Shared and study-specific dietary patterns and their association with head and neck cancer risk in an international consortium

Main text - word count: 4786.

Total word count: 7000.

Abstract – Word count: 250 words

Background: A few papers have considered reproducibility of *a posteriori* dietary patterns across populations, as well as pattern associations with head and neck cancer risk when multiple populations are available.

Methods: We used individual-level pooled data from 7 case-control studies (3,844 cases; 6,824 controls) participating in the International Head and Neck Cancer Epidemiology consortium. We simultaneously derived shared and study-specific *a posteriori* patterns with a novel approach named "multi-study factor analysis" applied on 23 nutrients. We derived odds ratios (ORs) and 95% confidence intervals (CIs) for cancers of the oral cavity and pharynx combined, and larynx, from logistic regression models.

Results: We identified 3 shared patterns that were reproducible across studies (75% variance explained): the *Anti-oxidant vitamins and fiber* and the *Fats* patterns were inversely associated with oral and pharyngeal cancer risk (OR=0.57, CI: 0.41-0.78; OR=0.80, CI: 0.67-0.95, respectively, highest vs. lowest score quintile category). The *Animal products and cereals* and the *Fats* patterns were positively associated with laryngeal cancer risk (OR=1.51, CI: 1.06-2.13; OR=1.83, CI: 1.44-2.33), whereas a linear inverse trend in laryngeal cancer risk was evident for the *Anti-oxidant vitamins and fiber* pattern. We also identified 4 additional study-specific patterns, one for each of the 4 US studies examined. We named them all as *Dairy products and breakfast cereals* and 2 were associated with oral and pharyngeal cancer risk.

Conclusion: Multi-study factor analysis provides insight into pattern reproducibility, and supports previous evidence on cross-country reproducibility of dietary patterns and on their association with head and neck cancer risk.

Keywords**:** diet, Mediterranean; diet, high-fat; diet, Western; head and neck neoplasms; laryngeal neoplasms; mouth neoplasms; pharyngeal neoplasms; reproducibility of results.

Introduction

Dietary patterns have long been recognized as a useful tool for assessing overall diet, or key aspects of the diet, and its contribution to health and disease, as they synthesize multiple related dietary components (food items, food groups, or nutrients) in one or more combined variables. Interest in dietary patterns is also motivated from observations that: 1. foods may have interactive effects on bioavailability and circulating levels of nutrients, and thus potentially on disease risk[1]; 2. data dimensionality and multiple estimation challenges affect the statistical analysis of many single dietary components.

Despite these advantages, little research in nutritional epidemiology has focused on statistical methods specific to dietary patterns. Most of the recent studies use standard multivariate statistical methods [principal component analysis, factor analysis, cluster analysis] to empirically identify *a posteriori* patterns, or consider *a priori*-defined ones, where scores are assigned according to an individual's adherence to dietary recommendations or other criteria. The lack of consistent methodology to derive dietary patterns has severely limited the ability to draw inferences from the results of studies based on this approach[2] and only the more recent version of the Dietary Guidelines for Americans has included evidence on patterns[3].

Reproducibility and validity of dietary patterns have not been extensively assessed[4]. Since 2012, the Dietary Patterns Methods Project supported standardized analyses on selected *a priori* patterns and mortality outcomes in three cohorts in the United States[2]. Other studies assessed the same issues on a study-level scale. In most of these papers, reproducibility of *a posteriori* patterns was investigated using dietary information from a common dietary assessment tool administered multiple times in the same sample; pairs of similar patterns were compared across occasions using correlation coefficients between scores[5], congruence coefficients between loadings[6,7], or measures of agreement between

clustering assignments[8]. Validity of the identified patterns was mostly assessed looking at their association with socio-demographic characteristics, lifestyle habits, nutrient/food profiles from the same dietary source, nutritional biomarkers, markers of disease, or a disease of interest[9,10]. In addition, relative validity of patterns was assessed within one study looking at the correlation coefficients between scores of similar patterns derived on a test and a reference dietary source[5].

A major drawback of dietary patterns is their applicability to different populations. This is especially true for the *a posteriori* patterns, which are meant to reflect existing dietary behavior in a population and may, therefore, be difficult to replicate in other settings[7]. "External reproducibility" measures the extent to which similar patterns are seen in diverse populations and it is here intended as synonymous with "replicability"[3]. External reproducibility of *a posteriori* patterns derived with principal component or factor analysis is of interest in this paper. In addition, we are interested in their validity, here intended as their real association with disease risk. A few papers have recently explored these issues[6,7,11,12] and supported the idea that some patterns are reproducible and valid across populations. However, there is still no statistical approach to answer the key issue of external reproducibility of dietary patterns. In addition, a focus on external reproducibility does not exclude that additional study-specific patterns could be relevant for descriptive purposes or for the assessment of disease risk.

de Vito et al.[13] recently proposed multi-study factor analysis as a generalization of factor analysis able to handle multiple studies simultaneously. Multi-study factor analysis leads to the identification of shared factors, which are common to all the studies, as well as additional study-specific factors for some of the studies in an integrated approach based on the maximum likelihood. This approach tackles the issue of external reproducibility of dietary patterns from a different perspective: the reproducible patterns are those that each study population shares with all the others, within a statistical model that integrates

information from all studies and allows to choose the best number of shared and study-specific patterns.

The International Head and Neck Cancer Epidemiology (INHANCE) consortium[14,15] was established in 2004 to elucidate the etiology of head and neck cancer through pooled analyses of individual-level data from several studies. Dietary habits have been previously investigated within the consortium[16-18]. A *posteriori* patterns were identified with a standard principal component factor analysis, where five study-specific datasets (~7,500 subjects) providing information on a common list of nutrients were standardized and analyzed as they were a single dataset[18]. In more recent versions of the consortium dataset, two other studies (~3,200 extra subjects) provide comparable information on nutrient intakes[17]. In addition, any consortia provide standardized definitions of exposures, outcomes, and confounding factors, and this limits possible sources of variability in the assessment of reproducibility and validity of dietary patterns.

The main objective of this paper is to assess external reproducibility and validity of *a posteriori* patterns derived with multi-study factor analysis within the International Head and Neck Cancer Epidemiology consortium. This will address the following research questions: 1) pattern reproducibility: are there consistent and empirically estimable eating patterns across all the International Head and Neck Cancer Epidemiology populations?; 2) are there one or more additional study-specific eating patterns?; 3) pattern validity: are the shared and study-specific patterns identified by multi-study factor analysis associated with the risk of cancers of the oral cavity and pharynx combined and larynx?

Materials and methods

Design and subjects

We extracted from version 1.5 of the International Head and Neck Cancer Epidemiology consortium pooled dataset seven case-control studies[19-25] that provided a sufficiently large

list of common nutrients to be used for multi-study factor analysis. Three studies were conducted in Europe and four in the United States. Other details on the individual studies, including number of cases and controls, harmonization of data, and data pooling methods have been previously described[14] and are summarized in **eTable 1**. Informed consent was obtained from all subjects within the framework of the original studies. The investigations were approved by the relevant institutional review boards, according to the rules specific to each country at the time of data collection.

Selection of subjects

Cases were included if their cancer had been classified as an invasive cancer of oral cavity, oropharynx, hypopharynx, oral cavity or pharynx not otherwise specified, larynx, or head and neck cancer unspecified. Cases with cancers of the salivary glands or of the nasal cavity/ear/paranasal sinuses were excluded[15]. Corresponding controls from the original studies were included too, and this gave a sample of 12,295 subjects initially considered for inclusion.

We removed from our analysis: 1. subjects with missing or implausible (<500 or >5,500 kcal) values of daily non-alcohol energy intake (1,381 and 224 subjects, respectively, 1,605 total subjects); 2. cases without information on the site of origin of cancer (22 cases). Thus, our analysis included a total of 10,668 subjects, with 3,844 head and neck cancer cases and 6,824 controls. There were 711 oral cavity cancer cases, 1,088 oropharyngeal and 343 hypopharyngeal cancer cases (1,431 pharyngeal cancer cases), 411 unspecified oral cavity/pharynx cases (2,553 cases of oral and pharyngeal cancers combined), and 1,291 laryngeal cancer cases.

Specification of variables

Intakes of total energy, several nutrients, and food components were derived by combining

information from study-specific food-frequency questionnaires (FFQs), assessing subjects' usual diet during a reference period preceding cancer diagnosis for cases or interview for controls, with information from country-specific food composition databases[26,27].

From the study-specific lists, we selected twenty-three nutrients and food components to assess their potential joint role in head and neck cancer risk. Checks on nutrient definitions, reference periods of intake and measurement units were conducted across studies. We increased comparability by using intakes from foods only[18]. Nutrient intakes were expressed on a daily basis.

Statistical analysis

In the following, we describe the main steps of the statistical analysis, which includes dietary pattern identification and assessment of related head and neck cancer risk. We included in the main text methods and results from a multi-study factor analysis based on controls only (controls-only multi-study factor analysis). We also carried out a parallel analysis based on the combined sample of head and neck cancer cases and controls (cases+controls multi-study factor analysis). Results of the latter analysis are presented in the **eAppendix** (**eAppendix – Results**, **eFigure 4**, **eTable 6**, and **eTable 7**).

Data preprocessing

We log-transformed (base e) the study-specific raw dietary data to improve adherence to the assumption of normality of the shared and study-specific factors, as well as of the study-specific errors, as required by multi-study factor analysis.

Factorability of the correlation matrices

We explored whether the seven study-specific correlation matrices and the merged data

correlation matrix (generated by combining data from all the studies) of the log-transformed intakes of control subjects were factorable. We considered Bartlett's test of sphericity, overall and individual measures of sampling adequacy[28]. Given the reassuring results obtained, we applied multi-study factor analysis to identify *a posteriori* shared and study-specific dietary patterns on the set of control subjects.

Identification of dietary patterns

We carried out multi-study factor analysis[13] on the correlation matrix of the study-specific log-transformed (base e) data from control subjects to describe the variance-covariance structure among the selected P nutrients in terms of a few underlying unobservable shared and study-specific factors, known as dietary patterns. Multi-study factor analysis explicitly identifies K factors shared among all the available S studies, as well as J_s potential study-specific factors, giving a total of $T_s = K + J_s$ factors, s=1,…,S. This is evident from the corresponding formula of the study-specific correlation matrix which reflects the simultaneous presence of the shared and study-specific factors: $\Sigma_s = \Phi \Phi^T + \Lambda_s \Lambda_s^T + \Psi_s$, where Φ is the (P × K) shared factor-loading matrix, Λ_s is the (P × J_s) study-specific factor-loading matrix and Ψ_s is the (P × P) study-specific covariance matrix of the error term.

The method adopts an integrated approach based on the maximum likelihood and takes advantage of the Expectation Conditional Maximization algorithm[29], which is a generalization of the Expectation Maximization algorithm used in standard maximum likelihood factor analysis.

The choice of the number of shared, K, and total, T_s (shared and study-specific), factors to be included in the model, as well as the final model selection, are handled within multi-study factor analysis.

In detail, to choose the number of factors to retain, we first estimated the total number of factors, T_s, for each of the studies, using a combination of standard techniques for factor

analysis, including Horn's parallel analysis, Cattell's scree plot, and the Steiger's root mean square error of approximation index[30]. Next, we used the Akaike Information Criterion[31] on the multi-study factor analysis model to select the number of shared factors, K. The number of study-specific factors, J_s, for each study was then found by difference as T_s - K. A global Akaike Information Criterion was also used to identify the optimal pair (K, J_s).

We applied a varimax rotation to the factor-loading matrix of the shared factors to achieve a better-defined loading structure. To name the 'dominant nutrients'[28], we used nutrients with a shared (rotated) factor loading of at least 0.60 or a study-specific (unrotated) factor loading of at least |0.25|.

Factor scores indicate the degree to which each subject's diet conforms to one of the identified patterns. We calculated factor scores in multi-study factor analysis by adapting the standard Bartlett and Thurstone methods for factor analysis[30,32]. In detail, we calculated a factor score for each subject (case or control) and factor within each study by using the study-specific correlation matrix of the log-transformed data (from the overall sample of cases and controls) and the factor-loading matrix [Φ | Λ_s] (obtained juxtaposing shared and study-specific factor loadings derived from the controls-only analysis).

The correlations between the two types of scores were 0.99 for all factors, so we continued with the Bartlett method, since its scores have zero sample mean vector and zero sample covariances[30].

We evaluated the internal consistency of patterns using standardized Cronbach's alpha and *alpha-when-item-deleted* coefficients[28]; we also assessed the internal stability of the patterns using a split-half approach (**eAppendix – Statistical analysis**). Similarly, we conducted stratified multi-study factor analyses by sex. Finally, we applied our approach to the subset of the International Head and Neck Cancer Epidemiology data used to derive *a posteriori* patterns in 2012[18], to assess if multi-study and standard maximum likelihood

factor analyses identify similar shared patterns with a similar performance. We compared the distributions and standard errors of the factor loadings computed from 100 bootstrapped random sets of our original five studies under the two approaches.

Estimates of the association between the identified dietary patterns and head and neck cancer

We grouped participants into either five or three categories according to quintiles (shared factors) or tertiles (study-specific factors) of factor scores among the controls.

We estimated the odds ratios (ORs) and the corresponding 95% confidence intervals (CIs) of oral and pharyngeal cancers combined and laryngeal cancer, separately, for each score quantile category using unconditional multiple logistic regression models[33]. We fitted separate models for each factor, a shared factors regression model, and a composite regression model including all the shared and one study-specific factor at a time. Study-specific factors were analyzed only for the studies in which they were identified. Models included adjustments for age, sex, race, study center (when appropriate), education, pack-years of cigarette smoking, cigar smoking status, pipe smoking status, and alcohol drinking intensity (see **Table 1** for the covariate categories used). We also considered extra adjustments for non-alcohol energy intake (entered as study-specific quintile categories built on control subjects) or for supplement use of vitamin C, vitamin E, or beta-carotene (one variable at a time entered as never/ever in lifetime, when information was available). For all the models, we adopted a complete-case approach to the analysis.

To accommodate heterogeneity of the shared patterns' associations across studies, we estimated the corresponding ORs and CIs using a random-slope generalized linear mixed model with logit link function and binomial family[34] (eAppendix – Statistical analysis). All statistical tests were two-sided. Calculations were carried out using the open-source statistical computing environment R[35], with its libraries "statmod"[36], "psych"[37], "nFactors"[38],

"ggplot2"[39], and "lme4"[40], and a specialized code for performing multi-study factor analysis as described in the "Statistical analysis" section.

Results

Table 1 shows selected characteristics of cancer cases and controls. Over 90% of the subjects were white. Studies from Europe contributed approximately 50% of cases of oral and pharyngeal cancers combined, 60% of cases of laryngeal cancer, and over 60% of controls. Cases were more often tobacco smokers and alcohol drinkers than controls.

Correlations among individual nutrients were strong enough to suggest that the seven study-specific correlation matrices and the merged data correlation matrix on controls were factorable (eAppendix – Results).

eTable 2 shows results from the model selection procedure for fixed pre-selected values of T_s. The selected model presented three patterns shared among all the studies and one additional study-specific pattern for each of the four studies from the United States (giving a total of four study-specific patterns).

Table 2 and e**Figure 1** present the factor-loading matrix for the shared factors. These factors explained 75% of the total variance in each study-specific dataset. The rotation made shared pattern loadings positive, in such a way that only the magnitude of each loading (and not its sign) was used to name the factors. The first factor, named *Animal products and cereals*, had the greatest loadings on total protein, zinc, phosphorus, riboflavin, sodium, niacin, thiamin, cholesterol, calcium, vitamin B6, iron, potassium, and total carbohydrates. The second factor, named *Anti-oxidant vitamins and fiber*, had the greatest loadings on vitamin C, total fiber, total folate, potassium, total carotene, and vitamin B6. The third factor, named *Fats*, had the greatest loadings on monounsaturated and polyunsaturated fatty acids, vitamin E, and saturated fatty acids.

eTable 3 and eFigure 1 show the factor-loading matrix for the four study-specific factors.

These factors explained 5%, 3%, 6%, and 3%, respectively, of the total variance in each of the four US datasets. The *Los Angeles-specific* pattern had the largest positive loadings on calcium, phosphorus, and saturated fatty acids and the largest negative loadings on niacin and vitamin B6. The *Boston-specific* pattern had the largest positive loading on calcium and the largest negative one on niacin. The *Memorial Sloan Kettering Cancer Center (MSKCC)-specific* pattern had the largest positive loadings on calcium and phosphorus and the largest negative loadings on vitamin E, niacin, thiamin, vitamin B6, zinc, folate, and iron. Finally, the *North Carolina (2002-2006)-specific* pattern had the largest positive loadings on calcium and phosphorus, whereas niacin had the largest negative loading at -0.21. Given the consistent presence of the calcium vs. niacin pair in the study-specific factor loadings, we named all the four study-specific factors as *Dairy products and breakfast cereals*.

Nutrient communalities were generally satisfactory, with most portions of the nutrient variances contributed by the retained factors of 0.60 or more (**eTable 4**).

The internal consistency of patterns was high: standardized Cronbach's alphas ranged from 0.91 (*Boston-specific* pattern) to 0.98 (*Animal products and cereals* pattern) and most of the standardized Cronbach's *alphas-when-item-deleted* were lower than the corresponding alphas for the same factor. The internal stability of the sets of patterns identified in the two split-samples was also good (eAppendix – Results). In addition, although the selected model differed for males and females (males: three shared patterns and one study-specific pattern for each of the four US studies, as in the main analysis; females: four shared patterns), the *Anti-oxidant vitamins and fiber* and the *Fats* patterns have similar factor loadings and percentages of explained variances in strata of males and females, and were in agreement with the main analysis (eAppendix – Results).

Finally, when we carried out multi-study factor analysis on the subset of five studies analyzed in Edefonti et al.[18], the three previously identified patterns were satisfactorily

reproduced in the form of our shared patterns. The boxplots representing the distribution of the shared pattern loadings on the 100 bootstrapped random sets were narrower with multi-study than with standard maximum likelihood factor analysis (e**Figure 2**). Multi-study factor analysis estimated one extra pattern for each of the US studies: the American study-specific patterns were similar to the corresponding ones from the more recent analysis on seven studies. Percentages of explained variances were similar for the corresponding patterns in both the analyses (**eFigure 3**).

Table 3 and eFigure 1 give separate ORs and the corresponding CIs for oral and pharyngeal cancers combined, and laryngeal cancer, by quintiles of factor scores for the shared patterns. In the presence of appreciable heterogeneity of the associations across studies for both cancers, we reported results from the mixed-effects composite models including the three shared patterns, together with potential confounders. The selected mixed-effects models included random-effects terms for the *Anti-oxidant vitamins and fiber* pattern only for oral and pharyngeal cancers combined and random-effects terms for both the *Animal products and cereals* and the *Anti-oxidant vitamins and fiber* patterns for laryngeal cancer. The *Animal products and cereals* pattern was positively associated with laryngeal cancer risk (OR=1.51, 95% CI: 1.06-2.13 for the highest versus the lowest score quintile category, p for trend<0.001). Higher intakes of the dominant nutrients for the *Anti-oxidant vitamins and fiber* pattern were inversely related to oral and pharyngeal cancers combined (OR=0.57, 95% CI: 0.41-0.78, p for trend=0.003) and to laryngeal cancer risk, for which a protection was found from the second quintile category onward, although the CI for the last quintile category includes 1 (OR = 0.62, 95 % CI 0.37–1.04). The *Fats* pattern was inversely associated with oral and pharyngeal cancers combined and positively associated with laryngeal cancer: the ORs were 0.80 (95% CI: 0.67-0.95, p for trend=0.019) and 1.83 (95% CI: 1.44-2.33, p for trend<0.001), respectively.

eTable 5 and eFigure 1 give separate ORs and the corresponding CIs for cancers of the

oral cavity and pharynx combined and larynx, by tertiles of the study-specific factor scores. Results refer to the fixed-effects composite models including one study-specific pattern at a time, together with the shared ones and potential confounders. The *Dairy products and breakfast cereals* pattern identified in the Los Angeles study was inversely associated with oral and pharyngeal cancer risk (OR=0.66, 95% CI: 0.44-0.99, p for trend=0.048), but that identified in the Boston study was positively associated with the same cancer site (OR=1.55, 95% CI: 1.06-2.28, p for trend=0.034).

After adjustment for non-alcohol energy intake, results were still in agreement with those of the main analysis: for the *Anti-oxidant vitamins and fiber* patterns, the ORs for the last quintile category were 0.58 (95% CI: 0.42-0.80) for oral and pharyngeal cancers and 0.57 (95% CI: 0.34-0.96) for laryngeal cancer; for the *Fats* pattern, the OR of laryngeal cancer was 1.59 (95% CI: 1.19-2.12). In addition, the strength of the associations was weaker for the *Fats,* Los Angeles-specific, and Boston-specific patterns and oral and pharyngeal cancers combined and for the *Animal products and cereals* pattern and laryngeal cancer (eAppendix – Results). Results from mixed-effects models were derived with an approximated solution based on penalized iteratively reweighted least squares. Similarly, the extra adjustment for supplement use of vitamin C, or vitamin E, or beta-carotene in study-specific models provided results that were similar to the main analysis, although based on 38% (Los Angeles study) and 45-61% (Boston study) of the original samples (eAppendix – Results).

Discussion

We introduce multi-study factor analysis in nutritional epidemiology to give insight into external reproducibility and validity of *a posteriori* dietary patterns. In our application, we found that study populations from Italy, Switzerland, and the United States share three reproducible patterns characterized by consumption of animal products and cereals,

vitamin-rich foods, and fats, respectively. In addition, each of the American studies is characterized by a somewhat similar additional pattern, which opposes calcium and niacin as dominant nutrients. We also found that five of the seven patterns are associated with the risk of oral and pharyngeal cancers and/or laryngeal cancer.

Several reasons may explain why the American studies express an additional study-specific pattern. There might be true differences between Europe and the United States in terms of breakfast energy and composition, cereal content, and fortification policies, as well as sources of dairy products and cereals. Indeed, the shortlist of foods rich in niacin and low in calcium (or viceversa), identified from our analysis of US food composition tables[26] and FFQ-specific food sources[41], includes: milk, cheese, yoghurt, instant and filter coffee, as well as cereal products in general, including most breakfast cereals. There might be also a FFQ effect: the studies from Europe share the same FFQ, whereas those from the United States do not, except for two versions of the Block FFQ. The length of the American FFQs is also different across studies. In addition, the European studies are hospital-based, whereas the American studies are population-based investigations, so it is not possible to disentangle the role of control source, country, and type of FFQ. Moreover, in the current application, the percentage of variance explained by the study-specific patterns is relatively small. To provide a comparison, we fitted a (sub-optimal) multi-study factor analysis where we forced all the studies to express one extra study-specific pattern, in addition to the three shared ones. The new study-specific patterns had a median percentage of explained variance of 6 and this is in line with what is expected when cultural particularities are expressed as dietary patterns. Similarly, the number of cases and controls was limited in some of the US studies, and this may limit our ability to identify associations between study-specific patterns and head and neck cancer risk, especially for laryngeal cancer. However, in general, one of the functions of the study-specific factors is to improve comparability of the shared factors. Thus, we expect that, when model

selection justifies their inclusion, study-specific factors would contribute positively to the analysis even when their interpretation is difficult and their association with disease is weak.

To our knowledge, four studies have explored external reproducibility and validity of *a posteriori* patterns derived with principal component or factor analysis[6,7,11,12]. One[11] assessed reproducibility of principal-component-analysis- and confirmatory-factor-analysis-derived patterns in two samples from France and Spain, but did not provide any cross-country comparison. A paper by the European Prospective Investigation into Cancer and Nutrition (EPIC) group[12] showed that an "overall principal component analysis" solution captures most of the variance in any center and further assessed the validity of the "overall principal component analysis"-derived patterns in terms of foods, nutrients and lifestyle factors. A third study[6] assessed reproducibility of principal-component-analysis-based patterns on the same food groups in two different Spanish studies with similar FFQs and concluded that at least the widely prevalent *Western* pattern was reproduced across studies. Another study from Spain[7] provided a sound statistical approach to compare the previous Spanish patterns[6] with other *a posteriori* ones identified from the literature, and assessed the related breast cancer risk. The previous papers showed that: 1. independently of the approach used, there is some evidence in support of reproducibility and validity of dietary patterns across studies; 2. when individual-level data are available[6,12], researchers should consider statistical models that allow for pattern sharing across studies.

Our analysis had several strengths. The International Head and Neck Cancer Epidemiology consortium offers an interesting set-up to apply multi-study factor analysis. Besides the harmonization of exposures, confounders, and outcomes, we have information from different studies within the same countries, some of which (the European ones and two of the American ones) share the same FFQ. We also observed reassuring

results in our comparison between multi-study and standard factor-analysis-derived patterns on the subset of five studies analyzed in Edefonti et al.[18]: constructing shared patterns across the studies with multi-study factor analysis identifies more stable patterns.

We acknowledge that the International Head and Neck Cancer Epidemiology consortium includes case-control studies and this may limit the strength of the evidence on head and neck cancer risk. Multi-study factor analysis is, however, applicable to any consortium or network of consortia, including cohort-based consortia. In addition, within version 1.5 of the pooled dataset, three other studies provided information on nutrient intakes[17]. However, they provided shorter lists of nutrients and/or information on single and total nutrients (i.e. total carotene or beta-carotene equivalents and/or single carotenoids) not completely comparable with the studies used here. If we used all the ten available studies, the number of common nutrients would have been reduced to ten.

Our experience, as well as evidence from the EPIC group[12], suggest that partial sharing of patterns can be a good compromise between forcing the studies to express the same set of patterns[18] and allowing them to express separate sets of patterns (to be combined *ad hoc* in a subsequent analysis). If this possibility is integrated within an approach that includes a formal assessment of the number of shared and study-specific patterns, the corresponding statistical model provides a principled way to derive a realistic but parsimonious representation of dietary behavior across populations. The use of objective criteria, like the Akaike Information Criterion, to choose the number of factors is also a step forward, as visual inspection of scree plots and the (less stringent) eigenvalue>1 criterion are usually used in standard factor analysis. Moreover, we integrated multi-study factor analysis with standard checks of internal stability and internal consistency of the identified patterns.

Multi-study factor analysis has also limitations. The best-fitting number of study-specific factors selected in the current application is 1; this prevents the possibility of rotating the

study-specific factor-loading matrices, at the expense of factor interpretation. In its current implementation, multi-study factor analysis does not deal with situations where some studies provide nutrient information on total components and others on single components only. Furthermore, it does not deal with situations where some studies share one or more additional factors that are not common to all the available studies. In particular, as in our analysis each of the four American studies showed a similar study-specific pattern, it would be interesting to explore whether there is one shared “American” pattern through an extension of the current approach.

In the current application we derived multi-study dietary patterns based either on controls only, or on cases and controls together. The choice between these two approaches is an open problem for standard factor analysis as well, and the literature lacks methodological analyses to guide this choice. Our results suggested that differences in factor loadings and percentages of explained variances were small. However, although the point estimates for head and neck cancer risk were generally similar, corresponding standard errors were higher under the controls-only analysis. Future studies should address the comparison between these approaches.

In conclusion, the use of multi-study factor analysis in nutritional epidemiology identified reproducible eating patterns across the International Head and Neck Cancer Epidemiology populations from Europe and the USA, as well as US-specific eating behaviors, that may be validly associated with head and neck cancer risk. These results may inform the next releases of national dietary guidelines in Europe and in the US and provide the basis for future efforts of integration of information on *a posteriori* patterns from different countries.

References

1. Hu FB. Dietary pattern analysis: a new direction in nutritional epidemiology. *Curr Opin Lipidol* 2002;**13**(1):3-9.
2. Liese AD, Krebs-Smith SM, Subar AF, et al. The Dietary Patterns Methods Project: synthesis of findings across cohorts and relevance to dietary guidance. *J Nutr* 2015;**145**(3):393-402.
3. Broman K, Cetinkaya-Rundel M, Nussbaum A, et al. Recommendations to Funding Agencies for Supporting Reproducible Research, American Statistical Association, Available online at: http://www.amstat.org/asa/files/pdfs/pol-reproducibleresearchrecommendations.pdf (16 April 2018, date last accessed).
4. Edefonti V, Randi G, La Vecchia C, Ferraroni M, Decarli A. Dietary patterns and breast cancer: a review with focus on methodological issues. *Nutr Rev* 2009;**67**(6):297-314.
5. Hu FB, Rimm E, Smith-Warner SA, et al. Reproducibility and validity of dietary patterns assessed with a food-frequency questionnaire. *Am J Clin Nutr* 1999;**69**(2):243-9.
6. Castello A, Lope V, Vioque J, et al. Reproducibility of data-driven dietary patterns in two groups of adult Spanish women from different studies. *Br J Nutr* 2016;**116**(4):734-42.
7. Castello A, Buijsse B, Martin M, et al. Evaluating the Applicability of Data-Driven Dietary Patterns to Independent Samples with a Focus on Measurement Tools for Pattern Similarity. *J Acad Nutr Diet* 2016;**116**(12):1914-1924 e6.
8. Lo Siou G, Yasui Y, Csizmadi I, McGregor SE, Robson PJ. Exploring statistical approaches to diminish subjectivity of cluster analysis to derive dietary patterns: The Tomorrow Project. *Am J Epidemiol* 2011;**173**(8):956-67.
9. Hu FB, Rimm EB, Stampfer MJ, Ascherio A, Spiegelman D, Willett WC. Prospective study of major dietary patterns and risk of coronary heart disease in men. *Am J Clin Nutr* 2000;**72**(4):912-21.
10. Fung TT, Rimm EB, Spiegelman D, et al. Association between dietary patterns and plasma biomarkers of obesity and cardiovascular disease risk. *Am J Clin Nutr* 2001;**73**(1):61-7.
11. Varraso R, Garcia-Aymerich J, Monier F, et al. Assessment of dietary patterns in nutritional epidemiology: principal component analysis compared with confirmatory factor analysis. *Am J Clin Nutr* 2012;**96**(5):1079-92.
12. Moskal A, Pisa PT, Ferrari P, et al. Nutrient patterns and their food sources in an International Study Setting: report from the EPIC study. *PLoS One* 2014;**9**(6):e98647.
13. de Vito R, Bellio R, Trippa L, Parmigiani G. Multi-study factor analysis. *arXiv preprint arXiv:1611.06350* 2016.
14. Conway DI, Hashibe M, Boffetta P, et al. Enhancing epidemiologic research on head and neck cancer: INHANCE - The international head and neck cancer epidemiology consortium. *Oral Oncol* 2009;**45**(9):743-6.
15. Hashibe M, Brennan P, Benhamou S, et al. Alcohol drinking in never users of tobacco, cigarette smoking in never drinkers, and the risk of head and neck cancer: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium. *J Natl Cancer Inst* 2007;**99**(10):777-89.
16. Chuang SC, Jenab M, Heck JE, et al. Diet and the risk of head and neck cancer: a pooled analysis in the INHANCE consortium. *Cancer Causes Control* 2012;**23**(1):69-88.
17. Edefonti V, Hashibe M, Parpinel M, et al. Natural vitamin C intake and the risk of head and neck cancer: A pooled analysis in the International Head and Neck

Cancer Epidemiology Consortium. *Int J Cancer* 2015;**137**(2):448-62.
18. Edefonti V, Hashibe M, Ambrogi F, et al. Nutrient-based dietary patterns and the risk of head and neck cancer: a pooled analysis in the International Head and Neck Cancer Epidemiology consortium. *Ann Oncol* 2012;**23**(7):1869-80.
19. Bravi F, Bosetti C, Filomeno M, et al. Foods, nutrients and the risk of oral and pharyngeal cancer. *Br J Cancer* 2013;**109**(11):2904-10.
20. Schantz SP, Zhang ZF, Spitz MS, Sun M, Hsu TC. Genetic susceptibility to head and neck cancer: interaction between nutrition and mutagen sensitivity. *Laryngoscope* 1997;**107**(6):765-81.
21. Levi F, Pasche C, La Vecchia C, Lucchini F, Franceschi S, Monnier P. Food groups and risk of oral and pharyngeal cancer. *Int J Cancer* 1998;**77**(5):705-9.
22. Bosetti C, Gallus S, Trichopoulou A, et al. Influence of the Mediterranean diet on the risk of cancers of the upper aerodigestive tract. *Cancer Epidemiol Biomarkers Prev* 2003;**12**(10):1091-4.
23. Peters ES, McClean MD, Liu M, Eisen EA, Mueller N, Kelsey KT. The ADH1C polymorphism modifies the risk of squamous cell carcinoma of the head and neck associated with alcohol and tobacco use. *Cancer Epidemiol Biomarkers Prev* 2005;**14**(2):476-82.
24. Cui Y, Morgenstern H, Greenland S, et al. Polymorphism of Xeroderma Pigmentosum group G and the risk of lung cancer and squamous cell carcinomas of the oropharynx, larynx and esophagus. *Int J Cancer* 2006;**118**(3):714-20.
25. Divaris K, Olshan AF, Smith J, et al. Oral health and risk for head and neck squamous cell carcinoma: the Carolina Head and Neck Cancer Study. *Cancer Causes Control* 2010;**21**(4):567-75.
26. US Department of Agriculture (USDA), Agricultural Research Service. USDA National Nutrient Database for Standard Reference, Release 26 and previous versions. Nutrient Data Laboratory Home Page: http://www.ars.usda.gov/Services/docs.htm?docid=8964.
27. Gnagnarella P, Salvini S, Parpinel M. Food Composition Database for Epidemiological Studies in Italy. Version 1.2015. Available online at: http://www.bda-ieo.it/ (16 April 2018, date last accessed).
28. Pett MA, Lackey NR, Sullivan JJ. *Making sense of factor analysis: the use of factor analysis for instrument development in health care research.* Thousand Oaks, CA: Sage, 2003.
29. Meng X-L, Rubin DB. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* 1993;**80**(2):267-278.
30. Johnson RA, Wichern DW. *Applied multivariate statistical analysis*. 5th ed. Upper Saddler River, NJ: Prentice Hall, 2002.
31. Akaike H. A new look at the statistical model identification. *IEEE transactions on automatic control* 1974;**19**(6):716-723.
32. DiStefano C, Zhu M, Mindrila D. Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation* 2009;**14**(20):1-11.
33. Hosmer DW, Lemeshow S. *Applied logistic regression, 2nd edn*. New York, NY: John Wiley & Sons, Inc, 2000.
34. Pinheiro JC, Bates DM. *Mixed-effects models in S and S-PLUS*. New York, NY: Springer-Verlag, 2000.
35. R Development Core Team. R: A Language and Environment for Statistical Computing 2017 Vienna, Austria: R Foundation for Statistical Computing, Available at: http:// www.R-project.org (16 April 2018, date last accessed)
36. Giner G, Smyth GK. Statmod: probability calculations for the Inverse Gaussian distribution. *arXiv preprint arXiv:1603.06687* 2016.

37. Revelle W. Psych: procedures for personality and psychological research. Version 1.7.3. 2017, Available at: https://CRAN.R-project.org/package=psych (16 April 2018, date last accessed), Evanston, IL: Northwestern University.
38. Raîche G, Walls TA, Magis D, Riopel M, Blais J-G. Non-graphical solutions for Cattell's scree test. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 2013;**9**(1):23.
39. Wickham H. *Ggplot2: elegant graphics for data analysis* Springer, 2016.
40. Bates D, Maechler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*. Vol. 67(1), 2015;1-48.
41. Block G, Dresser CM, Hartman AM, Carroll MD. Nutrient sources in the American diet: quantitative data from the NHANES II survey. I. Vitamins and minerals. *Am J Epidemiol* 1985;**122**(1):13-26.

Table 1. Distribution of cases of oral and pharyngeal cancers combined, laryngeal cancer, and controls according to selected variables. International Head and Neck Cancer Epidemiology (INHANCE) Consortium.

	Controls (6824)	(%)	Oral and pharyngeal cancer cases (2553)	(%)	Laryngeal cancer cases (1291)	(%)
Age (years)						
<40	411	6.0	113	4.4	24	1.9
>=40to<=44	381	5.6	130	5.1	42	3.3
>=45to<=49	623	9.1	328	13	107	8.3
>=50to<=54	1,091	16	456	18	159	12
>=55to<=59	1,207	18	520	20	237	18
>=60to<=64	981	14	361	14	238	18
>=65to<=69	986	14	337	13	227	18
>=70to<=74	848	12	210	8.2	183	14
>=75	294	4.3	98	3.8	74	5.7
Missing	2	0.0	0	0.0	0	0.0
Sex						
Female	2,166	32	607	24	193	15
Male	4,653	68	1,942	76	1,097	85
Missing	5	0.1	4	0.2	1	0.1
Race						
Black	312	4.6	203	8.0	116	9.0
Others (with Asians)	87	1.3	56	2.2	15	1.2
White (with Hispanics)	6,397	94	2,285	90	1,156	90
Missing	28	0.4	9	0.4	4	0.3
Study center						
Boston	611	9.0	313	12	71	5.5
Italy Multicenter						
Milan	621	9.1	169	6.6	24	1.9
Pordenone - Latina	1,953	29	566	22	409	32
Los Angeles	828	12	246	9.6	60	4.7
Milan (2006-2009)	691	10	131	5.1	200	15
MSKCC	123	1.8	74	2.9	32	2.5
North Carolina (2002-2006)	1,120	16	687	27	374	29
Switzerland	877	13	367	14	121	9.4
Education						
<= Junior high school	2,604	38	821	32	571	44
Some high school	694	10	399	16	209	16
High school graduate	887	13	456	18	191	15
Technical school, some college	1,235	18	447	28	195	15

>= College graduate	1,404	21	430	17	125	9.7
Pack-years						
Never smoker	2,866	42	445	17	79	6.1
1-10	1,074	16	215	8.4	60	4.6
11-20	796	12	216	8.5	130	10
21-30	629	9.2	339	13	165	13
31-40	532	7.8	367	14	217	17
41-50	322	4.7	298	12	210	16
>50	525	7.7	636	25	414	32
Missing	80	1.2	37	1.4	16	1.2
Cigar smoking status						
Never cigar user	6,436	94	2,336	92	1,174	91
Ever smoked>=100 cigars in a lifetime	367	5.4	203	8.0	110	8.5
Missing	21	0.3	14	0.5	7	0.5
Pipe smoking status						
Never pipe user	6,350	93	2,342	92	1,189	92
Ever smoked>=100 pipes in a lifetime	449	6.6	201	7.9	91	7.1
Missing	25	0.4	10	0.4	11	0.9
Alcohol drinking intensity (number of drinks/day)						
Never drinker	1,664	24	261	10	112	8.6
<1	2,011	30	454	18	199	15
>=1to3	1,772	26	490	19	281	22
>=3to5	766	11	369	15	206	16
>=5	611	9.0	979	38	493	38

ABBREVIATIONS: MSKCC: Memorial Sloan Kettering Cancer Center.

Table 2. Shared factor-loading matrix and explained variances (VAR)[a] for the three shared dietary patterns identified by controls-only multi-study factor analysis. International Head and Neck Cancer Epidemiology (INHANCE) Consortium.

Nutrient	*Animal products and cereals*	*Anti-oxidant vitamins and fiber*	*Fats*
Total protein[b,c] (g)	**0.85**	0.22	0.50
Cholesterol[b,c] (mg)	**0.70**	-	0.56
Saturated fatty acids (g)	0.57	0.11	**0.71**
Monounsaturated fatty acids (g)	0.41	0.20	**0.83**
Polyunsaturated fatty acids (g)	0.34	0.26	**0.74**
Total carbohydrates (g)	**0.61**	0.47	0.28
Calcium (mg)	**0.70**	0.33	0.17
Sodium (mg)	**0.72**	0.26	0.47
Potassium (mg)	**0.62**	**0.67**	0.23
Phosphorus (mg)	**0.83**	0.32	0.36
Iron (mg)	**0.65**	0.52	0.30
Zinc (mg)	**0.84**	0.26	0.47
Thiamin (vitamin B1, mg)	**0.71**	0.54	0.30
Riboflavin (vitamin B2, mg)	**0.78**	0.43	0.23
Vitamin B6 (mg)	**0.68**	**0.60**	0.24
Vitamin C (mg)	0.18	**0.79**	-
Total folate (μg)	0.54	**0.70**	0.19
Niacin (vitamin B3, mg)	**0.72**	0.42	0.35
Lutein (μg)	0.16	0.53	0.26
Total carotene (μg)	0.12	**0.66**	0.16
Lycopene (μg)	0.21	0.32	0.25
Vitamin E (mg)	0.19	0.57	**0.67**
Total fiber (g)	0.39	**0.77**	0.16
Proportion of VAR explained (%)	35	23	17
Cumulative VAR explained (%)	35	58	75

[a]Estimated from a multi-study factor analysis carried out on twenty-three nutrients. The magnitude of each loading measures the importance of the corresponding nutrient to the factor. [b]Loadings ≥0.60 define the dominant nutrients for each factor and were shown in bold typeface; loadings <0.1 were suppressed. [d]The units of the nutrients indicated their original scale, but loadings were derived from log-transformed and standardized nutrient intakes entered into the multi-study factor analysis model.

Table 3. Odds Ratios (ORs)[a] of oral cavity and pharyngeal cancers combined, and laryngeal cancer and corresponding Confidence Intervals (95% CIs) on shared factor scores quintile categories, as derived from the controls-only multi-study factor analysis. International Head and Neck Cancer Epidemiology (INHANCE) Consortium.

Shared dietary pattern	Controls	Oral and pharyngeal cases	$p_{studies}$[b]	OR	95% CI[c]	Laryngeal cases	$p_{studies}$[b]	OR	95% CI[c]
Animal products and cereals									
I Quintile (-6.35, -0.89]	1,337	371		1[d]	-	137		1[d]	-
II Quintile (-0.89, -0.32]	1,338	374		0.77	0.64 – 0.92	174		0.91	0.69 - 1.19
III Quintile (-0.32, 0.15]	1,337	489	0.43	0.96	0.81 - 1.15	218	<0.001	1.06	0.82 - 1.38
IV Quintile (0.15, 0.72]	1,339	567		1.07	0.90 - 1.27	317		1.31	0.93 - 1.84
V Quintile (0.72, 4.70]	1,343	687		1.06	0.89 - 1.25	411		1.51	1.06 - 2.13
$p_{for\ trend}$[e]					0.020				<0.001
Anti-oxidant vitamins and fiber									
I Quintile (-5.18, -0.74]	1,337	663		1[d]	-	350		1[d]	-
II Quintile (-0.74, -0.15]	1,340	506		0.79	0.68 - 0.93	272		0.75	0.61 - 0.93
III Quintile (-0.15, 0.35]	1,341	514	<0.001	0.83	0.71 -0.98	250	<0.001	0.67	0.53 - 0.84
IV Quintile (0.35, 0.90]	1,343	427		0.68	0.54 - 0.85	202		0.60	0.46 - 0.78
V Quintile (0.90, 4.78]	1,333	378		0.57	0.41 - 0.78	183		0.62	0.37 – 1.04
$p_{for\ trend}$[e]					0.003				0.049
Fats									
I Quintile (-8.08, -0.80]	1,337	490		1[d]	-	157		1[d]	-
II Quintile (-0.80, -0.22]	1,342	481		0.91	0.77 - 1.07	217		1.28	0.99 - 1.65
III Quintile (-0.22, 0.23]	1,342	484	0.15	0.98	0.82 - 1.16	249	0.06	1.70	1.32 - 2.19
IV Quintile (0.23, 0.75]	1,332	514		0.91	0.77 - 1.07	283		1.69	1.32 - 2.16
V Quintile (0.75, 4.33]	1,341	519		0.80	0.67 - 0.95	351		1.83	1.44 - 2.33
$p_{for\ trend}$[e]					0.019				<0.001

[a]Estimated from multiple logistic regression models adjusted for age, sex, race, study center, education, pack-years of cigarette smoking, cigar smoking status, pipe smoking status, alcohol drinking intensity (number of drinks per day). Results refer to the composite models including all the three shared factors simultaneously. [b]P for heterogeneity between studies. [c]For both cancer sites, we reported results from a generalized linear mixed model including a random-slope for the *Anti-oxidant vitamins and fiber* pattern with oral and pharyngeal cancers combined and a random slope for the *Animal products and cereals* and the *Anti-oxidant vitamins and fiber* patterns with laryngeal cancer. [d]Reference category. [e]P for linear trend.