possible to take this into account in the pattern identification by allowing the distribution of the latent pattern variable itself to depend on the covariate.

Differently from the above mentioned classical methods, LCA also allows first to inspect and then take into account for possible correlation between some dietary variables within classes. When the same measurement instrument is used for all foods, correlated errors are expected because of self-reporting bias and similar ways of wording items in the questionnaire. That leads to the importance of considering the effect of correlated errors between food intakes in the dietary pattern identification [2]. We were able to identify and allow correlated errors between some food groups that resulted from the nature of the assessment instrument and coding of food groups.

The study of the influence of a posteriori dietary patterns using food variables on OPC has been mostly addressed using PCFA. Our results were comparable to the evidence coming from these studies [3, 37-44,86]. A pattern related to high intake of fruit and vegetable, named 'Prudent' or 'Vegetable and Fruit' was found and associated to OPC in most of these studies. A second pattern characterized by high intake of meat was also often found. A part from meat, this pattern was also distinguished by other foods that varied in the different studies, which is why the label used changed from 'Western' to 'Snacks', for example. Apart from these two main patterns, the number and the types of further ones differed in the various publications. Patterns related to a combination of different food types varied from 'Traditional' country specific diet, 'Combination', 'Modern' 'Monotonous', 'Starchy', etc.

With regards to a potential 'Traditional' pattern, a characteristic of our study is the adjustment for non-alcoholic energy intake directly in pattern identification. This had influence in particular on foods having higher correlation with energy intake, like pasta, bread and sugary types of food. In fact, we note that adjusting for this covariate had the effect of not identifying a pattern strongly related to these foods as happened in different other studies. This effect was noted also in other papers analyzing the effect of energy correction in dietary patterning [91-92].

Concerning the influence of dietary patterns on the risk of OPC, we found a protective effect of the pattern characterized by high intake of leafy and fruiting vegetable and fruits ('Prudent pattern'). The Western cluster showed the highest risk of OPC. These results were comparable to those of other publications, that always found a protective effect of patterns related to high fruit and vegetables intake, and an increased risk of OPC cancer related to meat consumption in most of the cases [3,37-44,86].

Despite the above mentioned differences in aims, methods and dietary variables, our results were consistent with the previous publication with regards to the influence of diet on OCP risk [3]. The 'Prudent' pattern's diet was associated to all the nutrients that showed highest factor loadings for the pattern with the lowest risk for OPC ('Vitamin and fiber') in the previous study. Instead, the 'Western' pattern was associated to many the nutrients that showed highest factor loadings for a previous pattern associated to an increase of the risk of OCP ('Animal products'). The 'Higher consumers-combination' pattern exhibit high consume of nutrients associated to both previous protective ('Vitamin and fiber', 'Starch', 'Unsaturated fats') and risk factors patterns ('Animal products').

The clusters found with the LCA differed not only in terms of dietary intake but also in smoking and alcoholic consummation and in demographics (age, gender, education). Therefore, LCs on foods reflect typical diet-based groups in a population with all the side characteristics and also shows the importance of accounting for important risk factors in assessing the association between dietary patterns and health outcomes. Consequently, the ORs in the adjusted model showed smaller differences between classes. Tobacco and alcohol are, in fact, the major recognized risk factors for OPC [93-96], which emerged in our results too.

With reference to possible study limitations, hospital controls may be not representative of the general population for various aspects including dietary habits [97]. To limit this potential bias, controls were included according to a large spectrum of admission diagnoses, excluding the ones related to major know risk factors for OPC, such as tobacco and alcohol habits or long term dietary modifications. The recent diagnosis may affect patient's recall, but in our study, as the awareness of the role of diet on OPC risk was scarce, that this kind of misclassification was limited. Moreover, both cases and controls were interviewed in the same settings, by the same interviewers and with the same reproducible [88-89] and valid [87] FFQ. Among the strength of this study were the large sample size, the almost complete participation and the comparable catchment areas of cases and controls.

In conclusion, LCA gives further insights to dietary pattern research, allowing for the definition and estimation of the prevalence of different groups of subject characterized by different dietary choices, and comparing those groups in relation to important health outcomes like OPC. Thus, it adds a new perspective to the classical principal component/factor analysis which attempt to explain which foods are eaten in combination and their effect on health outcomes, and it has inferential advantages compared to cluster analysis.

4.5. Tables

Table 1 Content of food groups

Food Groups (labels)	Content
Milk	Milk, yoghurt
Coffee	Coffee
Tea and decaffeinated coffee	Tea, decaffeinated coffee
Bread	Bread, crackers, breadsticks, polenta
Pasta and rice	Pasta, rice
Soup	soups
Eggs	eggs
White meat	Chicken, turkey, rabbit
Red meat	Beef, horse, pork
Offal	Liver
Processed meat	Processed meat
Fish	Fishes
Cheese	Cheese
Potatoes	Potatoes
Pulses	Beans, peas, lentils, chickpeas
Leafy vegetables	Spinaches, sticks, salad, herbs
Fruiting vegetables	Tomatoes, zucchini, aubergines, peppers
Root vegetables	Carrots, onions
Cruciferous vegetables	Cabbages, cauliflowers, Brussels sprouts, turnip tops
Other vegetables	Artichokes, mixed salad
Fruits (not citrus)	Peaches, apricots, plums, melon, grapes, strawberries, cherries
Citrus fruits	Oranges, tangerines, grapefruits
Sugary drinks	Sugary drinks
Desserts	Biscuits, pies, pastries, croissants
Sugar	Sugar, sweeteners, candies

Table2 Probabilities of consumption for selected food items by dietary patterns derived from LCA. The latent class model was adjusted for non-alcoholic energy intake. Italy,1992-2009.

		Prudent %	Western %	Lower consumers-combination %	Higher consumers-combination %
Cluster's size		36,8	27,0	21,1	15,1
coffee	Below median	52,2	52,8	58,5	**60,6**
	Above median	47,8	47,2	41,5	39,4
tea	Not consumed	39,9	44,6	**62,3**	**62,8**
	Consumed	**60,1**	55,4	37,7	37,2
bread	Below median	57,8	44,1	**63,8**	22,9
	Above median	42,2	55,9	36,2	**77,1**
white	Below median	45,9	52,2	**61,3**	38,1
meat	Above median	54,1	47,8	38,7	**61,9**
red	Below median	**61,1**	30,9	**65,0**	35,7
meat	Above median	38,9	**69,1**	35,0	**64,3**
processed	Below median	50,7	47,3	50,8	**63,2**
meat	Above median	49,3	52,7	49,2	36,8
fish	Below median	45,6	49,2	**69,1**	**66,8**
	Above median	54,4	50,8	30,9	33,2
cheese	Below median	41,1	48,4	58,5	**62,6**
	Above median	58,9	51,6	41,5	37,4
potatoes	Below median	58,2	45,7	**61,7**	41,0
	Above median	41,8	54,3	38,3	59,0
leafy	Below median	35,5	59,8	**74,6**	22,5
vegetables	Above median	**64,5**	40,2	25,4	**77,5**
fruiting	Below median	24,4	**79,2**	**71,1**	31,0
vegetables	Above median	**75,6**	20,8	28,9	**69,0**
cruciferous	Not consumed	18,3	14,8	51,5	24,9
vegetables	Below median	25,1	**67,3**	19,8	46,6
	Above median	56,6	17,9	28,7	28,5
other	Not consumed	6,3	1,5	41,3	6,4
vegetables	Below median	36,1	55,9	54,3	22,7
	Above median	57,6	42,5	4,4	**70,9**
citrus	Not consumed	4,1	7,2	25,3	17,3
fruit	Below median	24,2	59,0	39,9	**62,8**
	Above median	**71,7**	33,8	34,8	19,9
other	Below median	29,8	**63,6**	**67,8**	50,2
fruits	Above median	**70,2**	36,4	32,3	49,8
sugary	Not consumed	54,2	41,2	59,9	**70,9**
drinks	Consumed	45,8	58,8	40,1	29,1
desserts	Below median	44,3	51,8	59,2	**68,0**
	Above median	55,7	48,2	40,8	32,0

Table 3 Dietary patterns' characteristics according to selected sociodemographic variables. Italy,1992-2009.

		Prudent %	Western %	Lower consumers-combination %	Higher consumers-combination %
	Cases	20,0	35,1	32,6	23,5
	Controls	80,0	64,9	67,4	76,5
Age (years)	<50	22,7	25,3	20,7	19,4
	50-59	30,1	31,4	26,8	34,8
	60-69	33,6	31,6	36,0	35,7
	>69	13,6	11,7	16,5	10,1
Sex	Male	53,1	73,8	65,6	81,1
	female	46,9	26,2	34,4	18,9
Education (years)	<7	46,4	55,7	55,3	64,7
	7-11	30,5	28,6	27,5	26,1
	>11	23,1	15,7	17,2	9,2
Alcoholic intake (weekly units)	0	24,5	12,4	22,0	9,0
	1-6	14,0	8,5	9,4	5,7
	7-13	15,0	10,5	11,1	8,3
	14-20	19,2	16,1	16,5	12,6
	>20	27,3	52,5	41,0	64,4
Smoking Habit	Never smoked	44,8	27,1	34,0	29,0
	Ex smoker	28,9	32,5	27,2	32,5
	<15 sig/d	10,8	13,7	12,2	12,1
	>14 sig/d	15,5	26,7	26,6	26,4
BMI	<18.5	2,1	2,1	3,4	1,8
	18.6-25.9	54,3	56,2	58,9	51,8
	26-29.9	32,0	29,2	25,6	32,8
	>29.9	11,6	12,5	12,1	13,6

Table 4 Dietary patterns' characteristics according to non-alcoholic energy intake and selected nutrients. Mean intake for each dietary pattern.Italy,1991-2009.

	Prudent	Western	Lower consumers-combination	Higher consumers-combination
Energy intake (kcal)	2252,4	2305,9	1838,0	**2582,6**
Animal protein (g)	59,2	**63,6**	50,7	**64,4**
Vegetable protein (g)	32,0	32,2	26,8	**38,6**
Animal fat (g)	41,7	**46,0**	36,3	**46,4**
Vegetable fat (g)	45,8	43,5	31,1	**56,1**
Cholesterol (mg)	301,7	**340,0**	253,1	336,7
Saturated fatty acids (g)	28,4	29,6	23,2	**31,5**
Monounsat. fatty acids(g)	41,3	39,5	30,0	**46,8**
Polyunsat. fatty acids (g)	1,2	1,3	1,0	**1,4**
Starch (g)	179,1	194,0	162,9	**229,2**
Soluble carbohydrate(g)	**115,5**	105,7	84,1	105,7
Sodium (mg)	2177,9	2318,1	1931,6	**2624,3**
Calcium (mg)	**987,8**	949,5	777,2	954,9
Potassium (mg)	4084,5	3873,1	3162,9	**4357,0**
Phosphorus (mg)	1554,0	1628,7	1317,2	**1728,1**
Iron (mg)	14,3	15,6	12,2	**18,0**
Zinc (mg)	12,9	13,8	10,9	**14,7**
Thiamin (vit. B1) (mg)	0,9	0,9	0,7	**1,0**
Riboflavin (vit. B2) (mg)	**1,7**	**1,7**	1,3	**1,7**
Vitamin B6 (mg)	2,0	2,1	1,6	**2,3**
Total folate (μg)	301,7	284,0	214,1	**316,5**
Niacin (mg)	19,1	19,9	15,9	**21,3**
Vitamin C (mg)	**180,5**	131,1	99,7	145,9
Retinol (μg)	770,0	**1268,4**	576,8	**873,7**
Beta-carotene equivalents (μg)	**4807,6**	3590,7	2501,7	4600,1
Lycopene (μg)	7172,4	7123,2	6324,0	**8782,7**
Vitamin D (μg)	3,3	3,3	2,5	3,1
Vitamin E (mg)	15,3	14,2	10,4	**18,2**
Total fiber (g)	**18,1**	15,0	12,3	**18,2**

Table 5 Odds ratios (OR) and corresponding 95% confidence intervals (CIs) for OPC for each cluster in models unadjusted and adjusted for known confounders. Italy,1992-2009.

		Unadjusted OR (95% CI)	Adjusted OR (95% CI) [a]
Dietary Patterns	Western	6.1 (4.4 -10.8)	2.56 (1.90 – 3.45)
	Lower consumers-combination	4.5 (3.2 – 6.2)	2.23 (1.64 – 3.02)
	Higher consumers-combination	3.9 (2.7 – 5.7)	1.28 (0.92 – 1.77)
	Prudent [b]	1	1

[a] Adjusted for sex, age, education, BMI, tobacco and alcohol intake

[b] Reference category.

5. ENERGY INTAKE ADJUSTMENT IN DIETARY PATTERN ANALYSIS THROUGH LATENT CLASS MODELS

5.1. Introduction

In nutritional epidemiology, total energy intake plays an important role in most of the studies of diet and diseases. The level of energy intake can be itself a determinant of some specific diseases, but even when it is not a direct cause, the association between diseases and specific nutrients may be confounded by total energy intake. Moreover, individual differences in total energy intake produce variation in the intake of specific nutrients unrelated to dietary composition [5]. All these aspects posed the necessity to consider total energy intake when interpreting association between diet and diseases and led to the suggestion that most research questions should focus on diet composition rather than the absolute amount of food consumed. Energy adjustment in dietary investigation reduces the variation of food intake resulting from differences in 'body size, metabolic efficiency and physical activity' [5].

In recent years, the interest in dietary patterns has grown as an alternative to the study of isolated components for the possibility to account for complex interactions among nutrients and foods. The primary objectives of a dietary pattern analysis are to characterise the eating habits of a population and to associate diet with disease. Latent class analysis (LCA) can achieve these objectives with additional advantages with respect to the traditional methods, such as principal components (PCA), factor (FA) and cluster analysis (CA).

One analytical decision, that has received little attention in the literature, is whether and how to adjust the models for energy intake in dietary pattern research using LCA. Traditionally, in the study of diet and diseases two kind of adjustments for energy intake are performed. A possible way to correct for energy intake is working with food already adjusted, usually with the residual method [5]. The second type of correction is made by entering the variable related to energy intake in the final model as confounder when assessing the association between diet and disease. In the framework of PCA, FA and CA applied to dietary pattern research another type of correction has often been used. It involves the inclusion of the energy variable directly in the classes/dimensions identification together with the other dietary indicators. Different from the above mentioned methods, LCA offers the possibility to correct for energy intake directly in pattern identification, distinguishing proper indicators from external variables and specifying different hypotheses on the effect of this variable.

Some studies using LCA have used correction for energy intake in class identification [2,36] or in the assessment of the association between dietary pattern and diseases [6,14], but no attention has been given to the effect that various types of adjustment may have and no comparison has been performed.

This study therefore set out to assess the effect of energy intake adjustment in dietary patterns identification through Latent Class Analysis. We also aimed to evaluate the effect of energy intake adjustment while assessing the influence of dietary patterns on oral and pharyngeal cancer risk, in a multicentric study conducted between 1992 and 2009 in Italy.

5.2. Materials and methods

5.2.1. Study population

We used data from a multicentric case-control study on oral and pharyngeal cancers (OPC) carried out between 1992 and 2009 in Italy, in the greater Milan area (northern Italy), the provinces of Pordenone (north-east Italy), and Rome and Latina (central Italy). The study included 946 patients (756 men, and 190 women; median age 58 years, range 22–79 years) admitted to major hospitals in the study areas with histologically confirmed oral cancer diagnosed within 1 year prior to the interview. Control were 2492 subjects (1497 men and 995 women; median age 58 years, range 19–82 years) admitted to the same hospitals in the same period for acute, non-neoplastic conditions, unrelated to known risk factors for oral cancer. Fewer than 5% of potential cases and controls contacted refused to participate. In each center the same structured questionnaire and coding material were used. Interviews were delivered by centrally trained and routinely supervised staff. Apart from the dietary habits, the questionnaire also collected information on various characteristics such as education, occupation, smoking and alcohol habits, physical activity, anthropometric measures, personal medical history and family history of cancer. The study was approved by the local ethical committees.

5.2.2. Dietary intake assessment

Dietary intake was assessed through a structured validated [87] and reproducible [88-89] food frequency questionnaire(FFQ) on the weekly consumption of 78 food items or recipes and five alcoholic beverages. All subject in the study had a complete FFQ. Italian food composition tables were used to calculate energy intake and nutrients [90]. Intake lower than once in a week, but at least once per month were coded as 0.5.

Food items and recipes were grouped into 25 food groups according to similar nutritional characteristics. Daily intake (g/d) was calculated for the food groups (Table 1) using standard portion sizes. The major part of food groups' distributions were skewed with a huge spike at zero (nonconsumers). We decided for categorization instead of transformation as we wanted to treat zeros differently from non-zeros. Especially with FFQ[2], they are expected to represent habitual non-consumption, therefore, they are likely to correspond to interesting population subgroups, e.g. vegetarians. Moreover, original variables were not continuous in nature. Categorization was done as follows. Indicators with a percentage of nonconsumers less than 10% (n=16) were categorized in a 2-level variable: below or above the median. Indicators with a proportion of non consumers between 10-50% (n=6) were categorized in a 3-level variable: nonconsumers and below or above the median among consumers (g/d>0). Indicators with a proportion of nonconsumers (n=3) equal or higher than 50% were dichotomized in consumers and nonconsumers. Categories were considered to be nominal, rather than ordinal due to a higher classification performance.

5.2.3. Statistical methods

Latent class analysis (LCA) was performed on the 25 foods groups. We specified correlated errors between coffee and sugar groups (in the FFQ sugar was related to hot beverages) and soup and pulses groups (the two groups shared one item). Those pairs showed the highest bivariate residual statistics(BVR).

LCA permits to account for total non-alcoholic energy intake (NAE) in pattern identification including it as an external covariate and allowing for its effect on the latent variable and/or on the single food items. These different formulations of the model correspond to different hypothesis on the effect of the NAE variable. We compared dietary patterns identified not adjusting or adjusting for NAE, using these three types of correction.

In the LC model, we have T response variables or indicators (food groups), denoted by y_{it} and a single categorical latent variable x, with K categories (dietary patterns). As we allowed for correlated errors

between some indicators, we use the symbol y_{ih} to denote one of the H subset of y_{it}. The y 's belonging to the same set h may correlate within latent classes. We consider a single exogenous variable (NAE) and we denote it by z_i.

0. LCA with no adjustment for NAE.

Assuming that the model does not contains the energy covariate, the general probability structure for the indicators is the one of the simple LCA model:

$$f(y_i)=\sum_{x=1}^{K} P(x) \prod_{h=1}^{H} f(y_{ih}|x)$$

Resulting patterns are not adjusted and therefore based on absolute food intake.

1. LCA with latent pattern variable depending on NAE.

The basic probability structure while allowing NAE to have effect on the latent variable is the following:

$$f(y_i|z_i)=\sum_{x=1}^{K} P(x|z_i) \prod_{h=1}^{H} f(y_{ih}|x)$$

It is be plausible that pattern prevalence may depend on NAE intake. This kind of adjustment therefore targets to correct for the association between the latent pattern variable and NAE intake.

2. LCA with single food items depending on NAE.

The general probability structure for this model that assumes NAE having effects on the single foods items is the following:

$$f(y_i|z_i)=\sum_{x=1}^{K} P(x) \prod_{h=1}^{H} f(y_{ih}|x, z_i)$$

This adjustment has the same aim of the residual methods, that is to derive dietary patterns based on food indicators that represents the relative intake of food.

3. LCA with both latent pattern variable and single food items depending on NAE.

The basic probability structure when NAE is assumed to affect both the latent variable and the indicators is the following:

$$f(y_i|z_i)=\sum_{x=1}^{K} P(x|z_i) \prod_{h=1}^{H} f(y_{ih}|x, z_i)$$

This last kind of correction combines both the aims of the previous two types of correction.

As we considered nominal indicators, we assume them to have a multinomial distribution with $M_h^* = \prod_{t\in h} M_t$ entries formed by the joint categorical variable y_{ih} obtained by cross-tabulating the categories of the variables in the subset h.

Logit functions are used for the linear predictors of $P(x|z_i)$ and $f(y_{ih}|x, z_i)$ and Wald tests on the related regression parameters were used to assess the strength of the influence of NAE on the food items and on the latent pattern variable.

Class parity for each model was determined fitting models from 1 to 10 classes and choosing the model with the lowest value of the BIC. A parity equal to 10 was chosen as the maximum to ensure substantial reduction in dimension from 25 food groups.

Names of the clusters were chosen according to the conditional distribution of food groups intake giving the latent classes (class-specific response probability). Instead of presenting class-specific marginal

probability, as it is usually done in traditional LCA, we choose to present the class-specific partial probability here to inspect the effect of the adjustment in dept. In models without local dependencies and external covariate adjustment, the two probabilities coincide. In other cases, while marginal probabilities should be obtained by aggregating over the other variables involved in the submodels for the response variable concerned, the partial probabilities represent the effect for a person with average values on all these variables. For our models, apart from local dependencies that are specified in the same way for all models, the main purpose of this kind of presentation is to show the conditional distribution of food groups intake in the latent classes for an average NAE intake, therefore not confounded by eventual differences in NAE intake between classes.

To determine the effect of adjustment in predicting a health outcome, we use the risk of OPC as example. A standard multiple logistic regression was fitted using the posterior membership probabilities estimated by LC models, through the three step approach [55] with proportional classification and ML correction [56] to evaluate the association between the classification and the risk of oral cancer. Five types of models were fitted.

Models with no adjustment for NAE in the pattern identification: dietary patterns from LC model unadjusted (0) with or without NAE included in the model as confounder were treated as exposures in the 3-step analysis. In the first case, that means that no correction for NAE intake was taken into account neither in pattern identification nor in the 3-step analysis. In the second case, the effect of NAE was taken into account just in the assessment of the association between dietary patterns and the health outcome.

Models with adjustment for NAE in the pattern identification: dietary patterns from LC model adjusted by types 1, 2, 3 of correction were evaluated as exposures for OPC risk.

In all the models, results were presented for both the solutions unadjusted and adjusted for selected known/potential risk factors for OPC.

Statistical analysis were performed using SAS 9.4 (SAS Institute, Cary, NC, USA) and Latent GOLD 5.1 (Vermunt & Magidson,2016) statistical software.

5.3. Results

According to BIC fit statistic, the best solutions identified a different number of classes according to the type of adjustment chosen. For model 0 we identified five classes, for model 1 seven classes, for model 2 five classes and for model 3 four classes.

Apart from very specific differences, we found two clusters that were robust in all the four correction solutions. One pattern was characterized by high intake of fruits and vegetables and avoidance of red meat. We called these clusters ‘Prudent pattern’. On the contrary, the second class that was common to all models exhibit a high intake of red meat and low intake of certain fruits and vegetables . We labelled it ‘Western pattern’.

All the other pattern were related to a combination-type of diet, with a strong difference in the amount of food consumed. Some of them showed a low intake of the majority of foods, therefore we labelled these clusters ‘Lower consumers- combination patterns’. People in the remaining classes reported a diet rich in red meat, bread, certain fruits and vegetables. We termed these classes ‘Higher consumers- combination patterns’.

Relevant differences in the four solutions regarded the number of these last two combination-diet patterns, which showed very specific differences within the two macro groups (lower-higher consumers). We found two Higher consumers–combination patterns in solution 0, 1, 2 and one pattern in solution 3. We found one Lower consumers-combination pattern in solution 0, 2, 3 and three patterns in solution 1.

Description of the classes for each model in terms of non-alcoholic energy intake and selected nutrients are given in Table 2. LC model 1, which holds the higher number of clusters, got also the more extreme groups in term of NAE with a daily intake of 3157,1 kj for the second 'Higher consumers-combination' cluster and 1344,7 kj for the first 'Lower consumers-combination' cluster. This clearly reflects on the amount of nutrients intake of the clusters. Instead, LC model 2 showed classes characterized by similar NAE intakes.

Changes in the correlations between foods in the two solutions were observed (data not shown). The main effect that appeared was a decrease of the highest correlation between foods (coffee vs tea, sugar vs desserts, desserts vs sugary drinks, red meat vs pasta) after the adjustment for energy intakes.

With regards to the influence of NAE on the latent variable and food groups, the Wald tests on the regression parameters showed strong associations in every model (Supplementary Table 5.1).

Table 3 reports the ORs and related CIs for the risk of oral cancer by dietary patterns for the five solutions, adjusted or not adjusted for known or potential risk factors. When not adjusting for risk factors, differences in the estimations for the five solutions were found. In general, adjustments for NAE results in a mitigation of the effects, thus remaining in the same order. Adjustments in the pattern identification (Models 1-3) resulted also in a shrinkage of the confidence intervals, especially in the model with the strongest correction (Model 3). On the contrary, when adjusting for known/potential risk factors, estimations of ORs and related CI remained consistent in all the models.

With respect to the Prudent one, the Western pattern was significantly related to the risk of OPC (ORs from 2.3 to 3.0) in all the models adjusted for known/potential risk factors. The Lower consumers-combination ones were also associated to a significant increase of the risk (ORs from 2.2 to 2.7) with the exception of Model 1 where two classes did not differ from the Prudent pattern (ORs 1.0 and 1.6). The Higher consumers- combination patterns did not differ significantly to the Prudent pattern in the majority of cases (ORs from 1.1 to 1.9).

5.4. Discussion

In epidemiological studies regarding the effect of diet on health outcomes total energy intake needs special attention and it should be considered when interpreting association between specific foods or nutrients and diseases [5].

Whether or not to adjust for energy intake in epidemiological studies is nowadays still debatable. Even though absolute amount of food or nutrients is biologically most relevant, adjusting for energy intake has the objective to determine when the effect exists *per se* [5] and it does not derive from the level of energy intake.

The traditional ways to account for energy intake in dietary patterning consist of two types of adjustment performed, respectively, before or after the dietary pattern identification phase of the analysis. The first regards performing the analysis on foods already adjusted, usually with the residual method [5]. The

second type is performed when assessing the association between dietary patterns and disease, by entering the variable related to energy intake in the final model as confounder.

In dietary pattern research another type of correction has been performed with factor, cluster or principal component analysis besides the above mentioned ones. This involves the inclusion of the energy intake variable in class/dimension identification together with the dietary indicators. This approach has been commonly applied and it was used also on a subset of our database in the study of nutrient dietary patterns[3]. Nevertheless, it is not fully correct from the theoretical point of view. In fact, in this way, the energy variable will influence the formation of classes and would, in essence become an indicator of the dietary patterns. In general, one wants to keep the part of dietary pattern identification (measurement part) and the assessment of the influence of external variables (structural part) separated.

Different from the above mentioned traditional methods for dietary pattern identification, LCA can be easily extended to include exogenous variables as covariates, permitting both the measurement part and the structural part of the model to be performed simultaneously using a single ML estimation algorithm.

LCA allows various means to take into account for total energy intake as its influence can be evaluated at the level of latent pattern variable and/or food indicators.

Regressing only food variables on energy intake is the analogous to working with food already adjusted (as is done with the residual method) for LCA. This correction results in classes with homogeneous energy intake and has the advantage of quantifying the energy effect differently from the residual method. Nevertheless, restricting to just this type of adjustment does not permit to improve the prediction of the latent pattern variable by covariate adjustment [98]. Pattern prevalence itself, especially in a population with homogeneous diet, may depend on total energy intake. However, performing alone this last type of correction does not permit to focus on dietary composition rather than absolute amount of food. Therefore, it may be preferable to first proceed with regressing food variables on NAE, and then to check and eventually allow for both types of correction. The appropriateness of this choice in our analysis was witnessed by significant parameters for both these types of associations assessed with Wald tests. We can conclude that when the aim is to correct for energy intake directly in the pattern identification part of the analysis, it is important to evaluate both the effects of the energy variable (on classes and on single food items) to focus on dietary composition and improve the prediction of the latent pattern variable by covariate adjustment [98].

Differences in terms of the number of classes extracted and class specific food intake were observed between the different adjusting solutions. The clusters that resulted robust in the different solutions were the ‘Prudent pattern’ and the ‘Western pattern’. We noted that the patterns which changed were those that showed highest/lowest non-alcoholic energy intake.

Balder et al.[92] found a similar issue while examining the stability of dietary patterns using factor analysis by different analytic decisions. They noted that the patterns with the high loadings on energy contributing foods in the unadjusted model were the one which changed. Northstone et al.[91] observed this happening while correcting food indicators, declaring that this was the result of the fact that adjusting for energy intake the food groups makes them being not correlated with energy.

With regard to OCP risk, differences in the estimations for the different solutions were found only when ORs were not corrected for known/potential risk factors. The correction in the identification phase of the analyses outperformed the other type of correction in controlling for energy intake. Differences between

the three types of correction were also found, ascribable to the different hypothesis on the effect of energy that they implied.

While assessing the association between dietary patterns and any health outcome is important to take into account for potential risk factors, as the groups identified through LCA may differ with respect to them and the resulted association may be confounded. When adjusting for known/potential risk factors, estimations of ORs and related CI remained consistent in all the models we fitted.

Our results were comparable to previous publications, which found a protective effect of patterns related to high fruit and vegetables intake, while with regards to meat consumption, it has been related to an increased risk of OPC cancer in many studies, although not all studies provided consistent results [3,37-44,86].

Balder et al.[92] affirmed that the pattern obtained through factor analysis using unadjusted food variables were comparable to those using energy adjusted data. Northstone et al. [91] in their study on the effect of energy adjustment in dietary patterning with PCA concluded that, although there were differences in the dietary patterning solutions obtained with unadjusted or energy-adjusted data, these differences did not appear to have major impact on the association with their health outcome.

With regards to factor analysis, Balder et al. [92] concluded as the dietary patterns they found through FA were robust to energy adjustment, they indeed were based on relative consumption of food rather than actual intake and there was no need for energy correction. Northstone et al.[91] using PCA recommended to make adjustments at a later stage when analyzing the effects of dietary patterns on the outcome of interest, although it is important to present both unadjusted and adjusted results, mostly because the residuals methods does not permit to clearly evaluate the effect of energy intake.

LCA overcomes this problem, permitting to quantification the effect of the energy intake both on the dietary variables and on the pattern variable.

Therefore, we conclude that dietary patterns identified through LCA were robust to energy adjustment when controlling for known/potential risk factors in the assessment of their association with the risk of OPC. The correction in the identification phase of the analysis has the aim to avoid the identification of clusters that discriminate mainly for the amount of energy intake and to estimate the effect of energy on dietary items/classes. When choosing to perform this kind of correction, we recommend to evaluate the effect of energy intake both on food items and latent classes to correctly specify the structure of dependencies between all the variables involved in the model.

5.5. Tables

Table 1. Probabilities of consumption for selected food items by dietary patterns derived from the four solutions of LCA. Italy 1992-2009.

		Model 0					Model 1							Model 2					Model 3			
	Classes	1	2	3	4	5	1	2	3	4	5	6	7	1	2	3	4	5	1	2	3	4
	Labels	P[a]	W[b]	H1[c]	H2[c]	L1[d]	P[a]	W[b]	H1[c]	H2[c]	L1[d]	L2[d]	L3[d]	P[a]	W[b]	H1[c]	H2[c]	L1[d]	P[a]	W[b]	H1[c]	L1[d]
milk	Not consum.	13.5	23.7	38.2	11.8	32.2	11.8	24.5	36.0	11.8	41.2	23.8	16.9	8.7	25.5	33.6	32.4	27.2	11.9	23.1	42.1	25.3
	Below median	53.1	43.4	44.8	44.9	46.8	50.5	45.9	42.4	41.3	46.4	53.9	48.0	49.0	46.6	44.8	52.3	43.5	49.9	46.3	47.1	44.5
	Above median	33.4	32.8	16.9	43.4	21.1	37.7	29.7	21.6	47.0	12.4	22.3	35.0	42.3	27.9	21.6	15.3	29.3	38.2	30.6	10.8	30.2
coffee	Below median	55.9	55.0	59.7	48.0	59.6	56.5	48.7	**60.7**	48.2	**63.9**	55.6	56.8	57.7	56.0	**61.2**	43.4	57.5	52.5	54.7	**62.1**	56.2
	Above median	44.1	45.0	40.4	52.0	40.4	43.5	51.3	39.3	51.8	36.1	44.4	43.2	42.3	44.0	38.8	56.6	42.5	47.5	45.3	38.0	43.8
tea	Not consum.	42.2	43.7	**62.8**	39.8	**62.5**	37.5	45.9	59.0	42.2	**67.3**	48.8	55.0	32.9	45.8	58.7	**70.0**	**60.4**	40.0	44.8	**63.5**	**61.5**
	Consumed	57.8	56.3	37.2	**60.2**	37.6	**62.5**	54.1	41.0	57.8	32.7	51.2	45.0	**67.1**	54.2	41.3	30.0	39.6	**60.1**	55.2	36.5	38.5
bread	Below median	**70.8**	38.0	33.0	29.9	**67.0**	55.1	58.1	15.4	19.3	**85.9**	**83.5**	51.9	**61.9**	42.2	33.9	39.3	47.5	59.6	44.3	25.6	47.7
	Above median	29.2	**62.0**	**67.0**	**70.1**	33.0	44.9	41.9	**84.6**	**80.7**	14.1	16.5	48.1	38.1	57.8	**66.1**	**60.7**	52.5	40.4	55.7	**74.5**	52.3
pasta	Below median	**60.0**	45.7	55.3	27.5	58.4	48.7	48.0	47.0	20.8	**71.6**	**75.4**	40.7	52.3	46.2	**63.2**	36.6	47.0	48.8	48.1	58.3	40.3
	Above median	40.0	54.3	44.7	**72.5**	41.6	51.3	52.0	53.0	**79.2**	28.4	24.6	59.3	47.7	53.9	36.8	**63.5**	53.0	51.2	51.9	41.7	59.7
soup	Below median	53.5	38.8	44.7	47.1	53.5	51.7	48.4	36.2	44.4	54.8	53.5	52.3	46.4	43.2	42.2	**62.5**	49.5	50.3	43.0	46.1	52.8
	Above median	46.5	**61.2**	55.3	52.9	46.5	48.3	51.6	**63.8**	55.6	45.2	46.5	47.7	53.6	56.8	57.8	37.6	50.5	49.7	57.0	53.9	47.2
eggs	Not consum.	16.7	5.6	15.1	5.3	23.4	10.9	5.0	11.9	5.9	25.1	20.9	20.7	9.7	5.9	15.2	18.9	22.5	11.1	5.7	18.6	21.6
	Below median	46.1	38.7	35.0	36.5	45.8	40.0	49.1	32.7	34.0	44.5	48.0	45.1	43.3	41.6	36.0	42.9	41.2	43.4	42.3	35.9	42.3
	Above median	37.2	55.8	49.9	58.2	30.8	49.0	45.8	55.4	**60.1**	30.4	31.1	34.2	47.0	52.5	48.8	38.3	36.3	45.6	52.0	45.5	36.1
white	Below median	47.1	47.9	43.0	41.5	**63.4**	45.1	**61.2**	40.3	38.5	**71.3**	50.5	54.5	44.6	50.8	49.2	42.7	**61.3**	46.0	52.9	40.8	58.0
meat	Above median	52.9	52.1	57.0	58.5	36.6	54.9	38.8	59.7	**61.5**	28.7	49.5	45.5	55.4	49.2	50.8	57.3	38.7	54.0	47.1	59.2	42.0
red	Below median	**79.5**	24.1	39.1	29.7	**66.4**	**68.1**	34.6	23.9	18.6	**78.7**	**78.0**	**60.1**	**65.3**	28.0	49.8	50.3	51.2	**62.9**	29.9	42.3	53.7
meat	Above median	20.5	**75.9**	**60.9**	**70.3**	33.6	31.9	**65.4**	**76.1**	**81.4**	21.3	22.0	39.9	34.7	**72.0**	50.2	49.7	48.8	37.1	**70.1**	57.8	46.3

offals	Not consum.	**82.4**	46.8	**77.2**	**60.5**	**77.9**	**77.3**	46.4	**68.1**	52.4	**73.5**	**84.3**	**83.5**	**75.9**	46.9	**85.5**	**74.0**	**79.9**	**75.8**	49.8	**75.1**	**79.0**
processed meat	Consumed	17.6	53.2	22.8	39.5	22.1	22.7	53.6	31.9	47.6	26.5	15.7	16.5	24.1	53.1	14.5	26.0	20.2	24.2	50.2	24.9	21.0
fish	Below median	**64.8**	47.5	**71.6**	27.9	55.7	56.2	49.3	56.0	27.7	**63.0**	**69.6**	43.9	53.1	48.2	**83.9**	39.3	48.5	51.0	48.4	**69.4**	43.7
cheese	Above median	35.2	52.5	28.5	**72.1**	44.3	43.8	50.7	44.0	**72.3**	37.1	30.4	56.1	46.9	51.8	16.1	**60.7**	51.5	49.0	51.6	30.6	56.3
potatoes	Below median	50.3	50.4	**73.6**	39.0	**67.5**	49.7	42.6	**70.9**	37.6	**72.5**	54.8	**66.0**	46.6	49.1	**81.0**	43.5	**72.5**	45.7	49.6	**68.7**	**67.1**
pulses	Above median	49.7	49.7	26.4	**61.1**	32.5	50.4	57.5	29.1	**62.4**	27.5	45.3	34.0	53.4	50.9	19.0	56.5	27.5	54.3	50.4	31.3	32.9
leafy vegetables	Below median	51.6	49.5	59.3	31.8	**61.1**	40.5	50.8	54.6	30.8	**71.4**	**63.9**	49.1	40.5	51.1	**61.3**	54.5	54.2	41.0	49.5	**68.8**	51.8
fruiting vegetables	Above median	48.4	50.5	40.7	**68.2**	38.9	59.5	49.2	45.4	**69.2**	28.6	36.1	50.9	59.5	48.9	38.7	45.5	45.9	59.0	50.5	31.2	48.2
root vegetables	Below median	**76.1**	45.1	34.7	34.0	**63.6**	58.6	52.0	36.4	33.1	**68.6**	**74.1**	56.2	**60.8**	46.0	37.4	54.0	56.2	58.8	46.7	46.1	55.2
cruciferous vegetables	Above median	23.9	54.9	**65.3**	**66.0**	36.4	41.4	48.0	**63.6**	**67.0**	31.4	26.0	43.8	39.2	54.0	**62.6**	46.0	43.8	41.2	53.3	53.9	44.8
	Below median	45.9	**60.3**	48.8	38.4	**63.4**	42.5	**60.2**	53.2	40.0	**68.7**	53.8	55.4	45.4	59.4	54.3	34.9	**63.0**	42.6	**60.7**	48.6	58.0
	Above median	54.1	39.8	51.2	**61.6**	36.6	57.5	39.8	46.8	**60.1**	31.3	46.2	44.6	54.7	40.7	45.7	**65.1**	37.0	57.4	39.3	51.4	42.0
	Below median	37.4	55.7	19.9	36.7	**75.3**	31.6	**75.3**	30.6	35.5	**85.5**	43.1	**64.9**	38.5	58.4	18.5	33.8	**73.7**	35.5	**60.7**	24.5	**71.5**
	Above median	**62.6**	44.3	**80.1**	**63.3**	24.7	**68.4**	24.7	**69.4**	**64.5**	14.5	56.9	35.1	**61.5**	41.6	**81.5**	**66.2**	26.3	**64.5**	39.3	**75.5**	28.5
	Below median	29.0	**86.1**	19.7	28.5	**73.8**	18.8	**79.6**	51.2	42.3	**81.7**	37.9	**66.6**	29.3	**79.4**	20.8	21.6	**76.6**	24.4	**79.5**	32.3	**69.7**
	Above median	**71.0**	13.9	**80.3**	**71.5**	26.2	**81.2**	20.4	48.8	57.7	18.3	**62.1**	33.4	**70.7**	20.6	**79.2**	**78.4**	23.4	**75.6**	20.5	**67.7**	30.3
	Not consum.	8.5	15.5	30.3	17.4	53.5	8.6	15.8	33.4	14.4	57.1	17.3	55.5	8.1	16.7	24.5	35.1	58.4	11.2	17.7	30.9	57.2
	Below median	36.8	47.5	27.5	35.4	35.9	32.5	49.6	35.6	37.0	37.0	38.9	31.1	34.6	48.6	28.1	36.6	31.7	34.7	48.9	30.5	32.4
	Above median	54.7	37.1	42.3	47.3	10.6	58.9	34.6	31.1	48.6	5.9	43.9	13.5	57.3	34.8	47.4	28.3	10.0	54.1	33.3	38.6	10.4
	Not consum.	20.8	14.7	25.0	17.0	45.2	16.3	13.7	24.5	16.1	45.0	25.0	53.1	18.7	14.2	26.4	22.7	52.5	18.3	14.9	25.2	50.9
	Below median	25.7	75.8	53.8	25.1	27.0	29.2	**65.5**	58.6	37.6	32.0	29.5	15.2	27.6	67.0	68.8	6.1	27.8	25.2	**67.4**	47.2	19.5
	Above median	53.6	9.5	21.2	57.9	27.8	54.6	20.8	17.0	46.3	23.0	45.5	31.8	53.7	18.9	4.8	71.2	19.7	56.5	17.7	27.6	29.7

other vegetables	Not consum.	5.1	1.0	4.6	10.6	32.7	4.8	0.4	7.4	5.4	33.7	9.5	42.8	6.8	1.6	2.3	13.4	37.4	6.3	1.5	6.7	41.0
	Below median	38.2	55.8	11.6	38.2	56.9	31.2	55.5	35.3	40.4	58.2	41.3	54.3	37.8	56.6	6.5	40.3	55.1	36.2	56.2	23.5	54.3
	Above median	56.7	43.2	**83.8**	51.3	10.4	**64.0**	44.1	57.3	54.2	8.1	49.2	2.9	55.3	41.8	91.3	46.3	7.5	57.5	42.2	**69.9**	4.7
citrus fruit	Not consum.	5.8	6.4	17.8	2.0	25.8	1.5	5.3	17.8	2.7	27.7	15.0	19.8	3.6	7.5	17.0	9.0	26.7	4.0	7.3	18.7	22.5
	Below median	24.8	**63.6**	**67.3**	28.7	44.2	23.2	53.3	**68.6**	34.1	48.6	40.0	37.7	21.4	60.3	71.6	39.2	43.2	24.1	59.9	**64.5**	38.6
	Above median	**69.4**	30.0	14.9	**69.3**	30.0	**75.3**	41.4	13.6	**63.3**	23.7	45.0	42.5	75.0	32.2	11.4	51.8	30.1	**71.9**	32.8	16.9	39.0
other fruits	Below median	30.5	59.3	55.0	33.9	**72.1**	22.5	**68.5**	58.5	33.0	**82.5**	51.7	56.5	27.1	65.1	51.9	48.2	65.2	29.4	**65.1**	55.3	**62.6**
	Above median	**69.5**	40.7	45.0	**66.1**	27.9	**77.5**	31.5	41.5	**67.0**	17.6	48.3	43.5	73.0	34.9	48.1	51.8	34.8	**70.6**	34.9	44.7	37.4
sugary drinks	Not consum,	**63.5**	39.7	**69.7**	40.1	**63.6**	56.6	44.2	57.3	34.9	**70.5**	**71.4**	50.8	51.8	45.8	67.4	71.7	51.6	54.6	42.0	**76.6**	53.4
	Consumed	36.6	**60.4**	30.3	59.9	36.4	43.4	55.8	42.7	**65.2**	29.5	28.6	49.2	48.2	54.2	32.7	28.3	48.4	45.4	58.0	23.4	46.6
desserts	Below median	55.7	48.1	**64.5**	34.2	**66.1**	45.6	59.0	55.3	26.8	**85.4**	**71.4**	41.1	36.6	59.7	59.7	78.0	46.0	43.6	53.7	**78.4**	47.6
	Above median	44.3	51.9	35.5	**65.8**	33.9	54.4	41.0	44.7	**73.2**	14.6	28.6	58.9	63.4	40.3	40.3	22.0	54.0	56.4	46.3	21.6	52.4
sugar	Below median	**63.4**	39.5	51.2	28.5	59.8	46.1	58.3	36.3	24.7	**69.9**	**73.5**	43.6	42.0	47.3	50.4	72.2	38.2	49.2	44.8	58.2	43.2
	Above median	36.7	**60.5**	48.8	**71.5**	40.2	53.9	41.7	**63.7**	**75.3**	30.1	26.5	56.4	58.0	52.7	49.6	27.8	61.8	50.8	55.2	41.8	56.8

[a]Prudent patterns, [b]Western patterns, [c]Higher consumers-combination patterns, [d]Lower consumers-combination patterns.

Table 2. Sizes and mean weekly intake of selected nutrients and energy for LCs in the four solutions. Italy, 1992-2009

	Model 0					Model 1							Model 2					Model 3			
Labels	P[a]	W[b]	H1[c]	H2[c]	L1[d]	P[a]	W[b]	H1[c]	H2[c]	L1[d]	L2[d]	L3[d]	P[a]	W[b]	H1[c]	H2[c]	L1[d]	P[a]	W[b]	H1[c]	L1[d]
Sizes (%)	22.1	19.2	11.2	22.4	25.1	17.3	12.5	15.5	18.1	9.9	16.2	10.6	30.3	27.2	8.6	14.8	19.1	36.8	27.0	15.1	21.1
Energy intake (kJ)	2010	2401	2317	2731	1805	2255	2043	2592	3157	1345	1584	2111	2226	2243	2248	2215	2217	2252	2306	2583	1838
Animal protein (g)	52.7	66.0	58.2	71.2	50.0	58.0	58.0	66.5	80.2	39.9	45.3	56.7	58.9	62.4	54.6	58.1	58.8	59.2	63.6	64.4	50.7
Vegetable protein (g)	29.1	33.5	34.6	38.4	26.4	32.3	29.2	37.9	43.3	20.3	23.7	30.3	31.4	31.8	33.1	33.7	31.3	32.0	32.2	38.6	26.8
Animal fat (g)	35.6	47.9	42.0	52.7	35.4	40.8	40.4	49.5	61.8	25.9	27.9	42.0	41.7	44.1	41.4	38.2	45.1	41.7	46.0	46.4	36.3
Vegetable fat (g)	41.4	44.6	51.2	54.4	31.8	46.7	38.8	51.3	62.9	23.8	32.5	35.9	43.8	43.4	48.7	49.7	36.6	45.8	43.5	56.1	31.1
Starch (g)	158.6	203.0	200.7	226.6	159.4	177.7	171.1	230.5	263.9	118.2	126.5	185.5	175.5	189.5	192.3	191.2	197.4	179.1	194.0	229.2	162.9
Soluble carb. (g)	106.3	110.9	96.4	129.4	80.4	118.5	91.7	109.6	148.9	58.2	78.3	99.1	117.7	99.8	99.3	92.2	103.4	115.5	105.7	105.7	84.1

[a]Prudent patterns, [b]Western patterns, [c]Higher consumers-combination patterns, [d]Lower consumers-combination patterns.

Table 3. Odds ratios (OR) and corresponding 95% confidence intervals (CIs) for oral cancer risk for LCs by different types of adjustment. Italy, 1992-2009

		UNADJUSTED DIETARY PATTERNS		UNADJUSTED DIETARY PATTERNS with NAE included in the model		ADJUSTED DIETARY PATTERNS					
		Model 0				Model 1		Model 2		Model 3	
		ORs (95%CI)	ORs (95%CI[a])	ORs (95%CI)	ORs (95%CI[a])	ORs (95%CI)	ORs (95%CI[a])	ORs (95%CI	ORs (95%CI[a])	ORs (95%CI)	ORs (95%CI[a])
Prudent pattern		1[b]	1[b]	1[b]	1[b]	1[b]	1[b]	1[b]	1[b]	1[b]	1[b]
Western pattern		10 (6.4 - 18.8)	2.9 (2.0 – 4.2)	9.5 (5.7 - 15.9)	3.0 (2.0 – 4.3)	5.2 (3.2 – 8.6)	2.3 (1.5 – 3.7)	7.2 (4.8 – 10.8)	2.6 (1.8 – 3.5)	6.1 (4.4 – 8.5)	2.6 (1.9 - 3.5)
Higher consumers combination patterns	H1	6.6 (3.5 - 11.9)	1.4 (0.9 – 2.2)	5.8 (3.2 - 10.3)	1.4 (0.9 – 2.2)	6.9 (4.3 – 11.0)	1.9 (1.3 – 2.8)	4.8 (2.9 – 7.7)	1.5 (0.7 – 1.2)	3.9 (2.7 – 5.7)	1.3 (0.9 – 1.8)
	H2	3.5 (2.0 - 6.4)	1.3 (0.9 – 2.0)	2.9 (1.6 - 5.0)	1.5 (1.0 – 2.2)	2.5 (1.6 – 4.0)	1.1 (0.8 – 1.7)	2.8 (1.7 – 4.7)	1.2 (0.8 – 1.9)	-	-
Lower consumers combination patterns	L1	8.0 (4.7 - 13.9)	2.5 (1.8 – 3.7)	7.8 (4.6 - 13.0)	2.5 (1.7 – 3.6)	6.3 (4.0 – 9.9)	2.7 (1.8 – 4.1)	7.1 (4.7 – 10.5)	2.6 (1.8 – 3.5)	4.5 (3.2 – 6.2)	2.2 (1.6 – 3.0)
	L2	-	-	-	-	1.2 (0.7 – 2.1)	1.0 (0.7 – 1.6)	-	-	-	-
	L3	-	-	-	-	3.1 (1.9 – 5.2)	1.6 (1.0 – 2.5)	-	-	-	-

[a]Adjusted for sex, age, BMI, education, tobacco and alcohol consumption. [b]Reference category.

6. DIETARY PATTERNS INSPECTION THROUGH LATENT CLASS TREE: AN APPLICATION TO MULTICENTRIC CASE-CONTROL STUDIES ON SELECTED DIGESTIVE TRACT CANCER

6.1. Introduction

In the last years, Latent Class Analysis (LCA) has become a popular method in social and behavioural sciences. Even though less popular than the traditional methods such as principal components, factor and cluster analysis, LCA can provide interesting insight to detect mutual exclusive groups of subjects who share the same dietary behaviour. Moreover, recent developments of the methods can be used to address important issues in dietary patterning.

However, with large datasets (in terms of cases and/or variables), the fit of a LC model may improve until it identifies a large number of classes, as the model takes into account a large number of dependencies. Some classes may therefore differ from others in a very specific/less interesting way, so the interpretation of the final model could be troublesome and the comparison between classes could be arduous. Moreover, the choice of the goodness of fit measure (e.g. AIC or BIC) can result in identifying a completely different number of classes that are substantially hard to compare. This often ends up in different grouping solutions even on the same database.

Recently, Van den Bergh et al.[66,99-100] proposed a development of the classical LCA, called Latent Class Tree (LCT) to address these issues. This approach, based on an algorithm for density estimation developed by Van der Palm et al.[104], consists of a stepwise hierarchical partitioning of the data, imposing a hierarchical tree structure on the latent classes. Therefore, it leads to a solution that allows for direct interpretation of the relationships between classes and solutions with different numbers of classes. An advantage of this method is that the class characteristics remain the same for every chosen solution. The choice of a different goodness of fit measure results only in deciding the relative importance of a split, and consequently, in a different 'length' of the tree.

The same authors deepened the issue of relating latent classes to external variables through the adjusted three-step analysis in LCT modelling. The LCT way to detect dietary patterns allows different granularity that permit to inspect the relative importance of subgroups in their relation with health outcomes.

These recent developments of the traditional LCA has never been used in dietary patterns research. Therefore, the aim of this analysis is to identify dietary patterns through LCT to add a new perspective on the research on dietary patterns and their association with the risk of selected digestive tract cancer in Italy.

6.1.1. Materials and methods

6.1.2. Study population

We use data from two multicentric case-control studies respectively on oral/pharyngeal cancer(OPC) and esophageal squamous cell cancer (ESCC). The one related to OPC was carried out between 1992 and 2009, in the greater Milan area (northern Italy), the provinces of Pordenone (North-East Italy), and Rome and Latina (Central Italy) in a multicentric case-control. The study included 946 patients (756 men, and 190 women; median age 58 years, range 22–79 years) admitted to major hospitals in the study areas with

incident, histologically confirmed OPC diagnosed within 1 year prior to the interview. The study on ESCC was conducted from 1992 to 1997 in the provinces of Milan, Pordenone and Padua (northern Italy). Cases were 304 patients (275 men and 29 women; median age 60 years, range 39–77 years) admitted to the major teaching and general hospitals in the areas under investigation with incident (diagnosed within 1 year before inclusion in the study) for histologically confirmed squamous cell cancer of the esophagus, and with no history of cancer.

Controls were 2492 subjects for OPC (1497 men and 995 women; median age 58 years, range 19–82 years) and 743 subjects for ESCC (593 men and 150 women; median age 60 years, range 36–77 years)admitted to the same hospital networks in the same period for acute, non-neoplastic conditions, unrelated to alcohol drinking, tobacco smoking or long term dietary modifications. Fewer than 5% of potential cases and controls refused to participate for both the studies. Centrally trained interviewers used the same structured questionnaire and coding material in all centers. Apart from the dietary habits, the questionnaire collected information on socio-demographic variables such as education, occupation, tobacco and alcohol consumption, physical activity, anthropometric measures, personal medical history and family history of cancer. The study protocol was approved by the local ethical committees and all participants gave informed consent to participate.

6.1.3. Dietary intake assessment

Dietary intake was assessed through a structured validated [87] and reproducible [88-89] food frequency questionnaire(FFQ) including weekly consumption of 78 food items or food groups and five alcoholic beverages. Intake frequencies lower than once in a week, but at least once per month were coded as 0.5. Italian food composition tables were used to calculate energy intake and nutrients[90].

Food items and recipes were grouped into 25 food groups according to similar nutritional characteristics. Daily intake (g/d) was calculated for the food groups (Table 1) using standard portions' sizes. The major part of food groups' distributions were skewed with a huge spike at zero (nonconsumers). We decided for categorization instead of transformation as we wanted to treat zeros differently from non-zeros. Especially with FFQ[2], they are expected to represent habitual non-consumption, therefore, they are likely to correspond to interesting population subgroups, e.g. vegetarians. Moreover, original variables were not continue in nature. Categorization was done as follows. Indicators with a percentage of nonconsumers less than 10% (n=16) were categorized in a 2-level variable: below or above the median. Indicators with a proportion of non consumers between 10-50% (n=6) were categorized in a 3-level variable: nonconsumers and below or above the median among consumers (g/d>0). Indicators with a proportion of nonconsumers (n=3) equal or higher than 50% were dichotomized in consumers and nonconsumers. Categories were considered to be nominal, rather than ordinal due to a higher classification performance.

Categorization was performed according to the distribution of the variables in the dataset on which the specific analysis was performed.

6.1.4. Covariate adjustment and local dependency inspection

Total non-alcoholic energy intake influence was evaluated in the pattern identification by using Wald test on the regression parameters related to its association with single food groups and cluster membership. The adjustment for energy intake permit to focus on dietary composition rather than the absolute amount of food consumed and to improve the prediction of the latent variable by covariate adjustment[98].

We evaluated the within-class residual correlations (local dependencies) among food groups intake checking the bivariate residuals (BVR) between pairs of food groups and allowed for correlated errors between food groups that showed high BVR. This step is particularly important in this type of analysis

because potential violation of the assumption of independence between indicators within latent classes (conditional independence) may give rise to further classes in the class identification.

6.1.5. Latent Class Analysis solution inspection

We first inspected the solution obtained with the LCA approach. Analysis was performed on all cases of cancer (oral/pharyngeal and esophageal) and controls. As some controls were utilized in both studies, duplicated records were not included.

We first fitted the trivial 1-class LCA model, where all individuals belong to the same class, and then the number of classes was successively increased by 1 in each subsequent model until the value of a specific goodness of fit measure ceased to monotonically decrease or until the number of classes reached 10. This parity was chosen as the maximum to ensure substantial reduction in dimension from 25 food groups.

We also preformed LCA separately on the single case-control studies databases for oral/pharyngeal cancer and esophageal cancer, to check robustness of the previous solution and have insight on the choice of the number of classes for the first split of the LCT solution.

6.1.6. Latent Class Tree model

We fitted the LCT on the combined sample of the two studies on OPC and ESCC, excluding repeated controls.

In general, the hierarchical structure of a LCT is obtained by sequentially splitting each ‘parental’ class into two ‘child’ subclasses, starting from the complete sample. If a 2-class solution is preferred over the 1-class solution according to a model selection criteria, the ‘parental’ sample is split into 2 ‘child’ subsamples which contain the posterior membership probabilities for the class concerned as case weight. Then, each ‘child’ subsample is treated as a ‘parental’ one and the process is repeated on each of the weighted datasets. The process continues until only 1-class solutions are preferred, producing the hierarchical latent class tree.

The divisive algorithm that produces the LCT is based on the posterior membership probabilities for the two child classes conditional on the parental one. Therefore, a proportional split is done for each class or node. The weight at each node equals the weight at its parental node times the posterior probability of belonging to the concerned node conditional on belonging to the parent node.

6.1.7. Choice of the number of starting classes in LCT

As the first split is the one which picks up the most relevant associations in the data[100], we allowed the first split made on the entire sample to bear a maximum number of classes higher than 2. Therefore, the comparison for the first split are made for the 1-class vs 2-class,2-class vs 3-class,...,n-1-class vs n-class, with n as the maximum number chosen.

We decided the starting number of classes for the first split of the LCT according to substantial reasoning derived from the assessment of the LCA performed on the whole sample and on the two separate case-control studies databases, and to the relative improvement of the goodness of fit statistics.

We evaluated the relative improvement of the fit for the chosen solution for different fit measures (relative decrease of BIC, AIC3 and AIC), to confirm the goodness of the solution chosen. In any case, this type of model guarantees final model fit even for choices of the first split parity made irrespective of their statistical fit, as remaining associations are picked up splitting up further down the tree[99].

6.1.8. Model interpretation and fit statistics

As one strength of this model is to permit the inspection of class formation basing on substantial interest and reasoning according to the aim of the research, we presented the results for different fit statistics.

We chose 3 different fit statistics with a different level of penalization: BIC, AIC3 and AIC. The three statistics which aim to balance model fit with parsimony [105-106], are defined in Chapter 3, par.32.2

It is clear that the BIC results in a stronger penalization, followed by AIC3 and finally AIC. We also evaluated the BIC penalized for N equal to the total sample size instead of each 'node' class size, as this is the preferred fit measure for LCT models. The stronger the penalization, the less importance is given to further splits that results in a final tree solution with a smaller number of classes.

Class interpretation at each node is done based on class specific response probabilities as in LCA, but considering the conditioning on its parental nodes.

6.1.9. Assessment of the association between dietary patterns and the risk of selected digestive tract cancer: a 3 Step analysis

The association between the dietary pattern tree identified through LCT and oral/pharyngeal and esophageal cancers risk was assessed through a modified three step analysis proposed by Van den Bergh [66].

After the dietary pattern tree identification (1-step), subjects were classified and classification errors were assessed (2-step). Then, the final step (3-step) consists of relating the class membership and cancer risk while correcting for the classification errors. As proportional assignment was used to build the LCT, subjects were classified according to the same criterion, leading to a 'soft' assignment with a weight for subject equal to posterior membership probability for that class. The maximum likelihood (ML) – based correction was used to correct for the classification errors.

We performed two separate groups of three step analyses, one for each original case-control study datasets, maintaining the classification previously derived from the combined database and therefore basing the likelihood estimation on the whole sample. Each analysis included cases of the selected cancer and their original controls.

The three-step approach was applied at every split of the tree, for each cases-control study, yielding ORs and related 95% CIs comparing classes belonging to the same parental node at each level of the tree. Results were presented for both the unadjusted and the composite adjusted models including terms for age, sex, education, body mass index (BMI), tobacco and alcohol consumption.

Statistical analysis were performed using SAS 9.4 (SAS Institute, Cary, NC, USA), Latent GOLD 5.1 (Vermunt & Magidson,2016) and RStudio 3.5.1(R Foundation for Statistical Computing, Vienna, Austria) statistical software.

6.2. Results

6.2.1. Covariate adjustment and local dependency inspection

Cluster prevalence and food group consumption were conditioned on total non-alcoholic intake in the final models as there were strong association according to the Wald test on the corresponding regression parameters (data not shown).

The BVR statistics showed high correlated errors between sugar and coffee food groups and between pulses and soups food groups. As the FFQ questions on sugar were related to hot beverages and in the construction of food group variables pulses and soup shared an item, we specified correlated errors between coffee and sugar groups and between soup and pulses groups in the final model.

The above mentioned corrections were applied in both LCA and LCT analysis.

6.2.2. Latent Class Analysis solution inspection

LCA performed on the oral/pharyngeal case-control study dataset showed the best fit according to BIC for a 4 class solution (see Chapter 4). One class was characterized by a higher intake of fruit and vegetable and lower intake of red meat. People in the second class reported higher consumption of red meat and lower consumption of fruits and some vegetables. The remaining two classes were characterized by a mixed diet but with a difference in the amount of intake. One class showed lower intake of various food while the other higher intake with respect to the other classes.

LCA performed on the esophageal case-control study dataset showed the best fit according to BIC for a 2 class solution (Supplementary table 6.1). One class exhibited a higher intake of white meat, fruit, fruiting, root and other types of vegetable, while people in the other class reported a low intake of fish, leafy, fruiting and other vegetables, and all types of fruit (Supplementary table 6.2).

LCA performed on the two datasets together showed the best fit according to BIC for a 5 class solution (Supplementary table 6.3). One class was characterized by avoidance of red meat and high consumption of some fruits/vegetables. In contrast, two classes showed higher intake of red meat. The first one showed also high intake of sugary foods. The second one a mixed attitude towards fruit and vegetables: certain of these foods were eaten in high proportion and other in lower proportion comparing to the rest of population. The last two clusters showed a low-intake combination diet, differing from each other in the types of food avoided (Supplementary table 6.4).

6.2.3. Choice of the number of starting classes in LCT

Based on the results of LCA on the single studies, we opted for a 4 class solution as a maximum number of classes for the first split, as it should reasonably represent the principal associations in the whole data. Supplementary Table 6.3 reports the fit statistics and their relative improvement for the whole sample LCA, for models with increased number of classes. While AIC and AIC3 improving measures showed a slower decrease, we notice that for the 4 class solution the relative improvement of the BIC is still relevant, while adding more classes improves the fit, albeit marginally.

6.2.4. Model interpretation and fit statistics

Figure 1 reports the complete structure of the LCT fitted on the data with the class sizes displayed for every level of the tree, according to the three fit measures chosen. According to the BIC statistics (N= total sample or N=node sample) no further splits are needed in addition to the four initial classes. According to AIC3 a further split is needed for each of the four classes, resulting in a final 8 class solution. The AIC statistics allowed further splits ending up in the solution with 13 classes. We labelled only the classes relevant according to all the fit statistics.

Table 2.1 reports the conditional distribution of food group intake giving the latent classes of the first split for the food group more relevant in discriminating and labelling the clusters. A more comprehensive table is given in Supplementary Table 6.5. Cluster 1 labelled ‘Prudent pattern’, showed higher probability to consume more tea, leafy and fruiting vegetables, desserts and lower probability to consume bread.

Subjects in Cluster 2, that we named 'Western pattern', reported higher consumption of red meat and lower consumption of fruiting, leafy cruciferous and other vegetables, citrus fruits and other fruits. We termed Cluster 3 'Lower consumers-combination pattern' as people in it were less likely to consume especially tea, bread, white meat, fish, pulses, leafy and fruiting vegetables, citrus fruit and other types of fruits and in general showed higher proportion of people eating a less than average amount of every food. Cluster 4 had higher probability to eat bread, white meat, red meat, leafy and fruiting vegetables and lower probability to eat desserts, sugary drinks citrus fruit and other fruits, cheese and tea. We called this cluster 'Higher consumers-combination pattern'. Estimated cluster prevalence were 31.5% of the population (n=1322) for the 'Prudent pattern', 29.9% (n=1251) for the 'Western pattern', 19.6% (n=820) for the 'Lower consumers-combination pattern' and 19.1% (n=799) for the 'Higher consumers-combination pattern'.

Table 2.2 and 2.3 report the conditional distribution of food groups intake giving the latent classes of the second level splits for the food groups more relevant in discriminating the clusters. The complete table is given in Supplementary Table 6.6-7. It must be noted that the conditional distribution of foods which did not discriminate between 'child' subclasses, where similar to the their respective distribution in the 'parental' one.

For the second split performed on the Prudent pattern, the conditional distributions of food groups in these 'child' patterns were concordant with their 'parent's one, with the exception of the foods that discriminated the two clusters. Cluster 1.1 showed a lower intake of coffee, pasta, red and processed meat and potatoes. By contrast, people in cluster 1.2 reported higher intake of coffee, pasta, eggs, red and processed meat, fish, pulses and sugar.

For the second split performed on the Western pattern, the conditional distributions of food groups in these 'child' patterns were concordant with their 'parent's one, with the exception of an higher intake of pasta, red and processed meat, potatoes in class 2.1 and a lower intake of coffee, processed meat, fish, cheese, potatoes, root vegetables and desserts in class 2.2.

Food consumption in the child classes of the Lower-intake combination diet pattern, were similar to the parental one, except for a lower intake of tea, soup, root vegetables in class 3.1 and a lower intake of coffee, pasta and red meat in class 3.2.

Food consumption in the child classes of the Higher intake combination pattern, differed from each other and from the parental class, for a high intake of soup, red meat, potatoes and other vegetable and lower consumption of cruciferous vegetables in class 4.1 and a lower intake of potatoes and sugar in class 4.2.

At the third level of the tree, according to AIC only classes 1.1, 1.2, 3.1 and 4.1 were split. Class 1.1.1 was similar to the parental class 1.1 except for a low intake of citrus fruit and higher intake of dessert. Class 1.1.2, differed from the parental class 1.1 and the class 1.1.1 for a higher intake of pulses, cruciferous vegetables, citrus fruit and a lower intake of cheese. Classes 1.2.1 and 1.2.2 discriminated for a different intake of many vegetable and sugar. Class 1.2.1 reported higher consume of leafy, fruiting vegetable and sugar, while class 1.2.2 reported lower intake of the same kind of vegetable and cruciferous ones, with a preference for other types of vegetables (Table 2.4 and Supplementary Table 6.8 as extended version).

Belonging to branch of the tree characterized by low intake of foods, clusters 3.1.1 and 3.1.2 differed for the types of food less consumed. Class 3.1.1 showed low consumption of sugar, desserts, potatoes, bread and red meat and avoidance of sugar drinks, cruciferous and other types of vegetables. Class 3.1.2 reported low intake of white meat, fish, other vegetables, citrus fruit and sugar drinks. People in class 4.1.1 tended

to eat more pasta, processed meat, fish and pulses. In contrast, class 4.1.2 reported lower intake of processed meat and fish (Table 2.5 and Supplementary Table 6.9 as extended version)

According to the AIC fit measure, another split from the branch of the Prudent pattern was needed. People in the child classes of class 1.2.1, showed specific preferences for certain types of vegetables. Class 1.2.1.1 reported high intake of leafy, fruiting vegetable and sugar. People in class 1.2.1.2 preferred root and other types of vegetable, consuming lower amount of leafy, fruiting and cruciferous vegetables (Table 2.6 and Supplementary Table 6.10 as extended version).

Figures 2 and 3 report ORs and corresponding CIs for each cancer by the classification in dietary patterns at each split of the tree, from the composite model including the relevant confounding and risk variables. Compared to the Prudent pattern, the Western one was positively related to oral/Pharyngeal cancer (OR=1.91, 95% CI: 1.41-2.58) and to esophageal cancer (OR=3.22, 95% CI: 1.78 – 5.82). The Lower consumers-combination pattern was positively associated to oral/pharyngeal cancer (OR=2.14, 95% CI: 1.58-2.91) and to esophageal cancer (OR=2.85, 95%CI: 1.47-5.55). No significant association was found between the Higher consumers-combination pattern and oral/pharyngeal cancer (1.04, 95% CI: 0.74-1.46) and esophageal cancer (OR=0.89, 95% CI: 0.39-1.99).

At the second level of the tree, no significant differences were found between the classes in the risk of both types of cancer. At the third level of the split, class 1.1.1 reported a conditional higher risk for both types of cancer (respectively, OR=1.85, 95% CI:1.07-3.19 for oral/pharyngeal cancer and OR=5.37, 95% CI: 1.48-19.44 for esophageal cancer). No other significant differences were found between the other pairs of classes in the third level of the tree in the risk of both cancers. At the fourth level of the tree, the two classes didn't differ significantly for the risk of both types of cancer.

ORs for unadjusted model are given in Supplementary figures 6.1-2.

6.3. Discussion

In the research on dietary patterns, conceived as mutually exclusive groups of people with different dietary habits, in an explorative setting, the objective is to find clusters that describe the data fairly well and that are reasonable and easy to interpret from the epidemiological point of view.

To achieve this goal in LCA, the researcher fits a sequence of models with different numbers of classes and select the one that performs best according to a chosen goodness of fit measure, like BIC or AIC. As these two statistics differ in the level of penalty, the solution obtained are often different and hard to compare.

With large databases sometimes the fit improves until the model has a large number of classes. In these cases, the final solution usually includes also very specific classes that are not relevant for the research.

In our application of LCA on the combined dataset of the two studies on oral/pharyngeal and esophageal cancer, we found the best fit for a solution that was difficult to interpret and included minor differences between clusters. We found two classes characterized by a low intake combination diet that differed only in the type of food avoided most. The other two clusters, characterized by high meat intake were similar. Moreover, the analysis on the single databases resulted in both cases in a clustering solution clearly interpretable and more interesting from the epidemiological point of view.

The Latent Tree model, developed by Van den Bergh et al.[66,99-100] is a possible solution to these problems. LCT is a sequential algorithm similar to those used in hierarchical cluster analysis, but that maintains all the properties and the strength of LCA. The main advantage of this procedure is that it gives a clear insight on how the cluster are related and makes it possible to compare solutions with different numbers of classes.

In our fitting of the LCT, the choice of one of the selected three fit measures we considered resulted in a different length of the three. The two BIC measures are usually preferred in LCA; as they are the most penalizing they restricted the tree to only the first split. Looking at the LCA performed on the two databases we saw that a 4 class solution showed the best fit for the OPC study while a 2 class solution was the best solution for ESCC according to BIC. Therefore, in the traditional way to perform LCA and LCT, a 4 class solution describe properly the main part of variability of the combined datasets. In fact, with the traditional analysis one generally wants the lowest number of classes that can adequately describe the heterogeneity and additional classes can complicate interpretation and further analysis (e.g. assessment of the association between patterns and health outcomes).

At the first level of the tree, we found a pattern characterized by high intake of leafy and fruiting vegetable and fruits (‘Prudent pattern’), a pattern with a high intake of red meat and low intake of specific fruits and vegetables (‘Western pattern’) and two patterns which showed a combination-type of diet. The first ‘combination’ pattern showed a low intake of the majority of foods (‘Lower consumers-combination pattern’), and the other high intake of varying foods (‘Higher consumers-combination pattern’). These classes were also related to the principal differences in terms of both cancers risk.

As in standard LCA, sometimes it is reasonable to rely not only on statistical criteria but also to inspect further solutions and to verify if a split is meaningful for the research at hand. A more ‘relaxed’ fit measure like AIC3 or AIC can be useful in permitting a deeper inspection of class formation and to think more theoretically on the meanings of classes.

In this work, classes from the second split onward, did not differ in the risk of both cancer types with just one exception. It has to be noted that the majority of these classes were not pairs that discriminate completely for certain foods (e.g. avoidance of one food vs high intake of the same food), instead presented pairs with one class reporting high preference/disfavour for selected foods and the other one medium intake of the same foods and vice versa. Moreover, the contrasts between pairs were defined by many different food, making it difficult to clearly identify ‘healthy/unhealthy’ groups.

Nevertheless, we found a significant difference between classes 1.1.1 and 1.1.2 in the risk of both types of cancer. Apart from specific preferences in each classes that didn’t find any opposite correspondence in the other, the two classes markedly differed for the intake of citrus fruit. Therefore, the effect can be interpreted as follows. Conditioning on being in the Prudent pattern (which showed the lowest risk of both types of cancer), and also in the subgroup which eat less pasta, potatoes, coffee, red and processed meat, consuming low quantities of citrus fruit instead of high quantities, is associated with an increased risk for both cancers.

Our results were comparable to the evidence coming from the studies on the influence of *a posteriori* dietary patterns on OPC and ESCC. Patterns related to high intake of fruits and vegetables were mostly inversely associated with the risk of OPC [37-44] and ESSC [43,45-48,50-53]. Patterns related to high intake of meat were often found in both studies of OPC [37-44] and ESCC [43,45-48,50-53]. These kinds of patterns were also associated with other types of foods that varied in the different studies, therefore the

association of these patterns with the two types of cancer was less consistent the studies. Patterns related to a combination of different foods in these studies varied, often according to country specific diets and consequently the results of their association with the two types of cancer [37-38,40,42-46,50,52-54]. Evidence on the benefits of the intake of citrus fruit with regard to both cancer sites is also recognized in the literature [101-103].

This study has some limitations. As we analyzed only Italian data, having a population that generally shares a common dietary behaviour may be the cause that the most important dietary choices were mostly described by the first initial split, and further splits identified very specific differences that mostly were not very relevant from the clinical point of view. For example, this may be a reason why we found 4 classes in the LCA on the OPC study that was conducted in northern and central Italy and only 2 classes for the LCA on the ESCC study which was conducted just in northern Italy. Moreover, the small size of certain groups is also a signal of very specific dietary groups. A more complex study with a stronger variability in terms of population, and consequently, of dietary habits would probably offer an interesting object for LCT analysis.

LCA is not so common in dietary pattern research as factor, principal component and cluster analysis. A strength of this study is that it proposed the first application, to our knowledge, of LCT to this field.

In summary, LCA can be a powerful tool in this field of research with many advantages with respect to classical methods. In cases of difficulty in the interpretation of the solution obtained with the standard methodology, we propose LCT as an interesting instrument to inspect classes formation and mutual relation in the research of dietary patterns.

6.4. Tables and Figures

Figure 1.Layout of the LCT according to BIC, AIC3, AIC with number of cases per groups, Italy 1992-2009.

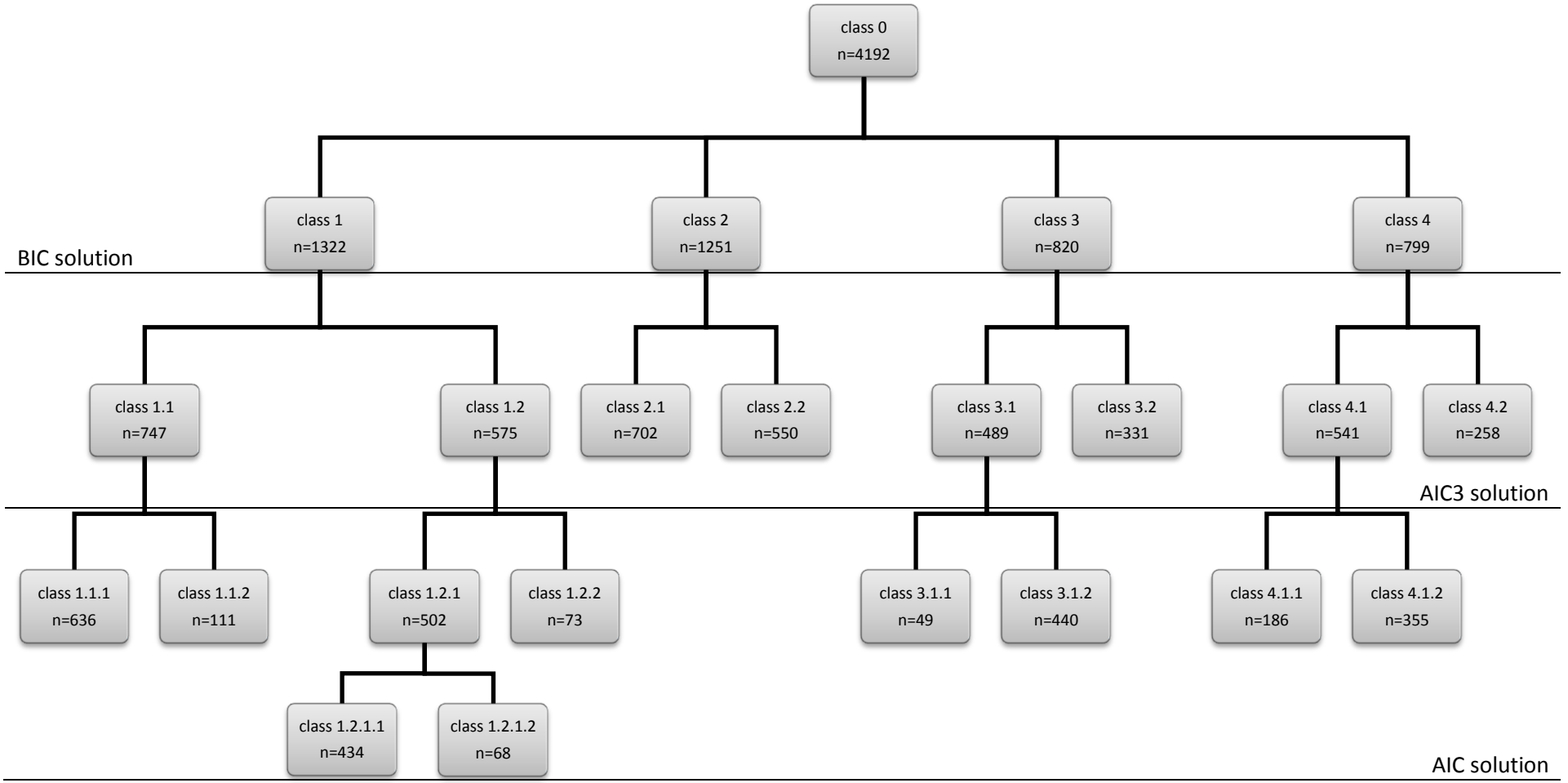


Table 1 Distribution of cases of OPC and ESCC and controls by selected covariates. Italy,1992-2009.

		OPC N(%)		ESCC N(%)	
		cases	controls	cases	controls
Age	<50	190 (20.1)	583 (23.39)	29 (9.5)	78 (10.50)
(years)	50-59	313 (33.1)	734 (29.45)	112 (36.8)	266 (35.80)
	60-69	329 (34.8)	837 (33.59)	119 (39.1)	288 (38.76)
	>69	114 (12.1)	338 (13.56)	44 (14.5)	111 (14.94)
Sex	Male	756 (79.9)	1497 (60.07)	275 (90.5)	593 (79.81)
	female	190 (20.1)	995 (39.93)	29 (9.5)	150 (20.19)
Education	<7	558 (59.0)	1283 (51.48)	217 (71.4)	456 (61.37)
(years)	7-11	260 (27.5)	726 (29.13)	65 (21.4)	189 (25.44)
	>11	128 (13.5)	483 (19.38)	22 (7.2)	98 (13.19)
Alcoholic	0	86 (9.1)	546 (21.91)	5 (1.6)	70 (9.42)
intake	1-6	47 (5.0)	307 (12.32)	3 (1.0)	52 (7.00)
(weekly	7-13	55 (5.8)	356 (14.29)	6 (2.0)	89 (11.98)
units)	14-20	125 (13.2)	452 (18.14)	21 (6.9)	130 (17.50)
	>20	633 (66.9)	831 (33.35)	269 (88.5)	402 (54.10)
Smoking	Never smoked	137 (14.5)	1079 (43.30)	33 (10.9)	245 (32.97)
Habit	Ex smoker	269 (28.4)	764 (30.66)	109 (35.9)	287 (38.63)
	<15 sig/d	128 (13.5)	287 (11.52)	39 (12.8)	86 (11.57)
	≥15 sig/d	412 (43.6)	362 (14.53)	123 (40.5)	125 (16.82)
BMI	<18.5	43 (4.6)	37 (1.48)	7 (2.30)	6 (0.81)
	18.6-25.9	601 (63.5)	1304 (52.33)	170 (55.92)	361 (48.59)
	26-29.9	223 (23.6)	809 (32.46)	91 (29.93)	283 (38.09)
	>29.9	79 (8.4)	342 (13.72)	36 (11.84)	93 (12.52)
N		946	2492	304	743

Table2.1 Probabilities of consumption for selected food groups by dietary patterns derived from LCT. First level split, nodes 1, 2, 3, 4.

		Class 1 Prudent %	Class 2 Western %	Class 3 Lower consumers-combination %	Class 4 Higher consumers-combination %
Size %		31.5	29.9	19.6	19.1
tea	Not consumed	38.0	40.4	**63.5**	**63.4**
	Consumed	**62.0**	59.6	36.5	36.6
bread	Below median	**62.6**	44.4	**60.1**	28.6
	Above median	37.4	55.6	39.9	**71.4**
white	Below median	45.8	54.0	**65.8**	38.4
meat	Above median	54.2	46.0	34.2	**61.6**
red	Below median	59.7	37.7	58.1	34.1
meat	Above median	40.4	**62.4**	41.9	**65.9**
fish	Below median	45.4	55.0	**72.1**	55.8
	Above median	54.6	45.0	27.9	44.2
cheese	Below median	43.7	55.5	59.3	**61.9**
	Above median	56.3	44.5	40.7	38.1
pulses	Below median	43.8	57.6	**63.2**	44.6
	Above median	56.2	42.4	36.8	55.4
leafy	Below median	37.7	55.3	**76.7**	25.3
vegetables	Above median	**62.4**	44.7	23.3	**74.7**
fruiting	Below median	34.8	**84.0**	**79.1**	38.7
vegetables	Above median	**65.2**	16.0	20.9	**61.3**
root	Not consumed	10.7	13.0	59.3	26.9
vegetables	Below median	35.3	**61.2**	33.8	33.6
	Above median	54.1	25.9	7.0	39.5
cruciferous	Not consumed	19.4	16.3	51.1	20.9
vegetables	Below median	52.6	**79.5**	39.2	56.3
	Above median	28.0	4.2	9.7	22.8
other	Not consumed	6.9	1.0	41.1	7.1
vegetables	Below median	38.3	**73.1**	54.0	33.5
	Above median	54.8	25.9	4.9	59.5
citrus	Not consumed	5.1	5.8	25.4	15.2
fruit	Below median	52.8	**79.7**	**61.8**	**72.3**
	Above median	42.1	14.5	12.9	12.5
other	Below median	40.4	**70.8**	**78.0**	**63.2**
fruits	Above median	59.6	29.2	22.0	36.9
sugary	Not consumed	52.1	43.5	59.7	**71.1**
drinks	Consumed	47.9	56.5	40.4	28.9
desserts	Below median	34.8	55.9	53.8	**73.2**
	Above median	**65.2**	44.1	46.2	26.8

Table2.2 Probabilities of consumption for selected food groups by dietary patterns derived from LCT. Second level splits, nodes 1.1,1.2 and 2.1, 2.2

		Parental class: 1		Parental class: 2	
		Class 1.1 %	Class 1.2 %	Class 2.1 %	Class 2.2 %
Size %		56.5	43.5	56.1	43.9
coffee	Below median	**64.5**	39.3	45.5	**60.3**
	Above median	35.5	**60.7**	54.5	39.7
pasta	Below median	**62.4**	39.2	37.9	56.6
	Above median	37.6	**60.8**	**62.1**	43.4
eggs	Not consumed	15.0	7.8		
	Below median	49.7	30.2		
	Above median	35.3	**62.0**		
red	Below median	**75.3**	39.8	23.7	55.3
meat	Above median	24.7	**60.2**	**76.3**	44.7
processed	Below median	**70.8**	29.6	37.6	**73.4**
meat	Above median	29.2	**70.4**	**62.4**	26.6
fish	Below median	55.4	32.5	45.7	**66.8**
	Above median	44.7	**67.5**	54.3	33.2
cheese	Below median			47.1	**66.2**
	Above median			52.9	33.8
potatoes	Below median	**68.6**	44.5	38.2	**75.8**
	Above median	31.4	55.6	**61.9**	24.2
pulses	Below median	51.1	34.4		
	Above median	48.9	**65.6**		
root	Not consumed			15.1	10.3
vegetables	Below median			52.6	**72.1**
	Above median			32.3	17.6
sugar	Below median	52.0	39.1		
	Above median	48.0	**60.9**		
desserts	Below median			47.1	**67.2**
	Above median			52.9	32.8

Table2.3 Probabilities of consumption for selected food groups by dietary patterns derived from LCT. Second level splits, nodes 3.1,3.2 and 4.1, 4.2

		Parental class: 3		Parental class: 4	
		Class 3.1 %	Class 3.2 %	Class 4.1 %	Class 4.2 %
Size %		59.6	40.4	67.7	32.3
coffee	Below median	41.9	**80.2**		
	Above median	58.1	19.8		
tea	Not consumed	**79.2**	40.4		
	Consumed	20.8	59.7		
pasta	Below median	45.3	**61.6**		
	Above median	54.7	38.5		
soups	Below median	**63.6**	49.9	36.3	58.6
	Above median	36.4	50.1	**63.7**	41.4
red	Below median	49.9	**70.6**	23.4	56.3
meat	Above median	50.1	29.4	**76.6**	43.8
potatoes	Below median			30.1	**69.2**
	Above median			**69.9**	30.9
root	Not consumed	**65.7**	49.8		
vegetables	Below median	28.4	41.6		
	Above median	5.9	8.6		
cruciferous	Not consumed			20.2	22.4
vegetables	Below median			**61.5**	45.6
	Above median			18.4	32.1
other	Not consumed			6.2	8.9
vegetables	Below median			24.8	51.7
	Above median			**69.0**	39.4
sugar	Below median			47.4	**72.2**
	Above median			52.6	27.8

Table2.4 Probabilities of consumption for selected food groups by dietary patterns derived from LCT. Third level splits, nodes 1.1.1, 1.1.2 and 1.2.1 ,1.2.2

		Parental class: 1.1		Parental class: 1.2	
		Class 1.1.1 %	Class 1.1.2 %	Class 1.2.1 %	Class 1.2.2 %
Size %		85.1	14.9	87.3	12.7
cheese	Below median	46.7	**65.1**		
	Above median	53.3	34.9		
pulses	Below median	56.4	21.0		
	Above median	43.6	**79.0**		
leafy	Below median			32.4	**68.3**
vegetables	Above median			**67.6**	31.7
fruiting	Below median			26.9	**80.4**
vegetables	Above median			**73.1**	19.6
other	Not consumed			7.5	0.1
vegetables	Below median			40.9	17.9
	Above median			51.6	**82.0**
cruciferous	Not consumed	26.7	4.6	15.1	8.3
vegetables	Below median	58.6	25.6	45.3	**90.8**
	Above median	14.7	**69.8**	39.6	0.9
citrus	Not consumed	7.6	2.1		
fruits	Below median	**60.5**	34.8		
	Above median	31.9	**63.2**		
desserts	Below median	34.8	56.6		
	Above median	**65.3**	43.4		
sugar	Below median			37.4	50.6
	Above median			**62.6**	49.4

Table2.5 Probabilities of consumption for selected food groups by dietary patterns derived from LCT. Third level splits, nodes 3.1.1,3.1.2 and 4.1.1, 4.1.2

		Parental class: 3.1		Parental class: 4.1	
		Class 3.1.1 %	Class 3.1.2 %	Class 4.1.1 %	Class 4.1.2 %
Size %		10.0	90.0	90.0	10.0
bread	Below median	**84.0**	52.2	52.2	**84.0**
	Above median	16.0	47.8	47.8	16.0
pasta	Below median				
	Above median				
white	Below median	43.6	**71.0**	**71.0**	43.6
meat	Above median	56.5	29.0	29.0	56.5
red	Below median	**79.5**	46.6	46.6	**79.5**
meat	Above median	20.5	53.4	53.4	20.5
fish	Below median	50.8	**70.7**		
	Above median	49.2	29.3		
potatoes	Below median	**70.7**	55.1		
	Above median	29.3	44.9		
cruciferous	Not consumed	**73.8**	45.4		
vegetables	Below median	11.8	43.4		
	Above median	14.5	11.2		
other	Not consumed	**92.1**	31.9		
vegetables	Below median	7.4	**62.2**		
	Above median	0.5	5.9		
sugar	Not consumed	**90.5**	**60.9**	**60.9**	**90.5**
drinks	Consumed	9.5	39.1	39.1	9.5
desserts	Below median	**77.5**	56.1	56.1	**77.5**
	Above median	22.5	43.9	43.9	22.5
sugar	Below median	**74.4**	52.2	52.2	**74.4**
	Above median	25.6	47.9	47.9	25.6

Table 2.6 Probabilities of consumption for selected food groups by dietary patterns derived from LCT. Fourth level split, nodes 1.2.1.1, 1.2.1.2.

		Parental class: 1.2.1	
		Class 1.2.1.1 %	Class 1.2.1.2 %
Size %		86.5	13.6
leafy	Below median	33.0	**69.5**
vegetables	Above median	**67.0**	30.5
fruiting	Below median	28.5	**78.5**
vegetables	Above median	**71.5**	21.5
root	Not consumed	13.9	0.4
vegetables	Below median	33.9	8.4
	Above median	52.2	**91.2**
cruciferous	Not consumed	15.5	8.8
vegetables	Below median	47.5	**90.4**
	Above median	37.0	0.8
other	Not consumed	7.9	0.1
vegetables	Below median	42.2	17.0
	Above median	49.9	**82.9**
sugar	Below median	34.1	54.2
	Above median	**65.9**	45.8

Figure 2. Odds ratios (OR) and corresponding 95% confidence intervals (CIs) for oral/pharyngeal cancer risk at each split. Models were adjusted for sex, age, education, BMI, tobacco and alcohol intake. Italy,1992-2009. [a]Reference category for the split.

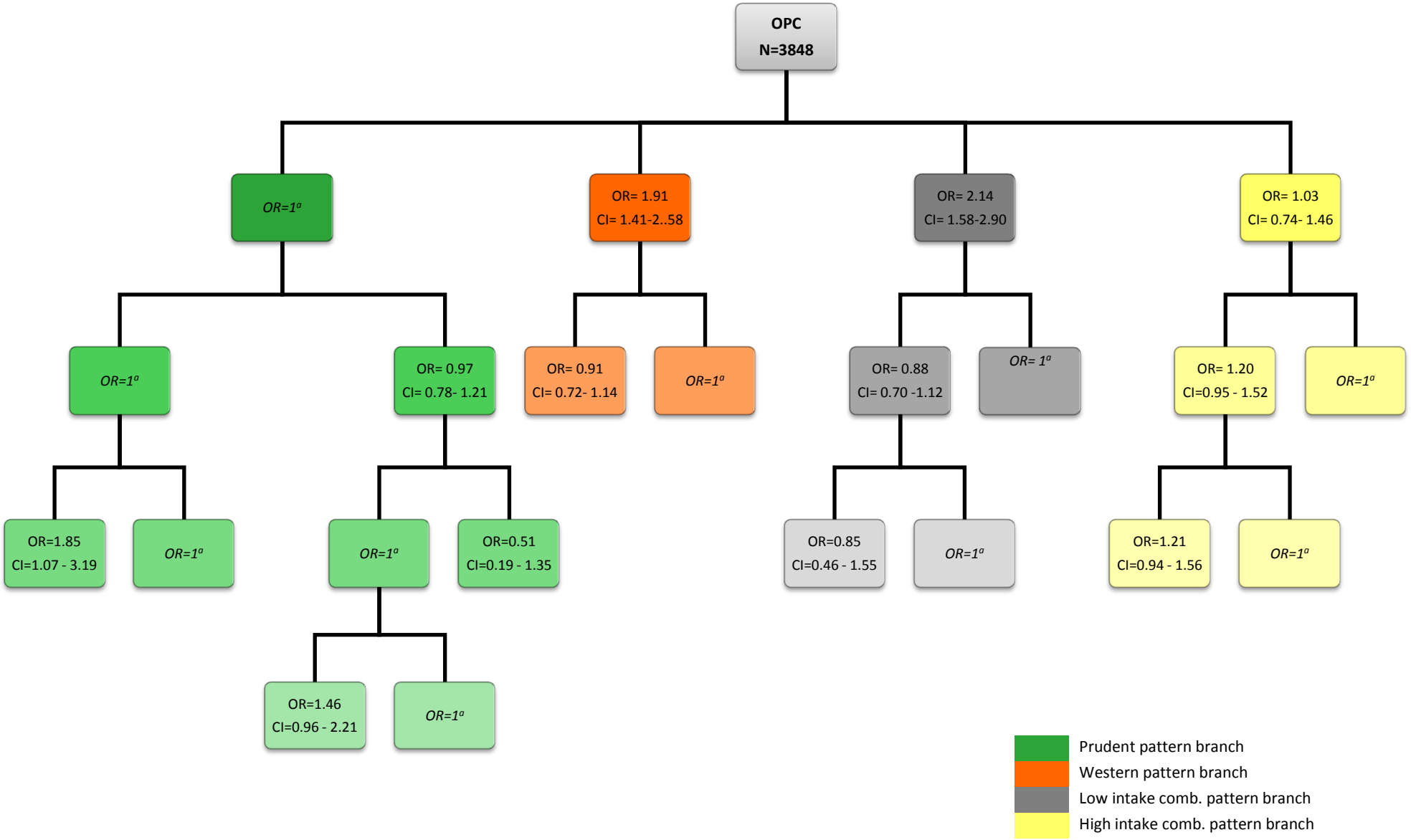
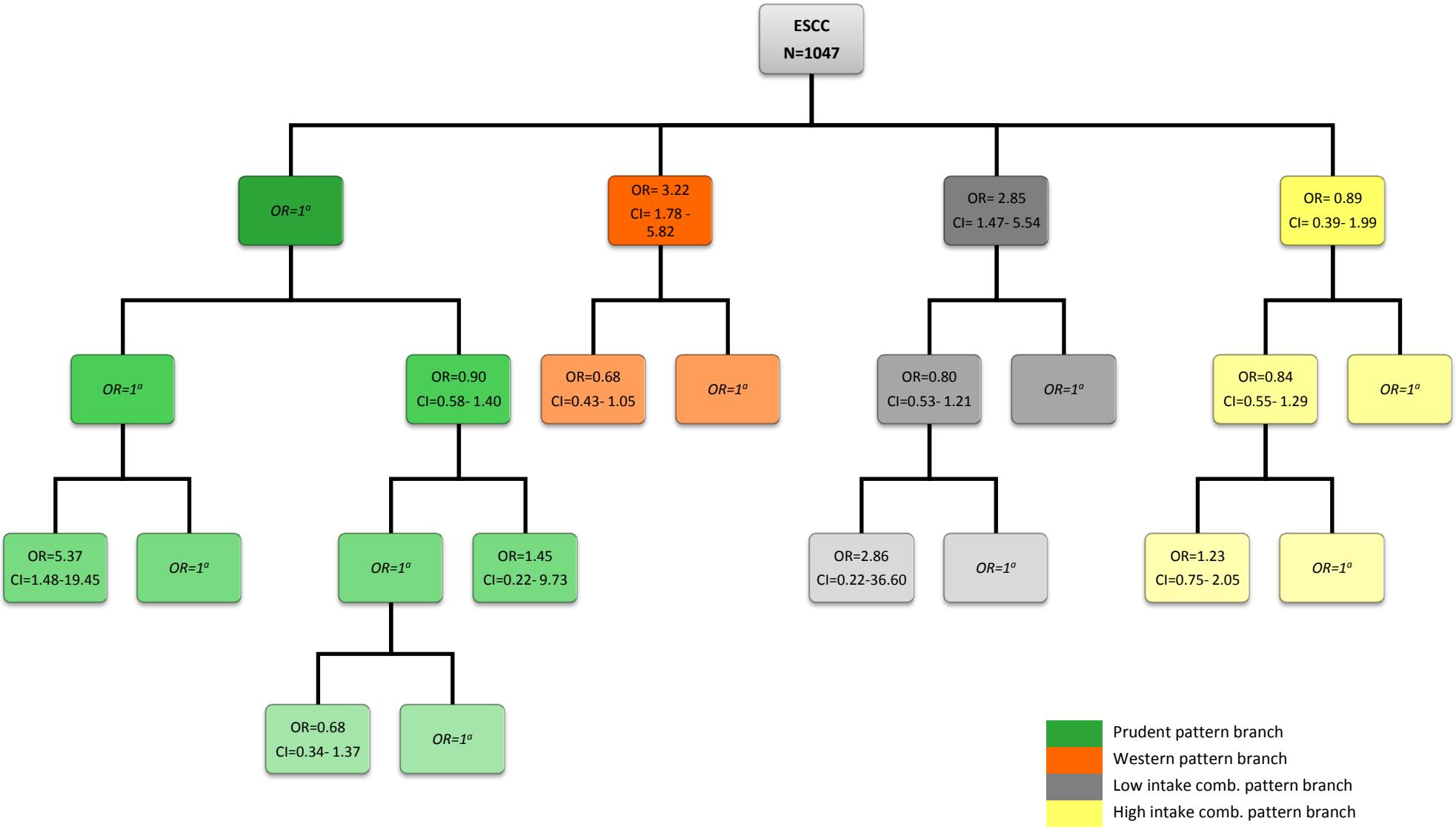


Figure 2. Odds ratios (OR) and corresponding 95% confidence intervals (CIs) for esophageal cancer risk at each split. Models were adjusted for sex, age, education, BMI, tobacco and alcohol intake. Italy,1992-1997. [a]Reference category for the split.



7. CONCLUSIONS

7.1. General conclusions

The scope of this dissertation is contextualized in the frame of dietary patterns research and particularly, on the assessment of the relation between dietary patterns and the risk of selected types of cancer. A Latent Class solution was proposed as an alternative to the traditionally used empirical methods such as factor, principal component and cluster analysis, and differences and advantages with respect to them were also presented.

The topics faced in this dissertation focused on three main issues. First, dietary patterns identification using Latent Class Analysis was targeted, followed by assessment of their influence on oral/pharyngeal cancer risk. Second, the robustness of the identified dietary patterns to total non-alcoholic energy intake adjustment was investigated. Finally, a new Latent Class approach, named Latent Class Tree, was presented, as a tool to help classes interpretation and analysis at different levels of details.

7.1.1. Dietary patterns and the risk of oral and pharyngeal cancer

Using data from a multicentric case-control study on OPC carried out between 1992 and 2009 which collected information on diet through a food frequency questionnaire, we found 4 dietary patterns, conceived as mutually exclusive groups of people which shared common dietary behaviour within groups. The first pattern, labelled 'Prudent pattern', showed higher probability to consume more leafy and fruiting vegetables, citrus fruit and all other kinds of fruits, tea and lower probability to consume red meat. The second pattern, that we named 'Western pattern', reported higher consumption of red meat and lower consumption of fruits, cruciferous and fruiting vegetables. We termed the third pattern 'Lower consumers-combination pattern' as people in it were less likely to eat fruits, leafy and fruiting vegetables, pulses, potatoes, fish, white and red meat, bread and tea/decaffeinated coffee. The last pattern had higher probability to eating fruiting, leafy and other vegetables, white and red meat and bread, while showed a lower probability to consume coffee, tea, processed meat, cheese, fish, sugary drinks and desserts. We called this last pattern 'Higher consumers-combination pattern'. Dietary patterns were adjusted for total non-alcoholic energy intake and correlation between certain foods item (sugar-coffee, soups-pulses) was allowed in class identification. Compared to the Prudent pattern, the Western and the Lower consumers-combination ones were positively related to the risk of OPC (OR=2.56, 95% CI: 1.90 – 3.45 and OR=2.23, 95% CI: 1.64 – 3.02). Higher consumers-combination pattern did not differ significantly from the Prudent pattern (OR=1.28, 95% CI: 0.92 – 1.77).

7.1.2. Energy intake adjustment in dietary pattern research using Latent Class Analysis

Using data from a multicentric case-control study on OPC carried out between 1992 and 2009 which collected information on diet through a food frequency questionnaire, we identified and compared dietary patterns adjusting or not for total energy intake in the class identification phase of the analysis. Three possible ways to correct for total energy intake in class identification were presented. In general unadjusted and adjusted solutions were comparable. The main difference was related to the patterns that showed highest/lowest non-alcoholic energy intake, that resulted in a variation of number of classes (5/7/4 patterns for the different adjusted solutions and 5 patterns for the unadjusted one).

Then, to determine the effect of adjustment in predicting an health outcome, we compared the effect of unadjusted dietary patterns, unadjusted dietary patterns with non-alcoholic energy intake also included in the model as a confounder, and adjusted dietary patterns on the risk of OPC . Differences in the estimations

for the distinct solutions were found when ORs were not corrected for known/potential risk factors. In general, adjustments for non-alcoholic energy intake results in a mitigation of the effects, thus remaining in the same order. When adjusting for known/potential risk factors, estimations of ORs and related CI remained consistent in all the models we fitted.

In the end, specific suggestions on how to perform energy correction in dietary patterns research using LCA are delivered, based on the results of the current analysis.

7.1.3. Dietary inspection through Latent Class Tree

In our application of LCA on the combined dataset of the two studies on oral/pharyngeal and esophageal cancer (Italy, 1992-2009), we found the best fit for a solution that was difficult to interpret and included minor differences between clusters. To address these issues Latent Class Tree method was applied. Three fit statistics (AIC, AIC3, BIC) were used for their different level of penalty that resulted in different lengths of the tree. For the first split we allowed for a 4-class solution which identified a pattern characterized by high intake of leafy and fruiting vegetable and fruits ('Prudent pattern'), a pattern with a high intake of red meat and low intake of certain fruits and vegetables ('Western pattern') and two patterns which showed a combination-type of diet. The first 'combination' pattern showed a low intake of the majority of foods ('Lower consumers-combination pattern'), and the other high intake of varying foods ('Higher consumers-combination pattern'). Compared to the Prudent pattern, the Western one was positively related to oral/Pharyngeal cancer (OR=1.91, 95% CI: 1.41-2.58) and to esophageal cancer (OR=3.22, 95% CI: 1.78 – 5.82). The Lower consumers-combination pattern was positively associated to oral/pharyngeal cancer (OR=2.14, 95% CI: 1.58-2.91) and to esophageal cancer (OR=2.85, 95%CI: 1.47-5.55). No significant association was found between the Higher consumers-combination pattern and oral/pharyngeal cancer (1.04, 95% CI: 0.74-1.46) and esophageal cancer (OR=0.89, 95% CI: 0.39-1.99). In the 'Prudent pattern' branch of the tree, we found two classes that differed in the risk of both cancer types only at subsequent splits. The two classes differed mainly for the intake of citrus fruit, showing respectively, OR=1.85, 95% CI:1.07-3.19 for oral/pharyngeal cancer and OR=5.37, 95% CI: 1.48-19.44 for esophageal cancer for the class that reported low intake of citrus fruit with respect to the class which exhibits a high intake of citrus fruit. No other significant differences were found between the other pairs of classes at any other level of the tree.

In conclusion, we presented LC methods as powerful tools to characterize eating habits of a population and to associate diet with specific health outcomes. These methods have some advantages that can address important issues in dietary pattern research, like, for example, pattern prevalence estimation, energy intake adjustment in pattern identification, and class formation inspection and comparison between different solutions though Latent Class Tree.

7.2. Future works

In this thesis an application of some methodologies belonging to the LC approach were proposed, addressing the issue of dietary patterning and its relation to the incident of certain type of cancer. Based on the results and the knowledge achieved during this work, some extensions of the ideas presented in this thesis can be proposed.

We here presented contributions using LCA with food groups as indicators in each analysis. Food groups can be correctly conceived as categorical (nominal/ordinal) items. When the distribution of food intake is extremely skewed, due to an high peak in 0 that characterized the not consumers as it was in our datasets,

the categorization of food intake is often the best choice. Keeping a separate category for not consumers can be a decision due to maintaining 'natural' groups in the population (e.g. vegetarians vs meat eaters). Moreover, the transformation of the variables to obtain a symmetrical distribution, like the logarithmic transformation, may sometimes not be optimal, especially in presence of a high peak in correspondence to 0. The usual way to treat this problem is to add a certain constant to food intakes (usually 1 in dietary variables), but the choice of the constant has a strong influence on the distribution of the transformed variable due to a different weight given to zeros. Moreover, food groups can be thought as categorical in nature, as they are a collection of different foods intakes.

FA cannot be applied to categorical variables without a risk of biased estimations. Therefore, for the above mentioned reason, this method may not be appropriate for identifying dietary patterns on food group indicators. In contrast, traditional LCA is based on categorical indicators, and further extensions permit dealing with different scales of indicators, in the general framework of finite mixture models.

Dealing with nutrients intake, instead of food groups, gives less problems related to the distribution of the variables. FA can be fairly applied on nutrients indicators, just taking into account the different scale of macro-micro nutrients and minor problems of skewness.

As a proof of concept, we fitted a conditional Gaussian mixture model for dietary pattern on the oral/pharyngeal cancer database using 28 selected micro-macro nutrients as indicators. Analysis were carried out both maintaining the original scale of the variables (LCA is scale invariant) and with logarithmic transformation for skewness correction of the indicators. We first fitted the trivial 1-class model and subsequently increased the parity. Using the BIC as a fit criterion, we noticed that the more classes we allowed in the model, the lower was the value of the BIC. That is to say, we didn't find a good fit for a reasonable grouping. One possible reason of this may derive from a violation of the local independence assumption, that is indicators should not be correlated given the classes. An effect of the violation of this assumption is exactly the formation of further groups. Considering the strong pattern of intercorrelation that exists among nutrients, we could hypothesize that this makes them more appropriate for a FA which is based on the correlation between indicators. On the opposite, in LCA the pattern of correlations existing among the items may be too extreme to be solved allowing a reasonable set of local dependencies, leading to difficulty in identifying distinct groups of people.

When it is possible to conceive groups of people with strong dietary choices, in preference of certain foods and in avoidance of others, this kind of discrimination may be weaker in terms of nutrients. People with different dietary habits may not have that clearly separated patterns of nutrients. Moreover, this huge number of classes may suggest that the latent variable underneath the indicators may be more appropriately described by a continuous one than a categorical one like it is assumed in LCA. This aspect deserves a further deepening, eventually with simulations.

Moreover, in the study of nutrient-based dietary patterns, another suggestion for future research can be given. As it was shown, FA and LCA can offer two different perspectives on the research of the association of dietary habits and certain diseases, an interesting option for future research could be combining the two methods. That is first applying FA on nutrients, explaining which nutrients are taken together, then fitting LCA on the factor scores obtained to classify the individuals in mutually exclusive groups.

A last suggestion for future work is related to one of the major limits of this study. The data we analyzed regard only Italian people. Differences in dietary habits may for this reason not be so relevant as we noticed when exploring in depth class formation with LCT. LCA is a method very sensitive in detecting heterogeneity

in the data, that is why we found differences in eating patterns in some different case-control studies. It would be more interesting to apply the model when the major part of the variability belongs to strong structural differences in eating patterns rather than minor differences belonging to a more homogeneous population. For this reason, our last proposal for future research could be the application of LC methods to data coming from different countries with a stronger variation in terms of diet, to compare the resulting dietary patterns in term of the risk of cancer.

REFERENCES

[1] Hu, F. B. (2002). Dietary pattern analysis: a new direction in nutritional epidemiology. Current opinion in lipidology, 13(1), 3-9.

[2] Fahey, M. T., Thane, C. W., Bramwell, G. D., & Coward, W. A. (2007). Conditional Gaussian mixture modelling for dietary pattern analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society), 170(1), 149-166.

[3] Edefonti, V., Bravi, F., La Vecchia, C., Randi, G., Ferraroni, M., Garavello, W., ... & Decarli, A. (2010). Nutrient-based dietary patterns and the risk of oral and pharyngeal cancer. Oral oncology, 46(5), 343-348.

[4] Bravi, F., Edefonti, V., Randi, G., Garavello, W., La Vecchia, C., Ferraroni, M., ... & Decarli, A. (2011). Dietary patterns and the risk of esophageal cancer. Annals of oncology, 23(3), 765-770.

[5] Willett, W. (2012). Nutritional epidemiology. Oxford University Press.

[6] Affret, A., Severi, G., Dow, C., Rey, G., Delpierre, C., Boutron-Ruault, M. C., ... & Fagherazzi, G. (2017). Socio-economic factors associated with a healthy diet: results from the E3N study. Public health nutrition, 20(9), 1574-1583.

[7] Albuquerque, G., Lopes, C., Durão, C., Severo, M., Moreira, P., & Oliveira, A. (2016). Dietary patterns at 4 years old: Association with appetite-related eating behaviours in 7 year-old children. Clinical Nutrition.

[8] Sotres-Alvarez, D., Siega-Riz, A. M., Herring, A. H., Carmichael, S. L., Feldkamp, M. L., Hobbs, C. A., ... & National Birth Defects Prevention Study. (2013). Maternal dietary patterns are associated with risk of neural tube and congenital heart defects. American journal of epidemiology, 177(11), 1279-1288.

[9] Bezerra, I. N., Bahamonde, N. M. S. G., Marchioni, D. M. L., Chor, D., de Oliveira Cardoso, L., Aquino, E. M., ... & de Matos, S. M. A. (2018). Generational differences in dietary pattern among Brazilian adults born between 1934 and 1975: a latent class analysis. Public health nutrition, 1-12.

[10] Casini, L., Contini, C., Marone, E., & Romano, C. (2013). Food habits. Changes among young Italians in the last 10 years. Appetite, 68, 21-29.

[11] Durao, C., Severo, M., Oliveira, A., Moreira, P., Guerra, A., Barros, H., & Lopes, C. (2017). Association between dietary patterns and adiposity from 4 to 7 years of age. Public health nutrition, 20(11), 1973-1982.

[12] Harrington, J. M., Dahly, D. L., Fitzgerald, A. P., Gilthorpe, M. S., & Perry, I. J. (2014). Capturing changes in dietary patterns among older adults: a latent class analysis of an ageing Irish cohort. Public health nutrition, 17(12), 2674-2686.

[13] Hohman, E. E., Paul, I. M., Birch, L. L., & Savage, J. S. (2017). INSIGHT responsive parenting intervention is associated with healthier patterns of dietary exposures in infants. Obesity, 25(1), 185-191.

[14] Leech, R. M., Timperio, A., Livingstone, K. M., Worsley, A., & McNaughton, S. A. (2017). Temporal eating patterns: associations with nutrient intakes, diet quality, and measures of adiposity. The American journal of clinical nutrition, 106(4), 1121-1130.

[15] Martin, C. L., Siega-Riz, A. M., Sotres-Alvarez, D., Robinson, W. R., Daniels, J. L., Perrin, E. M., & Stuebe, A. M. (2016). Maternal Dietary Patterns during Pregnancy Are Associated with Child Growth in the First 3 Years of Life–3. The Journal of nutrition, 146(11), 2281-2288.

[16] Martin, C. L., Siega-Riz, A. M., Sotres-Alvarez, D., Robinson, W. R., Daniels, J. L., Perrin, E. M., & Stuebe, A. M. (2016). Maternal dietary patterns are associated with lower levels of cardiometabolic markers during pregnancy. Paediatric and perinatal epidemiology, 30(3), 246-255.

[17] Mueller, M. P., Anzman-Frasca, S., Blakeley, C. E., Folta, S. C., Wilde, P., & Economos, C. D. (2017). Ordering patterns following the implementation of a healthier children's restaurant menu: A latent class analysis. Obesity, 25(1), 192-199.

[18] Padmadas, S. S., Dias, J. G., & Willekens, F. J. (2006). Disentangling women's responses on complex dietary intake patterns from an Indian cross-sectional survey: a latent class analysis. Public Health Nutrition, 9(2), 204-211.

[19] Rose, C. M., Savage, J. S., & Birch, L. L. (2016). Patterns of early dietary exposures have implications for maternal and child weight outcomes. Obesity, 24(2), 430-438.

[20] Saneei, P., Esmaillzadeh, A., Keshteli, A. H., Feizi, A., Feinle-Bisset, C., & Adibi, P. (2016). Patterns of dietary habits in relation to obesity in Iranian adults. European journal of nutrition, 55(2), 713-728.

[21] Uzhova, I., Woolhead, C., Timon, C. M., O'Sullivan, A., Brennan, L., Peñalvo, J. L., & Gibney, E. R. (2018). Generic Meal Patterns Identified by Latent Class Analysis: Insights from NANS (National Adult Nutrition Survey). Nutrients, 10(3), 310.

[22] Noori, M. A., Ghiasvand, R., Maghsoudi, Z., Feizi, A., Esmaillzadeh, A., Adibi, P., & Keshteli, A. H. (2016). Evaluation of dietary pattern stability and physical activity in three consecutive generations of women. International journal of public health, 61(1), 29-38.

[23] Sotres-Alvarez, D., Herring, A. H., & Siega-Riz, A. M. (2010). Latent Class Analysis Is Useful to Classify Pregnant Women into Dietary Patterns–3. The Journal of nutrition, 140(12), 2253-2259.

[24] De Vries, H., van't Riet, J., Spigt, M., Metsemakers, J., van den Akker, M., Vermunt, J. K., & Kremers, S. (2008). Clusters of lifestyle behaviors: results from the Dutch SMILE study. Preventive medicine, 46(3), 203-208.

[25] Durão, C., Severo, M., Oliveira, A., Moreira, P., Guerra, A., Barros, H., & Lopes, C. (2017). Association of maternal characteristics and behaviours with 4-year-old children's dietary patterns. Maternal & child nutrition, 13(2), e12278.

[26] Esmaillzadeh, A., Keshteli, A. H., Feizi, A., Zaribaf, F., Feinle-Bisset, C., & Adibi, P. (2013). Patterns of diet-related practices and prevalence of gastro-esophageal reflux disease. Neurogastroenterology & Motility, 25(10), 831-e638.

[27] Batis, C., Mendez, M. A., Sotres-Alvarez, D., Gordon-Larsen, P., & Popkin, B. (2014). Dietary pattern trajectories during 15 years of follow-up and HbA1c, insulin resistance and diabetes prevalence among Chinese adults. J Epidemiol Community Health, 68(8), 773-779.

[28] Gordon-Larsen, P., Koehler, E., Howard, A. G., Paynter, L., Thompson, A. L., Adair, L. S., ... & Herring, A. H. (2014). Eighteen year weight trajectories and metabolic markers of diabetes in modernising China. Diabetologia, 57(9), 1820-1829.

[29] Sotres-Alvarez, D., Herring, A. H., & Siega-Riz, A. M. (2013). Latent transition models to study women's changing of dietary patterns from pregnancy to 1 year postpartum. American journal of epidemiology, 177(8), 852-861.

[30] Pitt, E., Cameron, C. M., Thornton, L., Gallegos, D., Filus, A., Ng, S. K., & Comans, T. (2018). Dietary patterns of Australian children at three and five years of age and their changes over time: A latent class and latent transition analysis. Appetite, 129, 207-216.

[31] McCulloch, C. E., Lin, H., Slate, E. H., & Turnbull, B. W. (2002). Discovering subpopulation structure with latent class mixed models. Statistics in medicine, 21(3), 417-429.

[32] Rita Gaio, A., Costa, J. P. D., Santos, A. C., Ramos, E., & Lopes, C. (2012). A restricted mixture model for dietary pattern analysis in small samples. Statistics in medicine, 31(19), 2137-2150.

[33] Greve, B., Pigeot, I., Huybrechts, I., Pala, V., & Börnhorst, C. (2016). A comparison of heuristic and model-based clustering methods for dietary pattern analysis. Public health nutrition, 19(2), 255-264.

[34] Hsiao, P. Y., Mitchell, D. C., Coffman, D. L., Allman, R. M., Locher, J. L., Sawyer, P., ... & Hartman, T. J. (2013). Dietary patterns and diet quality among diverse older adults: the University of Alabama at Birmingham Study of Aging. The journal of nutrition, health & aging, 17(1), 19-25.

[35] Joy, E. J., Green, R., Agrawal, S., Aleksandrowicz, L., Bowen, L., Kinra, S., ... & Dangour, A. D. (2017). Dietary patterns and non-communicable disease risk in Indian adults: secondary analysis of Indian Migration Study data. Public health nutrition, 20(11), 1963-1972.

[36] Fahey, M. T., Ferrari, P., Slimani, N., Vermunt, J. K., White, I. R., Hoffmann, K., ... & Rodríguez-Barranco, M. (2011). Identifying dietary patterns using a normal mixture model: application to the EPIC study. Journal of Epidemiology & Community Health, jech-2009.

[37] Marchioni, D. M. L., Fisberg, R. M., Góis Filho, J. F. D., Kowalski, L. P., Carvalho, M. B. D., Abrahão, M., ... & Wünsch Filho, V. (2007). Dietary patterns and risk of oral cancer: a case-control study in São Paulo, Brazil. Revista de saude publica, 41(1), 19-26.

[38] Toledo, A. L. A. D., Koifman, R. J., Koifman, S., & Marchioni, D. M. L. (2010). Dietary patterns and risk of oral and pharyngeal cancer: a case-control study in Rio de Janeiro, Brazil. Cadernos de saude publica, 26, 135-142.

[39] De Stefani, E., Boffetta, P., Ronco, A. L., Correa, P., Oreggia, F., Deneo-Pellegrini, H., ... & Leiva, J. (2005). Dietary patterns and risk of cancer of the oral cavity and pharynx in Uruguay. Nutrition and cancer, 51(2), 132-139.

[40] De Stefani, E., Deneo-Pellegrini, H., Boffetta, P., Ronco, A. L., Aune, D., Acosta, G., ... & Ferro, G. (2009). Dietary patterns and risk of cancer: a factor analysis in Uruguay. International journal of cancer, 124(6), 1391-1397.

[41] Bradshaw, P. T., Siega-Riz, A. M., Campbell, M., Weissler, M. C., Funkhouser, W. K., & Olshan, A. F. (2012). Associations between dietary patterns and head and neck cancer: the Carolina head and neck cancer epidemiology study. American journal of epidemiology, 175(12), 1225-1233.

[42] Helen-Ng, L. C., Razak, I. A., Ghani, W. M. N., Marhazlinda, J., Norain, A. T., Raja Jallaludin, R. L., ... & Zain, R. B. (2012). Dietary pattern and oral cancer risk–a factor analysis study. Community dentistry and oral epidemiology, 40(6), 560-566.

[43] De Stefani, E., Boffetta, P., Correa, P., Deneo-Pellegrini, H., Ronco, A. L., Acosta, G., & Mendilaharsu, M. (2013). Dietary patterns and risk of cancers of the upper aerodigestive tract: a factor analysis in Uruguay. Nutrition and cancer, 65(3), 384-389.

[44] Amtha, R., Zain, R., Razak, I. A., Basuki, B., Roeslan, B. O., Gautama, W., & Purwanto, D. J. (2009). Dietary patterns and risk of oral cancer: a factor analysis study of a population in Jakarta, Indonesia. Oral oncology, 45(8), e49-e53.

[45] Bahmanyar, S., & Ye, W. (2006). Dietary patterns and risk of squamous-cell carcinoma and adenocarcinoma of the esophagus and adenocarcinoma of the gastric cardia: a population-based case-control study in Sweden. Nutrition and cancer, 54(2), 171-178.

[46] De Stefani, E., Boffetta, P., Fagundes, R. B., Deneo-Pellegrini, H., Ronco, A. L., Acosta, G., & Mendilaharsu, M. (2008). Nutrient patterns and risk of squamous cell carcinoma of the esophagus: a factor analysis in uruguay. Anticancer research, 28(4C), 2499-2506.

[47] De Stefani, E., Boffetta, P., Ronco, A. L., Deneo-Pellegrini, H., Correa, P., Acosta, G., & Mendilaharsu, M. (2008). Exploratory factor analysis of squamous cell carcinoma of the esophagus in Uruguay. Nutrition and cancer, 60(2), 188-195.

[48] Hajizadeh, B., Rashidkhani, B., Rad, A. H., Moasheri, S. M., & Saboori, H. (2010). Dietary patterns and risk of oesophageal squamous cell carcinoma: a case–control study. Public health nutrition, 13(7), 1107-1112.

[49] Hajizadeh, B., Jessri, M., Akhoondan, M., Moasheri, S. M., & Rashidkhani, B. (2012). Nutrient patterns and risk of esophageal squamous cell carcinoma: a case-control study. Diseases of the Esophagus, 25(5), 442-448.

[50] Ibiebele, T. I., Hughes, M. C., Whiteman, D. C., & Webb, P. M. (2012). Dietary patterns and risk of oesophageal cancers: a population-based case–control study. British Journal of Nutrition, 107(8), 1207-1216.

[51] Liu, X., Wang, X., Lin, S., Yuan, J., & Yu, I. T. (2014). Dietary patterns and oesophageal squamous cell carcinoma: a systematic review and meta-analysis. British journal of cancer, 110(11), 2785.

[52] Silvera, S. A. N., Mayne, S. T., Risch, H. A., Gammon, M. D., Vaughan, T., Chow, W. H., ... & West, A. B. (2011). Principal component analysis of dietary and lifestyle patterns in relation to risk of subtypes of esophageal and gastric cancer. Annals of epidemiology, 21(7), 543-550.

[53] Okada, E., Nakamura, K., Ukawa, S., Sakata, K., Date, C., Iso, H., & Tamakoshi, A. (2016). Dietary Patterns and Risk of Esophageal Cancer Mortality: The Japan Collaborative Cohort Study. Nutrition and cancer, 68(6), 1001-1009.

[54] Sewram, V., Sitas, F., O'Connell, D., & Myers, J. (2014). Diet and esophageal cancer risk in the Eastern Cape Province of South Africa. Nutrition and cancer, 66(5), 791-799.

[55] Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. Political Analysis, 12(1), 3-27.

[56] Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. Political analysis, 18(4), 450-469.

[57] Johnson RA, Wichern DW. (2007). Applied multivariate statistical analysis. 6th ed. Upper Saddle River, New Jersey: Prentice-Hall;.

[58] Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. Studies in Social Psychology in World War II Vol. IV: Measurement and Prediction, 362-412.

[59] Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part IA modified latent structure approach. American Journal of Sociology, 79(5), 1179-1259.

[60] Haberman, S.J. (1979). Analysis of Qualitative Data, Vol 2, New Developments. Academic Press. 9

[61]Hagenaars, J.A.(1990).Categorical Longitudinal Data - Loglinear Analysis of Panel, Trend and Cohort Data. Newbury Park: Sage.

[62] Vermunt, J.K. (1997). Log-linear Models for Event Histories. Thousand Oakes: Sage Publications.

[63] Hagenaars, J.A. and McCutcheon, A.L. (2002), Applied Latent Class Analysis. Cambridge University Press.

[64] Heinen, T. (1996). Latent class and discrete latent trait models: Similarities and differences. Sage Publications, Inc.

[65] Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots, and related graphical displays. Sociological methodology, 31(1), 223-264.

[66] van den Bergh, M. (2018). Latent class trees. PhD dissertation

[67] Vermunt, J. K., & Magidson, J. (2016). Technical guide for Latent GOLD 5.1: basic, advanced, and syntax. Statistical Innovations. Inc., Belmont Google Scholar.

[68] Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., ... & Bray, F. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. International journal of cancer, 136(5), E359-E386.

[69] AICR. (2008). World Cancer Research Fund/American Institute for Cancer Research. Food, nutrition, physical activity, and the prevention of cancer: A Global perspective.

[70] Boeing, H., Dietrich, T., Hoffmann, K., Pischon, T., Ferrari, P., Lahmann, P. H., ... & Skeie, G. (2006). Intake of fruits and vegetables and risk of cancer of the upper aero-digestive tract: the prospective EPIC-study. Cancer causes & control, 17(7), 957.

[71] Chuang, S. C., Jenab, M., Heck, J. E., Bosetti, C., Talamini, R., Matsuo, K., ... & La Vecchia, C. (2012). Diet and the risk of head and neck cancer: a pooled analysis in the INHANCE consortium. Cancer Causes & Control, 23(1), 69-88.

[72] Freedman, N. D., Park, Y., Subar, A. F., Hollenbeck, A. R., Leitzmann, M. F., Schatzkin, A., & Abnet, C. C. (2008). Fruit and vegetable intake and head and neck cancer risk in a large United States prospective cohort study. International Journal of Cancer, 122(10), 2330-2336.

[73] Lagiou, P., Talamini, R., Samoli, E., Lagiou, A., Ahrens, W., Pohlabeln, H., ... & Merletti, F. (2009). Diet and upper-aerodigestive tract cancer in Europe: the ARCAGE study. International journal of cancer, 124(11), 2671-2676.

[74] La Vecchia, C., Chatenoud, L., Franceschi, S., Soler, M., Parazzini, F., & Negri, E. (1999). Vegetables and fruit and human cancer: update of an Italian study. International journal of cancer, 82(1), 151-152.

[75] Franceschi, S., Bidoli, E., Vecchia, C. L., Talamini, R., D'Avanzo, B., & Negri, E. (1994). Tomatoes and risk of digestive-tract cancers. International Journal of Cancer, 59(2), 181-184.

[76] Franceschi, S., Favero, A., Conti, E., Talamini, R., Volpe, R., Negri, E., ... & La Vecchia, C. (1999). Food groups, oils and butter, and cancer of the oral cavity and pharynx. British Journal of cancer, 80(3-4), 614.

[77] Turati, F., Rossi, M., Pelucchi, C., Levi, F., & La Vecchia, C. (2015). Fruit and vegetables and cancer risk: a review of southern European studies. British Journal of Nutrition, 113(S2), S102-S110.

[78] La Vecchia C, Negri E, D'Avanzo B, Boyle P, Franceschi S. (1991). Dietary indicators of oral and pharyngeal cancer. International journal of epidemiology, 20(1), 39-44.

[79] Edefonti, V., Bravi, F., Garavello, W., La Vecchia, C., Parpinel, M., Franceschi, S., ... & Decarli, A. (2010). Nutrient-based dietary patterns and laryngeal cancer: evidence from an exploratory factor analysis. Cancer Epidemiology and Prevention Biomarkers, 19(1), 18-27.

[80] Tavani, A., La Vecchia, C., Gallus, S., Lagiou, P., Trichopoulos, D., Levi, F., & Negri, E. (2000). Red meat intake and cancer risk: a study in Italy. International Journal of Cancer, 86(3), 425-428.

[81] Di Maso, M., Talamini, R., Bosetti, C., Montella, M., Zucchetto, A., Libra, M., ... & Serraino, D. (2013). Red meat and cancer risk in a network of case–control studies focusing on cooking practices. Annals of oncology, 24(12), 3107-3112.

[82] Levi, F., Pasche, C., Lucchini, F., Bosetti, C., & La Vecchia, C. (2004). Processed meat and the risk of selected digestive tract and laryngeal neoplasms in Switzerland. Annals of oncology, 15(2), 346-349.

[83] Garavello, W., Lucenteforte, E., Bosetti, C., & La, C. V. (2009). The role of foods and nutrients on oral and pharyngeal cancer risk. Minerva stomatologica, 58(1-2), 25-34.

[84] Bravi, F., Polesel, J., Garavello, W., Serraino, D., Negri, E., Franchin, G., ... & Bosetti, C. (2017). Adherence to the World Cancer Research Fund/American Institute for Cancer Research recommendations and head and neck cancers risk. Oral oncology, 64, 59-64.

[85] Bravi, F., Edefonti, V., Randi, G., Ferraroni, M., La Vecchia, C., & Decarli, A. (2012). Dietary patterns and upper aerodigestive tract cancers: an overview and review. Annals of oncology, 23(12), 3024-3039.

[86] Edefonti, V., Hashibe, M., Ambrogi, F., Parpinel, M., Bravi, F., Talamini, R., ... & McClean, M. (2011). Nutrient-based dietary patterns and the risk of head and neck cancer: a pooled analysis in the International Head and Neck Cancer Epidemiology consortium. Annals of Oncology, 23(7), 1869-1880.

[87] Decarli, A., Franceschi, S., Ferraroni, M., Gnagnarella, P., Parpinel, M. T., La Vecchia, C., ... & Giacosa, A. (1996). Validation of a food-frequency questionnaire to assess dietary intakes in cancer studies in Italy results for specific nutrients. Annals of epidemiology, 6(2), 110-118.

[88] Franceschi, S., Negri, E., Salvini, S., Decarli, A., Ferraroni, M., Filiberti, R., ... & La Vecchia, C. (1993). Reproducibility of an Italian food frequency questionnaire for cancer studies: results for specific food items. European journal of cancer, 29(16), 2298-2305.

[89] Franceschi, S., Negri, E., Salvini, S., Decarli, A., Ferraroni, M., Filiberti, R., ... & La Vecchia, C. (1993). Reproducibility of an Italian food frequency questionnaire for cancer studies: results for specific food items. European journal of cancer, 29(16), 2298-2305.

[90] Gnagnarella, P., Parpinel, M., Salvini, S., Franceschi, S., Palli, D., & Boyle, P. (2004). The update of the Italian food composition database. Journal of Food Composition and Analysis, 17(3-4), 509-522.

[91] Northstone, K., Ness, A. R., Emmett, P. M., & Rogers, I. S. (2008). Adjusting for energy intake in dietary pattern investigations using principal components analysis. European journal of clinical nutrition, 62(7), 931.

[92] Balder, H. F., Virtanen, M., Brants, H. A., Krogh, V., Dixon, L. B., Tan, F., ... & Berrino, F. (2003). Common and country-specific dietary patterns in four European cohort studies. The Journal of nutrition, 133(12), 4246-4251.

[93] Bagnardi, V., Blangiardo, M., La Vecchia, C., & Corrao, G. (2001). A meta-analysis of alcohol drinking and cancer risk. British journal of cancer, 85(11), 1700.

[94] Boyle, P., & Levin, B. (2008). Head and neck cancers. In Boyle, P., & Levin, B. World cancer report 2008. IARC Press, International Agency for Research on Cancer.

[95] Boyle P, Levin B. (2008). Esophageal cancer. In Boyle, P., & Levin, B. World cancer report 2008. IARC Press, International Agency for Research on Cancer.

[96] La Vecchia C, Tavani A, Franceschi S et al. Epidemiology and prevention of oral cancer. Oral Oncol 1997; 33: 302–312.

[97] La Vecchia, C., Tavani, A., Franceschi, S., Levi, F., Corrao, G., & Negri, E. (1997). Epidemiology and prevention of oral cancer. Oral oncology, 33(5), 302-312.

[98] McClelland, R. L., & Kronmal, R. A. (2002). Regression-based variable clustering for data reduction. Statistics in medicine, 21(6), 921-941.

[99] van den Bergh, M., Schmittmann, V. D., & Vermunt, J. K. (2017). Building latent class trees, with an application to a study of social capital. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 13(S1), 13.

[100] van den Bergh, M., van Kollenburg, G. H., & Vermunt, J. K. (2018). Deciding on the starting number of classes of a latent class tree. Sociological Methodology, 0081175018780170.

[101] Cirmi, S., Navarra, M., Woodside, J. V., & Cantwell, M. M. (2018). Citrus fruits intake and oral cancer risk: a systematic review and meta-analysis. Pharmacological research.

[102] Wang, A., Zhu, C., Fu, L., Wan, X., Yang, X., Zhang, H., ... & Zhao, H. (2015). Citrus fruit intake substantially reduces the risk of esophageal cancer: a meta-analysis of epidemiologic studies. Medicine, 94(39).

[103] Zhao, W., Liu, L., & Xu, S. (2018). Intakes of citrus fruit and risk of esophageal cancer: A meta-analysis. Medicine, 97(13).

[104] Van der Palm, D. W., Van der Ark, L. A., & Vermunt, J. K. (2016). Divisive latent class modeling as a density estimation method for categorical data. Journal of Classification, 33(1), 52-72.

[105] Andrews, R. L., & Currim, I. S. (2003). A comparison of segment retention criteria for finite mixture logit models. Journal of Marketing Research, 40(2), 235-243.

[106] Nylund, K. L., Asparouhov, T. I. H. O. M. I. R., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Structural Equation Modeling, 14, 535-69.

[107] Mertz, W. (1984). Foods and nutrients. *Journal of the American Dietetic Association (USA)*.

[108] Everitt, B., & Hothorn, T. (2011). An introduction to applied multivariate analysis with R. Springer Science & Business Media.

[109] Kent, J. T., Bibby, J. M., & Mardia, K. V. (2006). Multivariate analysis (probability and mathematical statistics).

[110] Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. Sociological Methodology, 43(1), 272-311.

[111] Magidson, J., & Vermunt, J. (2002). Latent class models for clustering: A comparison with K-means. Canadian Journal of Marketing Research, 20(1), 36-43.

[112] Garre, F. G., & Vermunt, J. K. (2006). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. Behaviormetrika, 33(1), 43-59.

[113] Galindo-Garre, F., Vermunt, J. K., & Bergsma, W. P. (2004). Bayesian posterior estimation of logit parameters with small samples. Sociological Methods & Research, 33(1), 88-117.

SUPPLEMENTARY MATERIALS

Supplementary Table 4.1 Fit statistic and BIC improvement of the multiple LC models on oral/pharingeal cancer, Italy,1992-2009.

Nr of classes	logL	Nr P	BIC
1	-63574.1	64	127669.3
2	-63021.0	97	126831.8
3	-62754.7	130	126568.0
4	-62545.6	163	126418.4
5	-62414.5	196	126424.9
6	-62301.6	229	126467.9
7	-62158.7	262	126450.8
8	-62062.9	295	126527.9
9	-62012.8	328	126696.3
10	-61884.7	361	126708.8

Supplementary Table4.2 Probabilities of consumption for all food groups by dietary patterns derived from LCA. Italy, 1992-2009.

		Prudent %	Western %	Lower consumers-combination %	Higher consumers-combination %
Size		36,8	27,0	21,1	15,1
milk	Not consumed	12,0	22,7	28,8	39,0
	Below median	49,2	45,1	45,8	46,2
	Above median	38,8	32,3	25,5	14,8
coffee	Below median	52,2	52,8	58,5	**60,6**
	Above median	47,8	47,2	41,5	39,4
tea	Not consumed	39,9	44,6	**62,3**	**62,8**
	Consumed	**60,1**	55,4	37,7	37,3
bread	Below median	57,8	44,1	**63,8**	22,9
	Above median	42,2	55,9	36,2	**77,1**
pasta	Below median	48,8	46,9	52,0	49,5
	Above median	51,2	53,1	48,1	50,5
soup	Below median	47,3	46,6	58,6	43,3
	Above median	52,7	53,4	41,4	56,8
eggs	Not consumed	11,3	5,8	25,0	16,8
	Below median	43,0	41,6	43,0	33,8
	Above median	45,8	52,7	32,0	49,5
white	Below median	45,9	52,2	**61,3**	38,1
meat	Above median	54,1	47,8	38,7	**61,9**
red	Below median	**61,1**	30,9	**65,0**	35,7
meat	Above median	38,9	**69,1**	35,0	**64,3**
offals	Not consumed	**75,3**	49,0	**81,3**	**71,9**
	Consumed	24,7	51,0	18,7	28,1
processed	Below median	50,7	47,3	50,8	**63,2**
meat	Above median	49,3	52,7	49,2	36,8

		Prudent %	Western %	Lower consumers-combination %	Higher consumers-combination %
fish	Below median	45,6	49,2	**69,1**	**66,8**
	Above median	54,4	50,8	30,9	33,3
cheese	Below median	41,1	48,4	58,5	**62,6**
	Above median	58,9	51,6	41,5	37,4
potatoes	Below median	58,2	45,7	**61,7**	41,1
	Above median	41,8	54,3	38,3	59,0
pulses	Below median	42,7	57,3	**62,6**	44,0
	Above median	57,3	42,7	37,4	56,0
leafy	Below median	35,5	59,8	**74,6**	22,5
vegetables	Above median	**64,5**	40,2	25,4	**77,5**
fruiting	Below median	24,4	**79,2**	**71,1**	31,1
vegetables	Above median	**75,6**	20,8	28,9	**69,0**
root	Not consumed	11,2	17,6	57,5	30,1
vegetables	Below median	34,7	48,6	32,7	29,7
	Above median	54,2	33,7	9,8	40,2
cruciferous	Not consumed	18,3	14,8	51,5	24,9
vegetables	Below median	25,1	**67,3**	19,8	46,6
	Above median	56,6	17,9	28,7	28,5
other	Not consumed	6,3	1,5	41,3	6,4
vegetables	Below median	36,1	55,9	54,2	22,7
	Above median	57,6	42,5	4,4	**70,9**
citrus	Not consumed	4,1	7,2	25,3	17,4
fruit	Below median	24,2	59,0	39,9	**62,8**
	Above median	**71,7**	33,8	34,8	19,9
other	Below median	29,8	**63,6**	**67,7**	50,2
fruits	Above median	**70,2**	36,4	32,3	49,8
sugary	Not consumed	54,2	41,3	59,9	**70,9**
drinks	Consumed	45,8	58,8	40,2	29,1
desserts	Below median	44,3	51,8	59,2	**68,1**
	Above median	55,7	48,2	40,8	32,0
sugar	Below median	49,7	44,9	55,3	52,5
	Above median	50,3	55,1	44,7	47,5

Supplementary Table4.3 Log-likelihood and classification statistics for LCA solution. Italy, 1992-2009.

Log-likelihood statistics		Classification statistics			
BIC	AIC	Classification errors	Reduction of errors	Entropy R-square	Standard R-square
126418,4	125417,2	0.23	0.64	0.57	0.56

Supplementary Table 5.1. Results of Wald test on the effect of NAE on latent and/or indicators variables in the different LC models. Italy, 1992-2009.

		Model 1			Model 2			Model 3		
		par	Wald	p-value	par	Wald	p-value	par	Wald	p-value
Milk	Not consumed				-0.0007	100.78	<0.001	-0.0008	66.13	<0.001
	Below median				-0.0005			-0.0006		
	Above median				0[a]			0[a]		
Coffee	Below median				-0.0001	4.99	0.026	-0.0002	6.58	0.01
	Above median				0[a]			0[a]		
Tea - deca	Not consumed				-0.0001	5.82	0.016	-0.0001	2.04	0.15
	Consumed				0[a]			0[a]		
Bread	Below median				-0.002	564.03	<0.001	-0.0019	438.80	<0.001
	Above median				0[a]			0[a]		
Pasta - rice	Below median				-0.0011	310.43	<0.001	-0.0012	276.73	<0.001
	Above median				0[a]			0[a]		
Soup	Below median				-0.0002	14.33	<0.001	-0.0002	7.72	0.01
	Above median				0[a]			0[a]		
Eggs	Not consumed				-0.0008	114.21	<0.001	-0.0007	66.97	<0.001
	Below median				-0.0004			-0.0004		
	Above median				0[a]			0[a]		
White meat	Below median				-0.0005	76.93	<0.001	-0.0004	37.79	<0.001
	Above median				0[a]			0[a]		
Red meat	Below median				-0.0014	397.53	<0.001	-0.0014	288.07	<0.001
	Above median				0[a]			0[a]		
Offal	Not consumed				-0.0005	72.70	<0.001	-0.0004	37.51	<0.001
	Consumed				0[a]			0[a]		
Processed meat	Below median				-0.0006	121.18	<0.001	-0.0007	111.70	<0.001
	Above median				0[a]			0[a]		
Fish	Below median				-0.0003	30.06	<0.001	-0.0002	14.95	<0.001
	Above median				0[a]			0[a]		
Cheese	Below median				-0.0006	139.75	<0.001	-0.0007	92.17	<0.001
	Above median				0[a]			0[a]		
Potatoes	Below median				-0.0008	187.39	<0.001	-0.0007	139.71	<0.001
	Above median				0[a]			0[a]		
Pulses	Below median				-0.0004	45.00	<0.001	-0.0003	27.64	<0.001
	Above median				0[a]			0[a]		

		Model 1			Model 2			Model 3		
		par	Wald	p-value	par	Wald	p-value	par	Wald	p-value
Leafy veg.	Below median				-0.0007	136.31	<0.001	-0.0004	33.24	<0.001
	Above median				0[a]			0[a]		
Fruiting veg.	Below median				-0.0004	42.15	<0.001	-0.0002	4.05	0.04
	Above median				0[a]			0[a]		
Root veg.	Not consumed				-0.0006	61.56	<0.001	-0.0002	9.41	0.01
	Below median				-0.0003			-0.0002		
	Above median				0[a]			0[a]		
Cruciferous veg.	Not consumed				-0.0003	25.82	<0.001	-0.0001	2.83	0.24
	Below median				0.0000			-0.0001		
	Above median				0[a]			0[a]		
Other veg.	Not consumed				-0.0011			-0.0002		
	Below median				-0.0006	108.81	<0.001	-0.0001	3.86	0.14
	Above median				0[a]			0[a]		
Citrus fruits	Not consumed				-0.0007	52.51	<0.001	-0.0006	27.95	<0.001
	Below median				-0.0003			-0.0004		
	Above median				0[a]			0[a]		
Fruits (not citrus)	Below median				-0.0007	137.30	<0.001	-0.0006	81.49	<0.001
	Above median				0[a]			0[a]		
Sugary drinks	Not consumed				-0.0006	119.60	<0.001	-0.0007	84.31	<0.001
	Consumed				0[a]			0[a]		
Desserts	Below median				-0.0011	268.54	<0.001	-0.0013	161.82	<0.001
	Above median				0[a]			0[a]		
Sugar	Below median				-0.001	250.44	<0.001	-0.0011	224.94	<0.001
	Above median				0[a]			0[a]		
Latent classes	Cluster 1	0.0189	351.94	<0.001				-0.0006	61.43	<0.001
	Cluster 2	0.0146						-0.0005		
	Cluster 3	0.0038						-0.0018		
	Cluster 4	0.0170						0[a]		
	Cluster 5	0.0122								
	Cluster 6	0.0131								
	Cluster 7	0[a]								

[a]Reference for dummy conding.

Supplementary Table.6.1 Fit statistic and BIC improvement of the multiple LC models on esophageal cancer, Italy,1992-1997.

Nr of classes	logL	Nr P	BIC
1	-19013.8	64	38472.6
2	-18797.6	97	38269.6
3	-18695.3	130	38294.6
4	-18627.0	163	38387.5
5	-18562.3	196	38487.5
6	-18515.4	229	38623.1
7	-18469.9	262	38761.7
8	-18422.4	295	38896.1
9	-18382.4	328	39045.7
10	-18351.9	361	39214.2

Supplementary Table6.2 Probabilities of consumption for all food groups by dietary patterns derived from LCA. on the ESCC study. Italy,1992-1997.

		Class 1 %	Class 2 %
Size		55.6	44.4
milk	Not consumed	21.3	31.5
	Below median	44.9	45.4
	Above median	33.8	23.1
coffee	Below median	53.9	53.6
	Above median	46.1	46.4
tea	Not consumed	42.7	48.7
	Consumed	57.3	51.3
bread	Below median	57.3	41.1
	Above median	42.7	58.9
pasta	Below median	51.6	48.1
	Above median	48.4	51.9
soup	Below median	49.7	56.3
	Above median	50.3	43.7
eggs	Not consumed	11.6	12.7
	Below median	37.4	50.2
	Above median	51.1	37.2
white	Below median	38.1	59.5
meat	Above median	**61.9**	40.5
red	Below median	52.7	46.6
meat	Above median	47.3	53.4
offals	Not consumed	60.8	60.7
	Consumed	39.2	39.3
processed	Below median	55.2	55.5
meat	Above median	44.8	44.5
fish	Below median	49.0	**72.9**
	Above median	51.0	27.1
cheese	Below median	44.3	57.3
	Above median	55.7	42.7
potatoes	Below median	42.3	57.3
	Above median	57.7	42.7

		Class 1 %	Calss2 %
pulses	Below median	47.5	59.5
	Above median	52.5	40.5
leafy	Below median	42.1	**62.5**
vegetables	Above median	57.9	37.5
fruiting	Below median	35.2	**69.6**
vegetables	Above median	**64.8**	30.4
root	Not consumed	9.0	28.8
vegetables	Below median	27.8	58.0
	Above median	**63.2**	13.3
cruciferous	Not consumed	15.5	31.9
vegetables	Below median	48.3	57.4
	Above median	36.3	10.8
other	Not consumed	2.0	17.5
vegetables	Below median	31.1	**64.1**
	Above median	**66.9**	18.4
citrus	Not consumed	8.7	14.3
fruit	Below median	53.7	**63.3**
	Above median	37.6	22.4
other	Below median	40.0	**62.6**
fruits	Above median	**60.0**	37.4
sugary	Not consumed	56.6	53.0
drinks	Consumed	43.4	47.1
desserts	Below median	46.8	54.2
	Above median	53.2	45.8
sugar	Below median	52.7	46.6
	Above median	47.3	53.4

Supplementary Table. 6.3 Fit statistics and their relative improvement of the multiple LC models on OPC and ESCC studies, Italy,1992-2009.

Nr of classes	logL	Nr P	BIC	AIC	AIC3	R_{BIC}	R_{AIC}	R_{AIC3}
1	-76132.9	64	152799.7	152393.9	152457.9			
2	-75509.6	97	151828.2	151213.1	151310.1	1	1	1
3	-75217.4	130	151519.2	150694.8	150824.8	0.318114	0.438956	0.422825
4	-75003.4	163	151366.4	150332.8	150495.8	0.157281	0.306625	0.28669
5	-74860.2	196	151355.3	150112.5	150308.5	0.011395	0.186593	0.163207
6	-74745.8	229	151401.6	149952.8	150181.8	-0.04766	0.135226	0.110363
7	-74643.1	262	151471.5	149674.3	149936.3	-0.07193	0.235861	0.213891
8	-74499.3	295	151459.2	149575.5	149870.5	0.012602	0.083654	0.057308
9	-74422.9	328	151581.6	149450	149778	-0.12598	0.106331	0.080637
10	-74332.6	361	151676.2	149351.2	149712.2	-0.0974	0.083636	0.057289

Supplementary Table6.4 Probabilities of consumption for all food groups by dietary patterns derived from LCA. 5-classes solution, OPC and ESCC studies, Italy,1992-2009.

		Class1 %	Class2 %	Class3 %	Class4 %	Class5 %
Size		31.8	29.3	18.5	16.7	3.8
milk	Not consumed	11.2	23.0	37.1	29.9	29.7
	Below median	47.1	46.3	48.7	46.9	31.0
	Above median	41.7	30.7	14.2	23.3	39.3
coffee	Below median	54.5	53.6	57.6	55.8	47.8
	Above median	45.5	46.4	42.4	44.2	52.2
tea	Not consumed	38.0	41.5	**62.5**	**65.7**	46.9
	Consumed	**62.0**	58.6	37.5	34.3	53.1
bread	Below median	**65.4**	46.9	28.3	59.0	16.5
	Above median	34.6	53.1	**71.7**	41.0	**83.5**
pasta	Below median	54.0	49.0	44.4	47.8	32.7
	Above median	46.0	51.0	55.6	52.2	**67.3**
soup	Below median	47.4	48.3	41.9	57.4	49.5
	Above median	52.6	51.7	58.1	42.6	50.5
eggs	Not consumed	12.3	6.8	13.3	24.2	20.3
	Below median	41.9	48.7	36.3	40.6	29.1
	Above median	45.8	44.5	50.4	35.2	50.6
white	Below median	46.6	56.3	38.4	**64.6**	42.8
meat	Above median	53.5	43.7	**61.6**	35.4	57.2
red	Below median	**61.4**	41.4	31.2	55.7	31.4
meat	Above median	38.6	58.6	**68.9**	44.3	**68.6**
offals	Not consumed	**76.8**	50.1	**66.1**	**80.8**	**64.1**
	Consumed	23.2	49.9	33.9	19.2	35.9
processed	Below median	54.0	54.8	54.6	47.7	44.7
meat	Above median	46.0	45.2	45.4	52.3	55.3
fish	Below median	46.1	56.9	55.4	**70.3**	58.0
	Above median	53.9	43.1	44.6	29.7	42.0
cheese	Below median	45.5	57.3	**60.5**	58.7	40.3
	Above median	54.5	42.7	39.5	41.3	59.7
potatoes	Below median	**60.3**	56.7	38.9	58.0	48.6
	Above median	39.7	43.3	**61.1**	42.0	51.4
pulses	Below median	44.8	59.3	43.3	**60.7**	56.9
	Above median	55.2	40.7	56.7	39.3	43.1
leafy	Below median	38.8	58.3	25.6	**75.8**	37.7
vegetables	Above median	**61.2**	41.8	**74.4**	24.2	**62.3**
fruiting	Below median	36.2	**86.2**	39.6	**76.2**	56.4
vegetables	Above median	**63.8**	13.8	**60.4**	23.8	43.6

		Class1 %	Class2 %	Class3 %	Class4 %	Class5 %
root	Not consumed	10.6	15.2	25.0	**63.4**	25.1
vegetables	Below median	36.1	**62.3**	32.3	30.0	45.0
	Above median	53.3	22.5	42.7	6.6	29.9
cruciferous	Not consumed	19.9	17.7	19.5	51.4	34.5
vegetables	Below median	53.0	**78.4**	**60.1**	37.6	40.9
	Above median	27.1	4.0	20.5	11.0	24.6
other	Not consumed	7.1	1.6	5.5	45.4	14.0
vegetables	Below median	38.9	**76.8**	31.0	49.8	51.6
	Above median	54.0	21.7	**63.5**	4.8	34.4
citrus	Not consumed	5.4	7.4	13.8	24.7	18.2
fruit	Below median	53.8	**79.2**	**75.4**	**61.5**	49.0
	Above median	40.9	13.5	10.8	13.8	32.8
other	Below median	42.2	**72.7**	**63.7**	**78.5**	42.2
fruits	Above median	57.8	27.3	36.3	21.5	57.8
sugary	Not consumed	53.3	45.0	**70.3**	**60.2**	39.1
drinks	Consumed	46.7	55.0	29.7	39.9	**60.9**
desserts	Below median	37.4	58.0	**72.4**	52.4	32.4
	Above median	**62.6**	42.1	27.6	47.6	**67.6**
sugar	Below median	48.4	44.0	53.2	49.8	34.5
	Above median	51.6	56.0	46.9	50.3	**65.5**

Supplementary Table6.5 Probabilities of consumption for all food groups by dietary patterns derived from LCT. First level split, nodes 1, 2, 3, 4. OPC and ESCC studies, Italy 1992-2009.

		Class 1 Prudent %	Class 2 Western %	Class 3 Lower consumers-combination %	Class 4 Higher consumers-combination %
Size		31.5	29.9	19.6	19.1
milk	Not consumed	10.7	22.0	31.3	37.7
	Below median	46.0	46.2	45.7	48.8
	Above median	43.3	31.8	23.1	13.5
coffee	Below median	53.6	52.0	57.3	58.5
	Above median	46.4	48.0	42.7	41.5
tea	Not consumed	38.0	40.4	**63.5**	**63.4**
	Consumed	**62.0**	59.6	36.5	36.6
bread	Below median	**62.6**	44.4	**60.1**	28.6
	Above median	37.4	55.6	39.9	**71.4**
pasta	Below median	52.2	46.2	51.6	45.1
	Above median	47.8	53.8	48.4	54.9
soup	Below median	46.7	47.0	58.1	43.5
	Above median	53.3	53.1	42.0	56.5
eggs	Not consumed	11.9	6.0	24.1	15.3
	Below median	41.2	47.9	41.5	35.6
	Above median	46.9	46.2	34.4	49.1
white	Below median	45.8	54.0	**65.8**	38.4
meat	Above median	54.2	46.0	34.2	**61.6**
red	Below median	59.7	37.7	58.1	34.1
meat	Above median	40.4	**62.4**	41.9	**65.9**
offals	Not consumed	**76.3**	48.9	**79.6**	**68.3**
	Consumed	23.7	51.1	20.4	31.7
processed	Below median	52.9	53.3	51.5	54.1
meat	Above median	47.1	46.7	48.5	45.9
fish	Below median	45.4	55.0	**72.1**	55.8
	Above median	54.6	45.0	27.9	44.2
cheese	Below median	43.7	55.5	59.3	**61.9**
	Above median	56.3	44.5	40.7	38.1
potatoes	Below median	58.0	54.7	59.9	42.8
	Above median	42.0	45.4	40.1	57.2
pulses	Below median	43.8	57.6	**63.2**	44.6
	Above median	56.2	42.4	36.8	55.4
leafy	Below median	37.7	55.3	**76.7**	25.3
vegetables	Above median	**62.4**	44.7	23.3	**74.7**
fruiting	Below median	34.8	**84.0**	**79.1**	38.7
vegetables	Above median	**65.2**	16.0	20.9	**61.3**

		Class 1 Prudent %	Class 2 Western %	Class 3 Lower consumers-combination %	Class 4 Higher consumers-combination %
root vegetables	Not consumed	10.7	13.0	59.3	26.9
	Below median	35.3	**61.2**	33.8	33.6
	Above median	54.1	25.9	7.0	39.5
cruciferous vegetables	Not consumed	19.4	16.3	51.1	20.9
	Below median	52.6	**79.5**	39.2	56.3
	Above median	28.0	4.2	9.7	22.8
other vegetables	Not consumed	6.9	1.0	41.1	7.1
	Below median	38.3	**73.1**	54.0	33.5
	Above median	54.8	25.9	4.9	59.5
citrus fruit	Not consumed	5.1	5.8	25.4	15.2
	Below median	52.8	**79.7**	**61.8**	**72.3**
	Above median	42.1	14.5	12.9	12.5
other fruits	Below median	40.4	**70.8**	**78.0**	**63.2**
	Above median	59.6	29.2	22.0	36.9
sugary drinks	Not consumed	52.1	43.5	59.7	**71.1**
	Consumed	47.9	56.5	40.4	28.9
desserts	Below median	34.8	55.9	53.8	**73.2**
	Above median	65.2	44.1	46.2	26.8
sugar	Below median	46.4	42.0	50.9	55.5
	Above median	53.6	58.0	49.1	44.5

Supplementary Table6.6 Probabilities of consumption for all food groups by dietary patterns derived from LCT. Second level splits, nodes 1.1,1.2 and 2.1, 2.2 . OPC and ESCC studies, Italy, 1992-2009.

		Parental class: 1		Parental class: 2	
		Class 1.1 %	Class 1.2 %	Class 2.1 %	Class 2.2 %
Size		56.5	43.5	56.1	43.9
milk	Not consumed	11.4	9.8	21.7	22.4
	Below median	48.9	42.3	44.3	48.8
	Above median	39.8	47.9	34.1	28.9
coffee	Below median	**64.5**	39.3	45.5	**60.3**
	Above median	35.5	**60.7**	54.5	39.7
tea	Not consumed	33.9	43.3	47.5	31.2
	Consumed	**66.1**	56.7	52.5	**68.8**
bread	Below median	**68.2**	55.9	43.3	45.4
	Above median	31.8	44.1	56.7	54.6
pasta	Below median	**62.4**	39.2	37.9	56.6
	Above median	37.6	**60.8**	**62.1**	43.4
soup	Below median	50.8	41.4	49.3	43.9
	Above median	49.2	58.6	50.7	56.1
eggs	Not consumed	15.0	7.8	3.9	8.7
	Below median	49.7	30.2	41.6	56.0
	Above median	35.3	**62.0**	54.6	35.4
white	Below median	49.2	41.4	53.1	55.3
meat	Above median	50.8	58.6	46.9	44.7
red	Below median	**75.3**	39.8	23.7	55.3
meat	Above median	24.7	**60.2**	**76.3**	44.7
offals	Not consumed	83.3	67.3	41.8	57.9
	Consumed	16.7	32.7	58.2	42.1
processed	Below median	**70.8**	29.6	37.6	**73.4**
meat	Above median	29.2	**70.4**	**62.4**	26.6
fish	Below median	55.4	32.5	45.7	**66.8**
	Above median	44.7	**67.5**	54.3	33.2
cheese	Below median	49.4	36.3	47.1	**66.2**
	Above median	50.6	**63.7**	52.9	33.8
potatoes	Below median	**68.6**	44.5	38.2	**75.8**
	Above median	31.4	55.6	**61.9**	24.2
pulses	Below median	51.1	34.4	52.5	**64.1**
	Above median	48.9	**65.6**	47.5	35.9
leafy	Below median	38.1	37.0	56.0	54.5
vegetables	Above median	**61.9**	**63.0**	44.0	45.5
fruiting	Below median	35.7	33.7	**82.3**	**86.3**
vegetables	Above median	**64.4**	**66.3**	17.8	13.7

		Parental class: 1		Parental class: 2	
		Class 1.1 %	Class 1.2 %	Class 2.1 %	Class 2.2 %
root	Not consumed	9.6	12.1	15.1	10.3
vegetables	Below median	39.2	30.2	52.6	**72.1**
	Above median	51.3	57.7	32.3	17.6
cruciferous	Not consumed	23.4	14.2	13.7	19.6
vegetables	Below median	53.7	51.1	**81.3**	**77.1**
	Above median	22.9	34.7	5.0	3.3
other	Not consumed	7.2	6.6	0.7	1.4
vegetables	Below median	38.6	38.0	**66.6**	**81.3**
	Above median	54.2	55.5	32.7	17.3
citrus	Not consumed	6.8	3.0	4.9	7.0
fruit	Below median	56.7	47.8	**77.2**	**82.8**
	Above median	36.6	49.2	17.9	10.2
other	Below median	38.2	43.3	**72.3**	**68.9**
fruits	Above median	**61.8**	56.7	27.7	31.1
sugary	Not consumed	56.8	46.0	40.7	47.0
drinks	Consumed	43.2	54.0	59.4	53.0
desserts	Below median	37.9	30.5	47.1	**67.2**
	Above median	**62.1**	**69.5**	52.9	32.8
sugar	Below median	52.0	39.1	39.3	45.2
	Above median	48.0	**60.9**	**60.7**	54.8

Supplementary Table6.7 Probabilities of consumption for all food groups by dietary patterns derived from LCT. Second level splits, nodes 3.1,3.2 and 4.1, 4.2 . OPC and ESCC studies, Italy 1992-2009.

		Parental class: 3		Parental class: 4	
		Class 3.1 %	Class 3.2 %	Class 4.1 %	Class 4.2 %
Size		59.6	40.4	67.7	32.3
milk	Not consumed	36.7	23.3	34.8	43.8
	Below median	42.4	50.1	47.7	51.2
	Above median	20.9	26.6	17.5	5.0
coffee	Below median	41.9	**80.2**	54.9	**65.9**
	Above median	58.1	19.8	45.1	34.1
tea	Not consumed	**79.2**	40.4	**64.4**	**61.4**
	Consumed	20.8	59.7	35.6	38.7
bread	Below median	55.5	**67.4**	24.7	36.7
	Above median	44.5	32.6	**75.3**	**63.3**
pasta	Below median	45.3	**61.6**	42.8	49.4
	Above median	54.7	38.5	57.2	50.6
soup	Below median	**63.6**	49.9	36.3	58.6
	Above median	36.4	50.1	**63.7**	41.4
eggs	Not consumed	24.2	24.3	8.8	28.7
	Below median	37.8	46.5	32.3	42.5
	Above median	38.1	29.2	58.9	28.8
white	Below median	**68.2**	**62.1**	38.1	39.0
meat	Above median	31.8	37.9	**61.9**	**61.0**
red	Below median	49.9	**70.6**	23.4	56.3
meat	Above median	50.1	29.4	**76.6**	43.8
offals	Not consumed	**76.1**	**84.5**	**63.3**	**78.9**
	Consumed	23.9	15.5	36.7	21.1
processed	Below median	48.7	55.9	52.9	56.4
meat	Above median	51.3	44.1	47.1	43.6
fish	Below median	**68.7**	**77.2**	55.9	55.4
	Above median	31.3	22.8	44.1	44.6
cheese	Below median	58.1	**61.0**	57.4	**71.0**
	Above median	41.9	39.0	42.6	29.0
potatoes	Below median	56.7	**64.6**	30.1	**69.2**
	Above median	43.3	35.4	**69.9**	30.9
pulses	Below median	**61.3**	**66.0**	42.7	48.6
	Above median	38.8	34.0	57.4	51.4
leafy	Below median	**77.0**	**76.1**	24.9	26.1
vegetables	Above median	23.0	24.0	**75.1**	**73.9**
fruiting	Below median	**77.5**	**81.5**	38.0	40.1
vegetables	Above median	22.5	18.5	**62.0**	59.9

		Parental class: 3		Parental class: 4	
		Class 3.1 %	Class 3.2 %	Class 4.1 %	Class 4.2 %
root vegetables	Not consumed	**65.7**	49.8	27.0	26.7
	Below median	28.4	41.6	27.2	46.9
	Above median	5.9	8.6	45.8	26.4
cruciferous vegetables	Not consumed	48.2	55.3	20.2	22.4
	Below median	40.2	37.7	**61.5**	45.6
	Above median	11.6	7.1	18.4	32.1
other vegetables	Not consumed	37.9	45.7	6.2	8.9
	Below median	56.7	50.0	24.8	51.7
	Above median	5.4	4.3	**69.0**	39.4
citrus fruit	Not consumed	25.9	24.7	14.4	16.9
	Below median	**61.6**	**61.9**	**74.6**	**67.5**
	Above median	12.6	13.5	10.9	15.6
other fruits	Below median	**81.5**	**72.5**	**63.4**	**62.6**
	Above median	18.6	27.5	36.6	37.4
sugary drinks	Not consumed	**63.9**	53.4	**66.3**	**81.1**
	Consumed	36.1	46.6	33.7	18.9
desserts	Below median	58.3	48.0	**66.5**	**87.0**
	Above median	41.7	52.0	33.5	13.0
sugar	Below median	54.4	46.5	47.4	**72.2**
	Above median	45.6	53.5	52.6	27.8

Supplementary Table6.8 Probabilities of consumption for all food groups by dietary patterns derived from LCT. Third level splits, nodes 1.1.1, 1.1.2 and 1.2.1 ,1.2.2 . OPC and ESCC studies, Italy, 1992-2009.

		Parental class: 1.1		Parental class: 1.2	
		Class 1.1.1 %	Class 1.1.2 %	Class 1.1.1 %	Class 1.1.2 %
Size		85.1	14.9	87.3	12.7
milk	Not consumed	11.0	13.6	10.1	7.8
	Below median	48.9	48.6	41.9	45.0
	Above median	40.1	37.8	48.0	47.2
coffee	Below median	**65.1**	**61.7**	38.2	47.1
	Above median	35.0	38.3	**61.9**	52.9
tea	Not consumed	33.9	34.0	44.6	34.2
	Consumed	**66.1**	**66.0**	55.4	**65.8**
bread	Below median	**67.9**	**70.3**	54.5	**66.4**
	Above median	32.1	29.7	45.5	33.7
pasta	Below median	**62.3**	**62.9**	39.7	35.1
	Above median	37.7	37.1	**60.3**	**64.9**
soup	Below median	49.5	57.9	40.4	48.5
	Above median	50.5	42.1	59.6	51.5
eggs	Not consumed	15.9	10.3	7.5	9.8
	Below median	48.9	54.0	30.4	29.2
	Above median	35.2	35.7	**62.2**	**61.0**
white	Below median	49.0	50.6	43.2	29.0
meat	Above median	51.0	49.4	56.8	**71.0**
red	Below median	**73.5**	**85.6**	40.7	33.2
meat	Above median	26.5	14.5	59.3	**66.8**
offals	Not consumed	**82.4**	**88.4**	68.2	60.6
	Consumed	17.6	11.6	31.8	39.4
processed	Below median	**68.9**	**81.7**	30.4	24.4
meat	Above median	31.1	18.4	**69.7**	**75.6**
fish	Below median	57.8	41.6	34.6	17.9
	Above median	42.2	58.4	**65.4**	**82.2**
cheese	Below median	46.7	**65.1**	35.7	40.4
	Above median	53.3	34.9	**64.4**	59.6
potatoes	Below median	**64.9**	**89.4**	44.6	43.4
	Above median	35.1	10.6	55.4	56.6
pulses	Below median	56.4	21.0	34.5	33.8
	Above median	43.6	**79.0**	**65.6**	**66.2**
leafy	Below median	40.1	27.1	32.4	**68.3**
vegetables	Above median	59.9	72.9	**67.6**	31.7
fruiting	Below median	38.0	22.3	26.9	**80.4**
vegetables	Above median	**62.0**	**77.7**	**73.1**	19.6

		Parental class: 1.1		Parental class: 1.2	
		Class 1.1.1 %	Class 1.1.2 %	Class 1.1.1 %	Class 1.1.2 %
root	Not consumed	9.8	8.4	13.8	0.5
vegetables	Below median	39.1	39.2	33.4	8.1
	Above median	51.1	52.5	52.8	91.4
cruciferous	Not consumed	26.7	4.6	15.1	8.3
vegetables	Below median	58.6	25.6	45.3	**90.8**
	Above median	14.7	**69.8**	39.6	0.9
other	Not consumed	7.8	3.8	7.5	0.1
vegetables	Below median	37.6	44.0	40.9	17.9
	Above median	54.6	52.2	51.6	**82.0**
citrus	Not consumed	7.6	2.1	3.3	0.8
fruit	Below median	**60.5**	34.8	47.5	49.9
	Above median	31.9	**63.2**	49.2	49.3
other	Below median	38.9	34.0	43.6	40.7
fruits	Above median	**61.1**	**66.0**	56.4	59.3
sugary	Not consumed	55.8	**62.4**	46.3	44.3
drinks	Consumed	44.2	37.6	53.7	55.7
desserts	Below median	34.8	56.6	32.3	17.8
	Above median	**65.3**	43.4	**67.7**	**82.2**
sugar	Below median	50.8	59.1	37.4	50.6
	Above median	49.2	41.0	**62.6**	49.4

Supplementary Table6.9 Probabilities of consumption for all food groups by dietary patterns derived from LCT. Third level splits, nodes 3.1.1,3.1.2 and 4.1.1, 4.1.2. OPC and ESCC studies, Italy, 1992-2009.

		Parental class: 3.1		Parental class: 4.1	
		Class 3.1.1 %	Class 3.1.2 %	Class 4.1.1 %	Class 4.1.2 %
Size		10.0	90.0	90.0	10.0
milk	Not consumed	52.3	35.0	35.0	52.3
	Below median	28.7	43.9	43.9	28.7
	Above median	19.0	21.2	21.2	19.0
coffee	Below median	41.9	41.9	41.9	41.9
	Above median	58.1	58.1	58.1	58.1
tea	Not consumed	92.1	77.7	**77.7**	**92.1**
	Consumed	7.9	22.3	22.3	7.9
bread	Below median	**84.0**	52.2	52.2	**84.0**
	Above median	16.0	47.8	47.8	16.0
pasta	Below median	31.2	46.9	46.9	31.2
	Above median	**68.8**	53.1	53.1	**68.8**
soup	Below median	**67.4**	**63.1**	**63.1**	**67.4**
	Above median	32.6	36.9	36.9	32.6
eggs	Not consumed	56.3	20.6	20.6	56.3
	Below median	14.8	40.3	40.3	14.8
	Above median	28.9	39.1	39.1	28.9
white	Below median	43.6	**71.0**	**71.0**	43.6
meat	Above median	56.5	29.0	29.0	56.5
red	Below median	**79.5**	46.6	46.6	**79.5**
meat	Above median	20.5	53.4	53.4	20.5
offals	Not consumed	**96.6**	**73.8**	**73.8**	**96.6**
	Consumed	3.4	26.2	26.2	3.4
processed	Below median	45.2	49.0	49.0	45.2
meat	Above median	54.8	51.0	51.0	54.8
fish	Below median	50.8	**70.7**	**70.7**	50.8
	Above median	49.2	29.3	29.3	49.2
cheese	Below median	**69.8**	56.8	56.8	**69.8**
	Above median	30.2	43.2	43.2	30.2
potatoes	Below median	**70.7**	55.1	55.1	**70.7**
	Above median	29.3	44.9	44.9	29.3
pulses	Below median	51.3	**62.3**	**62.3**	51.3
	Above median	48.7	37.7	37.7	48.7
leafy	Below median	**77.2**	**76.9**	**76.9**	**77.2**
vegetables	Above median	22.8	23.1	23.1	22.8
fruiting	Below median	**63.5**	**79.0**	**79.0**	**63.5**
vegetables	Above median	36.5	21.0	21.0	36.5

		Parental class: 3.1		Parental class: 4.1	
		Class 3.1.1 %	Class 3.1.2 %	Class 4.1.1 %	Class 4.1.2 %
root vegetables	Not consumed	**76.6**	**64.5**	**64.5**	**76.6**
	Below median	18.6	29.5	29.5	18.6
	Above median	4.9	6.0	6.0	4.9
cruciferous vegetables	Not consumed	**73.8**	45.4	45.4	**73.8**
	Below median	11.8	43.4	43.4	11.8
	Above median	14.5	11.2	11.2	14.5
other vegetables	Not consumed	**92.1**	31.9	31.9	**92.1**
	Below median	7.4	**62.2**	**62.2**	7.4
	Above median	0.5	5.9	5.9	0.5
citrus fruit	Not consumed	33.7	25.0	25.0	33.7
	Below median	49.4	**62.9**	**62.9**	49.4
	Above median	16.9	12.1	12.1	16.9
other fruits	Below median	**73.2**	**82.4**	**82.4**	**73.2**
	Above median	26.8	17.6	17.6	26.8
sugary drinks	Not consumed	**90.5**	**60.9**	**60.9**	**90.5**
	Consumed	9.5	39.1	39.1	9.5
desserts	Below median	**77.5**	56.1	56.1	**77.5**
	Above median	22.5	43.9	43.9	22.5
sugar	Below median	**74.4**	52.2	52.2	**74.4**
	Above median	25.6	47.9	47.9	25.6

Supplementary Table6.10 Probabilities of consumption for all food groups by dietary patterns derived from LCT. Fourth level split, nodes 1.2.1.1, 1.2.1.2. OPC and ESCC studies, Italy, 1992-2009.

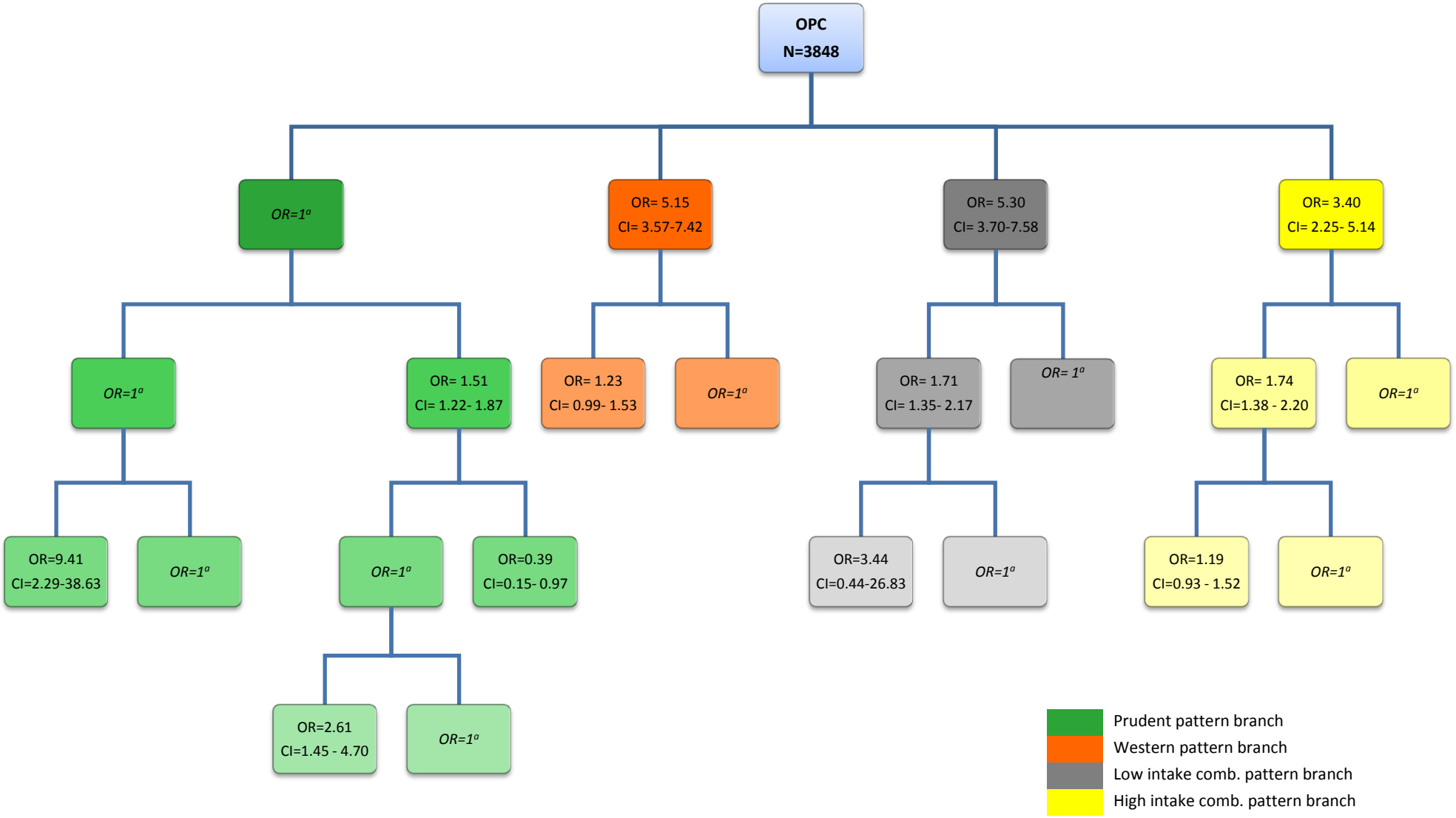
		Parental class: 1.2.1	
		Class 1.2.1.1 %	Class 1.2.1.1 %
Size		86.5	13.6
milk	Not consumed	10.6	7.7
	Below median	42.0	44.8
	Above median	47.3	47.5
coffee	Below median	38.6	46.7
	Above median	**61.4**	53.3
tea	Not consumed	42.9	35.3
	Consumed	57.1	**64.7**
bread	Below median	54.7	**67.6**
	Above median	45.3	32.4
pasta	Below median	41.6	34.2
	Above median	58.4	**65.8**
soup	Below median	38.9	49.3
	Above median	**61.1**	50.7
eggs	Not consumed	6.6	10.6
	Below median	29.5	30.2
	Above median	**63.9**	59.2
white	Below median	43.1	30.0
meat	Above median	56.9	**70.0**
red	Below median	40.0	33.5
meat	Above median	**60.0**	**66.5**
offals	Not consumed	**67.0**	**61.5**
	Consumed	33.0	38.5
processed	Below median	29.8	25.1
meat	Above median	**70.2**	**75.0**
fish	Below median	34.8	17.9
	Above median	**65.2**	**82.1**
cheese	Below median	35.9	40.7
	Above median	**64.1**	59.4
potatoes	Below median	43.9	43.7
	Above median	56.1	56.3
pulses	Below median	34.9	33.7
	Above median	**65.2**	**66.3**
leafy	Below median	33.0	**69.5**
vegetables	Above median	**67.0**	30.5
fruiting	Below median	28.5	**78.5**
vegetables	Above median	**71.5**	21.5

		Parental class: 1.2.1	
		Class 1.2.1.1 %	Class 1.2.1.1 %
root	Not consumed	13.9	0.4
vegetables	Below median	33.9	8.4
	Above median	52.2	**91.2**
cruciferous	Not consumed	15.5	8.8
vegetables	Below median	47.5	**90.4**
	Above median	37.0	0.8
other	Not consumed	7.9	0.1
vegetables	Below median	42.2	17.0
	Above median	49.9	**82.9**
citrus	Not consumed	3.6	0.6
fruit	Below median	49.3	49.7
	Above median	47.1	49.7
other	Below median	45.3	41.6
fruits	Above median	54.7	58.4
sugary	Not consumed	43.4	45.9
drinks	Consumed	56.6	54.2
desserts	Below median	32.5	19.2
	Above median	**67.5**	**80.8**
sugar	Below median	34.1	54.2
	Above median	**65.9**	45.8

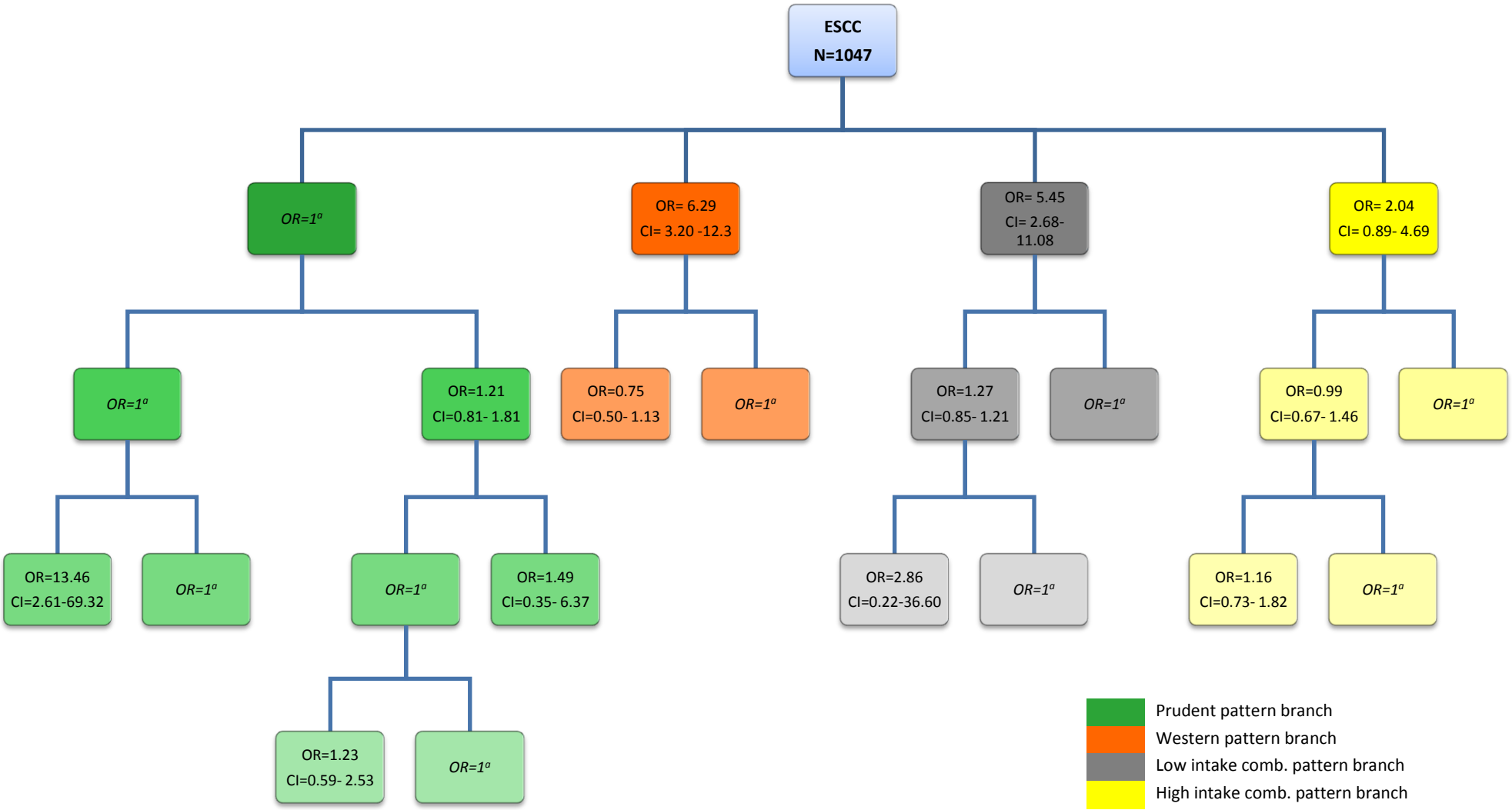
Supplementary Table. 6.11 Fit statistic for LCT splits (1-class vs 2-class model), till the last split according to each statistic. OPC and ESCC studies, Italy,1992-2009.

Parental class	Nr. classes	BIC	AIC	AIC3
1	1	47223.83	46891.88	46955.88
	2	47342.13	46839.02	46936.02
2	1	42137.62	41809.16	41873.16
	2	42223.7	41725.89	41822.89
3	1	28437.36	28135.98	28199.98
	2	28553.83	28097.04	28194.04
4	1	28182.48	27882.75	27946.75
	2	28301.29	27847.01	27944.01
1.1	1		26202.14	26266.14
	2		26195.25	26292.25
1.2	1		20167.17	20231.17
	2		20160.18	20257.18
2.1	1		23499.54	23563.54
	2		23509.92	23606.92
2.1	1		17737.18	17801.18
	2		17762.51	17853.94
3.1	1		16560.11	16624.11
	2		16550.23	16647.23
3.2	1		11167.74	11231.74
	2		11187.88	11284.88
4.1	1		18624.56	18688.56
	2		18609.44	18706.44
4.2	1		8955.236	9019.236
	2		8972.499	9069.499
1.1.1	1		22334.79	
	2		22351.57	
1.1.2	1		3622.194	
	2		3659.824	
1.2.1	1		18896.15	
	2		18888.44	
1.2.2	1		1149.922	
	2		1189.69	
3.1.1	1		14918.28	
	2		14929.09	
3.1.2	1		1503.369	
	2		1544.502	
4.1.1	1		12170.67	
	2		12197.18	
4.1.2	1		6238.297	
	2		6259.264	
1.2.1.1	1		16396.36	
	2		16420.36	
1.2.1.2	1		2310.993	
	2		2356.082	

Supplementary Figure 6.1. Unadjusted Odds ratios (OR) and corresponding 95% confidence intervals (CIs) for oral/pharyngeal cancer risk at each split. Italy,1992-2009. [a]Reference category for the split.



Supplementary Figure 6.2. Unadjusted Odds ratios (OR) and corresponding 95% confidence intervals (CIs) for esophageal cancer risk at each split. Italy,1992-1997. [a]Reference category for the split.



Supplementary Table.6.12 Odds ratios (OR) and corresponding 95% confidence intervals (CIs) for OPC at each split in models unadjusted and adjusted for known confounders with ML estimation. Italy,1991-2009. [a]Reference category for the split.

Level	Dietary Patterns	Unadjusted OR (95% CI)	Adjusted OR (95% CI) [b]
First level split	1 [a]	1	1
	2	5.16 (3.58 – 7.44)	1.91 (1.41 – 2.58)
	3	5.31 (3.71 – 7.60)	2.14 (1.58 – 2.91)
	4	3.41 (2.26 – 5.15)	1.03 (0.74 – 1.46)
Second level splits	1.1 [a]	1	1
	1.2	1.5 (1.22 – 1.87)	0.97 (0.78 – 1.21)
	2.1	1.23 (0.99 – 1.53)	0.91 (0.72 -1.14)
	2.2 [a]	1	1
	3.1	1.72 (1.35 – 2.18)	0.88 (0.70 – 1.12)
	3.2 [a]	1	1
	4.1	1.75 (1.38 – 2.20)	1.20 (0.95 – 1.52)
	4.2 [a]	1	1
Third level splits	1.1.1	12.05 (1.55 – 97.74)	1.87 (1.09 – 3.23)
	1.1.2 [a]	1	1
	1.2.1 [a]	1	1
	1.2.2	0.38 (0.14 – 0.98)	0.50 (0.18 -1.35)
	3.1.1	1.58 (0.77 -3.25)	0.85 (0.47 -1.55)
	3.1.2 [a]	1	1
	4.1.1	1.19 (0.93 – 1.52)	0.89 (0.94 – 1.57)
	4.1.2 [a]	1	1
Fourth level split	1.2.1.1	2.65 (1.46 – 4.08)	1.47 (0.97 – 2.23)
	1.2.1.2 [a]	1	1

[b] Adjusted for sex, age, education, BMI, tobacco and alcohol intake

Supplementary Table.6.13 Odds ratios (OR) and corresponding 95% confidence intervals (CIs) for ESCC at each split in models unadjusted and adjusted for known confounders with ML estimation. Italy,1992-1997. [a]Reference category for the split.

Level	Dietary Patterns	Unadjusted OR (95% CI)	Adjusted OR (95% CI) [b]
First level split	1 [a]	1	1
	2	6.35 (3.22 – 15.54)	3.24 (1.78 – 5.87)
	3	5.50 (2.69 – 11.24)	2.86 (1.47 – 5.58)
	4	2.05 (0.89 – 4.74)	0.88 (0.39 – 2.00)
Second level splits	1.1 [a]	1	1
	1.2	1.22 (0.81 -1.81)	0.90 (0.58 – 1.40)
	2.1	0.75 (0.50 – 1.12)	0.67 (0.43 – 1.05)
	2.2 [a]	1	1
	3.1	1.28 (0.86 – 1.90)	0.80 (0.53 – 1.21)
	3.2 [a]	1	1
	4.1	0.99 (0.67 – 1.46)	0.84 (0.55 – 1.29)
	4.2 [a]	1	1
Third level splits	1.1.1	4.17*e^10 (0-inf)	8.72 (1.07 – 70.75)
	1.1.2 [a]	1	1
	1.2.1 [a]	1	1
	1.2.2	1.55 (0.36 – 6.70)	1.41 (0.19 – 10.22)
	3.1.1	7.91 (0.21 -301.96)	7.49 (0.00 -16652.57)
	3.1.2 [a]	1	1
	4.1.1	1.16 (0.73 – 1.82)	1.24 (0.75 – 2.06)
	4.1.2 [a]	1	1
Fourth level split	1.2.1.1	1.23 (0.60 -2.54)	0.68 (0.34 – 1.38)
	1.2.1.2 [a]	1	1

[b] Adjusted for sex, age, education, BMI, tobacco and alcohol intake