Shared and Study-Specific Dietary Patterns: a Novel Approach to Replicability and Validity

R. De Vito[1], C. La Vecchia [2], G. Parmigiani [3,4], V. Edefonti [2*]

[1] Department of Computer Science, Princeton University, Princeton, NJ, USA;
[2] Branch of Medical Statistics, Biometry and Epidemiology ”G. A. Maccacaro”, Department of Clinical Sciences and Community Health, Università degli Studi di Milano, Milano, Italy;
[3] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA;
[4] Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA.

[*] The author list included all the INHANCE consortium investigators that participated in this project

Abstract
We adopt the “multi-study factor analysis” approach in order to identify dietary patterns (DPs) across different studies. The goal is 1) to distinctly estimate shared and study-specific DPs across multiple populations, and 2) to validate the DPs by examining their association with cancers of the oral cavity and pharynx (OCP) and larynx.
To assess these issues we use individual-level pooled data from 7 case-control studies (3,844 cases and 6,824 controls from Europe and the United States) participating in the International Head and Neck Cancer Epidemiology consortium.
We identified 3 DPs that were shared among all studies: the *Antioxidant vitamins and fiber* DP was inversely associated with risk of OCP and laryngeal cancers, the *Fats* DP was inversely associated with risk of OCP cancer, whereas the *Animal products and cereals* and the *Fats* DPs were positively associated with risk of laryngeal cancer. Each study from the US expressed an additional study-specific DP.
We provided a valuable tool to fully understand multi-study replicability in a DP analysis and insight into DP validity.

Key Words: dietary patterns, multi-study factor analysis, replicability, validity analysis, laryngeal cancer, oral cavity and pharyngeal cancer

1. Introduction

Dietary patterns (DPs) are an effective tool for assessing overall diet in individuals. One of the main contribution of DPs is focused on its association with health, and on its effects on disease risk (Hu, 2002).
In order to estimate DPs, research in nutritional epidemiology has been focused on statistical methods for *a posteriori* DPs, although most of the literature has adopted standard multivariate statistical methods, such as principal component analysis (PCA) (Varraso et al., 2012, Moskal et al., 2014, Castelló et al., 2016), and factor analysis (FA) (Gittelsohn et al., 1998, Togo et al., 2004, Judd et al., 2015).

However, an analysis that integrates different population and studies has not been extensively explored in this field. Considering multiple studies can leverage the power of sample size and can capture replicable signal and differences among diverse populations and, consequently, dietary habits. Indeed, when considering multiple studies, a crucial challenge is learning replicable features shared among studies while simultaneously identifying the component specific to each study. To our knowledge, few studies have focused on replicability of DPs. For example, some studies (Castelló et al., 2016, Castelló et al, 2016) compared similar DPs using correlation and congruence coefficients between DP loadings. Edefonti et al. (2012) analyzed five different populations shared exactly the same variables (e.g. nutrients) to be then merged in one single population. Then, they applied FA to this population to determine shared dietary patterns and their relation with head and neck cancer (HNC) risk.

These examples highlight the urgent need in this setting for a model able to handle multiple populations and to derive in a single analysis (1) factors that capture shared dietary patterns common to all available studies, and (2) study-specific factors.

In our paper (De Vito et al., 2018), we adopted the multi-study factor analysis (MSFA), developed by De Vito et al., 2016, in the International Head and Neck Cancer Epidemiology (INHANCE) consortium (Hashibe et al, 2007, Conway et al., 2009).

Multi-study factor analysis is a dimension-reduction approach that allows for the joint analysis of multiple studies. Specifically, MSFA is a generalization of FA able to handle multiple studies simultaneously. Indeed, MSFA estimates shared DPs common to all the studies, and identifies the study-specific DPs related to the specificity of each population. This method is based on the maximum likelihood estimation approach computed via an Expectation Conditional-Maximization (ECM) algorithm (Meng et al, 1993).

In the setting of the INHANCE consortium (Conway et al, 2009), MSFA tackles the issue of replicability of DPs. The INHANCE consortium was established in 2004 to elucidate the etiology of HNC through pooled analyses of individual-level data from several studies. In this way we are able to assess replicability of DPs in different studies presented in INHANCE and to validate the shared and study-specific DPs identified by MSFA by associating them with the risk of HNC subsites, including oral cavity and pharynx (OCP) and larynx.

Section 2 of this report introduces the data and describes MSFA, the estimation of model parameters based on maximum likelihood, implemented via the ECM algorithm. Section 3 presents the application to study DP replicability and validity. Section 4 contains the final discussion.

2. Materials and Methods

2.1 Design and Subjects

We extracted from version 1.5 of the INHANCE pooled dataset seven case-control studies (Bravi et al., 2013, Schantz et al., 1997, Levi et al., 1998, Bosetti et al., 2003, Peters et al., 2005, Cui et al., 2006, Hashibe et al., 2006, Divaris et al., 2010) that provided a sufficiently large list of nutrients. Three of the studies analyzed here were conducted in Europe and four in the United States (the case-control ratio is from 0.41 to 0.98 across studies). Other details on the individual studies, harmonization of data and data pooling methods have been previously described (Conway et al., 2009) and summarized (De Vito et al., 2018). Relevant institutional review boards approved the investigations, according to the specific rules applied to each country at the data collection time.

2.2 Model

We adopted MSFA to identify shared and study-specific DPs for the overall set of HNC cases and controls. Specifically, we considered S=7 studies, each represented by the same set of P=23 nutrients. Study s has n_s subjects, each represented by a P-dimensional log-transformed and standardized data vector, $x_{is,}$ with i=1,…,n_s, s=1,…,S. The x_{is} were expressed by MSFA in terms of K shared factors and J_s additional study-specific factors, giving a total T_s=K+J_s factors. Let f_{is} be the (K × 1) shared latent factor vector for subject i in study s, and Φ be the (P × K) shared factor loading matrix. Moreover, let l_{is} be the (J_s × 1) study-specific latent factor vector and Λ_s be the (P × J_s) specific factor loading matrix. Multi-study factor analysis assumes that the P-dimensional vector x_{is} is decomposed as:

$$x_{is} = \Phi\, f_{is} + \Lambda_s\, l_{is} + e_{is} \qquad i = 1, \dots, n_s \quad s = 1, \dots, S$$

where the (P × 1) error term e_{is} has a multivariate normal distribution with mean vector 0 and diagonal covariance matrix $\Psi_s = diag\,(\psi_{s1}, \dots, \psi_{sp})$.

As a result of the model assumption, the marginal distribution of x_{is} is multivariate normal with mean vector 0 and covariance matrix $\Sigma_s = \Phi\Phi^T + \Lambda_s\Lambda_s^T + \Psi_s$. The covariance matrix is decomposed in three different components, namely the shared factor's variance, the study- specific factor's variance and the variance of the error term.

2.2.1 Parameter Estimation

The parameters to be estimated within the MSFA approach are given by $\theta = (\Phi, \Lambda_s, \Psi_s)$. Let's assume that the observed variables x_{is} have been centered. Then the log-likelihood for the MSFA method is

$$l(\theta) = \sum_{s=1}^{S} -\frac{n_s}{2}\log|\Sigma_s| - \frac{n_s}{2} tr\,(\Sigma_s^{-1}\, C_{x_s x_s})$$

where $C_{x_s x_s}$ is the covariance matrix computed in each study.

The parameters Φ, Λ_s, and Ψ_s are estimated by a generalized version of the Expectation Maximization (EM) algorithm (Dempster et al., 1977). We use the ECM (Meng et al., 1993) to estimate the parameter vector by replacing the standard maximization step in the EM with a set of conditional maximization steps.

For a better interpretation of the loadings, the varimax rotation is then applied to the estimated factor loading matrices.

2.2.2 Dimension procedure (model selection)

To select the dimension of the model, we proceeded in two steps. Firstly, we determined the total number of factors, T_s, s=1,…, S, by adopting a combination of standard techniques for FA, such as Horn's parallel analysis, Cattell's scree plot, or the use of indexes, namely the root mean square error of approximation (Mulaik, 2009). Next, model selection techniques, such as Akaike Information Criterion (AIC) (Akaike, 1974), were applied to the MSFA model to select the number of shared factors, K. The number of study-specific factors, J_s s=1,…, S, is then obtained by difference as T_s - K, s=1,…,S.

Moreover, we determined a global AIC as a final criterion to identify the optimal pair (K, J_s).

2.2.3 Factor Scores

In order to estimate the degree of each subject's diet adherence to the previously identified DPs, we computed the factor scores from the MSFA model. Factor scores were computed within MSFA (De Vito et al., 2018) by adopting both the standard available methods for FA (Bartlett and Thurstone) (Johnson et al., 2002, DiStefano et al., 2009). Specifically, we determined a factor score for each subject and factor within each study by using the correlation matrix of each study, $C_{x_s x_s}$, and the overall factor-loading matrix $[\Phi, \Lambda_s]$. Since the correlation between the factors estimated by the two methods, Bartlett and Thrustone, was relevantly high (0.99) (De Vito et al., 2018), we proceeded with the Bartlett method, since it assumes unbiaseness and uncorrelation between factor scores (Johnson et al., 2002, DiStefano et al., 2009).

2.3 Association of the identified dietary patterns and head and neck cancer

Participants were grouped in five (quantiles for the shared factor) or three categories (tertiles for the study-specific factors) (De Vito et al., 2018).

The odds ratios (ORs) and the corresponding 95% confidence intervals (CIs) of OCP and laryngeal cancers were computed, separately, for each category using multiple logistic regression models (Hosmer et al., 2000). Separate models were fitted for each factor. In validation, we fitted a shared-factor regression model and a study-specific regression model including all the shared and one study-specific factor at a time. Each model 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 De Vito et al., 2018 for the covariate categories adopted).

The resulting ORs and CIs were computed adopting a random-slope generalized linear mixed model with logit link function and binomial family (Pinheiro et al., 2011), in order to include the heterogeneity of the shared DPs' associations across studies. The random-effects had the quintile categories effects (except for the reference) as random slopes and study center as common grouping factor, in total eight levels. We did not include in the model random intercepts or random effects correlations. Specifically, four random-effects terms (one for each quintile category, reference category excluded) could be considered in the model for each shared pattern, with a total of thirty-two random effects (one for each study center and quintile category, reference category excluded). Moreover, we considered only random-effects terms for patterns showing a good fit in favor of them.

3. Analysis and Results

In this section, we analyzed the seven studies in order to estimate the shared and the study-specific DPs and their association with risk of cancer of OCP and larynx.

Initially we conducted preliminary analysis to inspect the factorability of the data adopting techniques such as Bartlett's test of sphericity, the Kaiser-Meyer-Olkin measure, and individual measures of sampling adequacy (Pett et al., 2003).

Since all these measures produced reasonable values, we then proceeded with DP estimation. We firstly assessed the total factors dimensions, the number of the shared factors and the number of the study-specific factors. Comparing AICs from different models, we set the number of shared factors to three and the number of study-specific factors to one for the American studies [Los Angeles, Boston, Memorial Sloan Kettering Cancer Center, and North Carolina (2002-2006)].

Next, we focused on the DPs. The three shared factors explained 75% of the total variance. The first factor, *Animal products and cereals*, had highest loadings on phosphorus, riboflavin, zinc, total protein, calcium, niacin, thiamin, vitamin B6, sodium, potassium, iron, cholesterol, and total carbohydrates. The second factor, *Anti-oxidant vitamins and fiber,* had the highest loadings on vitamin C, total fiber, total carotene, total folate, vitamin E, and potassium. The third factor, *Fats,* had the highest loadings on monounsaturated, polyunsaturated, and saturated fatty acids, and vitamin E.

The study-specific factors explained 5% (Los Angeles), 3% (Boston), 6% (Memorial Sloan Kettering Cancer Center), and 3% (North Carolina (2002-2006)) of the total variance. These four factors, all named *Dairy products and breakfast cereals*, presented a similar pattern. Specifically, all of these four factors showed a high positive loading on calcium and a high negative one (in absolute value) on niacin.

Further, we compared MSFA and FA via bootstrap analysis. In particular, we analyzed the distributions and standard errors of the factor loadings computed from 100 bootstrapped random sets of the seven studies under the two approaches, i.e. MSFA and FA, after merging all the data-sets in one. Figure 1 depicts the bootstrap results: the boxplots produced by the MSFA are always less broad than the corresponding ones from FA. Additionally, the standard errors of the shared-factor loadings are always smaller under the MSFA approach.

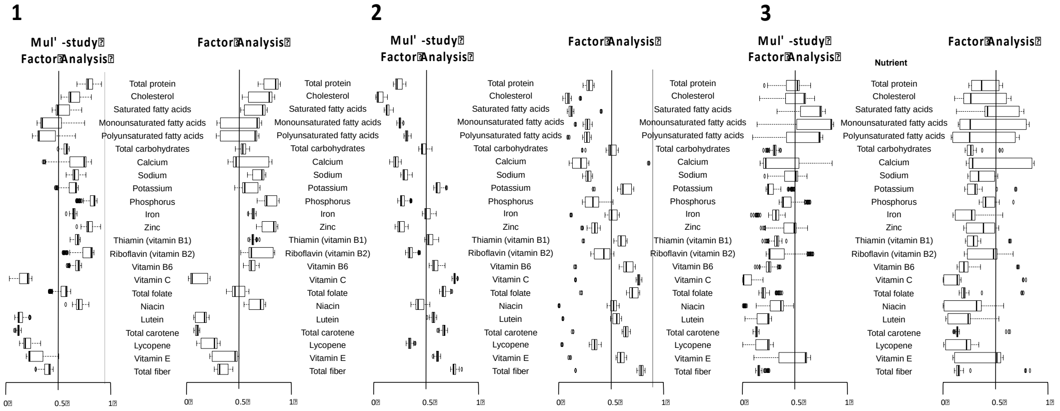

Figure 1: Distribution of 100 bootstraps of the sample of the estimated factor loadings for the three shared dietary patterns identified by both the multi-study factor analysis and factor analysis. The distribution of each loading is depicted in each boxplot.

The bootstrap analysis illustrates how MSFA borrows strength across studies in the estimation of the factor loadings, in such a way that the reliability in independent observations is not only preserved but even improved.

Finally, the ORs and the CIs for cancers of the OCP and larynx were computed. For the shared factors, as reported in detail in De Vito et al., 2018, the *Animal products and cereals* pattern was positively associated with laryngeal cancer risk, the *Anti-oxidant vitamins and fiber* pattern was inversely related to the risk of cancer at both OCP and larynx, and, finally, the *Fats* pattern was inversely associated with cancer of the OCP and positively associated with laryngeal cancer. For the study-specific factors, the Los Angeles specific DP was inversely associated with OCP cancer risk, however the pattern identified in the Boston study was positively associated with the same cancer site.

4. Discussion

Multi-study factor analysis provide a valuable tool for the analysis of reproducibility and validity of DPs in a nutritional epidemiology framework. The main idea is to estimate the shared factors across studies and to identify study-specific factors related to each population. De Vito et al., 2018 identifies four American study-specific patterns and no additional study-specific patterns in any European population. The reason can be found in differences between Europe and United States in terms of foods and breakfast habits. Indeed, the foods rich in niacin and low in calcium (or vice versa), as identified in our *study-specific* DPs analysis, includes: milk, cheese, yoghurt, instant and filter coffee, as well as cereal products in general, with most breakfast cereals. Secondly, there are some differences in the FFQ questionnaire: the European studies have the same FFQ, which differs from those of the American studies.

Differences in terms of risk for Los Angeles-specific and Boston--specific DPs are reported. The diverse direction of the risk for the *Dairy products and breakfast cereals* DP can be related to different food sources of the two FFQs.

Multi-study factor analysis can be applied to many settings, such as any consortium or network of consortia, including those of cohort studies where the aim is to identify shared and study-specific factors. The selection of the factor-loading matrix dimension via objective criteria, like AIC, allows us to check the model dimension with a valuable measure and tool. Moreover, MSFA is consolidated with standard checks of internal stability and internal consistency of the estimated DPs (Edefonti et al., 2010).

In conclusion, the application of MSFA in nutritional epidemiology allowed to identify relevant eating patterns across the INHANCE consortium populations from Europe and the US associated with risk of OCP and laryngeal cancers. This approach can be the basis for integrating information from different population in a DP framework. Moreover, these results may be relevant information for the next releases of national dietary guidelines.

Acknowledgements

We would like to thank the investigators of the INHANCE consortium: **Y. C. A. Lee** from the Division of Public Health, Department of Family & Preventive Medicine, University of Utah School of Medicine; **M. Parpinel** from the Department of Medicine, University of Udine; **D. Serraino** from the Epidemiology and Biostatistics Unit, CRO Aviano National Cancer Institute, IRCCS; **A. F. Olshan** from the University of North Carolina School of Public Health, Chapel Hill; **J. P. Zevallos** from the Department of Otolaryngology/Head and Neck Surgery, University of North Carolina at Chapel Hill; **F. Levi** from the Institute of Social and Preventive Medicine (IUMSP), Lausanne University Hospital (CHUV); **Z. F. Zhang** from the Department of Epidemiology, UCLA School of Public Health; **H. Morgenstern** from the Departments of Epidemiology and Environmental Health Sciences, School of Public Health and Comprehensive Cancer Center, University of Michigan; **W. Garavello** from the Department of Otorhinolaryngology, School of Medicine and Surgery, University of Milano – Bicocca; **K. Kelsey** from the Department of Epidemiology and Pathology and Laboratory Medicine, Brown University; **M. McClean** from the Department of Environmental Health, Boston University School of Public Health; **S. Schantz** from the Department of Otolaryngology, New York Eye and Ear Infirmary; **G. P. Yu** from the Medical Informatics Center, Peking University; **P. Boffetta** from the Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai; **S. C. Chuang** from the Institute of Population Health Sciences, National Health Research Institutes; and **M. Hashibe** from the Division of Public Health, Department of Family & Preventive Medicine and Huntsman Cancer Institute.

References

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

Varraso, R., et al. Assessment of dietary patterns in nutritional epidemiology: principal component analysis compared with confirmatory factor analysis. *The American journal of clinical nutrition* 96.5 (2012): 1079-1092.

Moskal, A., et al. Nutrient patterns and their food sources in an International Study Setting: report from the EPIC study. *PLoS One* 9.6 (2014): e98647.

Castelló, A., et al. Evaluating the applicability of data-driven dietary patterns to independent samples with a focus on measurement tools for pattern similarity. *Journal of the Academy of Nutrition and Dietetics* 116.12 (2016): 1914-1924.

Gittelsohn, J., Wolever, T. M., Harris, S. B., Harris-Giraldo, R., Hanley, A. J., & Zinman, B. (1998). Specific patterns of food consumption and preparation are associated with diabetes and obesity in a Native Canadian community. *The Journal of nutrition*, *128*(3), 541-547.

Togo, P., Osler, M., Sørensen, T. I. A., & Heitmann, B. L. (2004). A longitudinal study of food intake patterns and obesity in adult Danish men and women. *International journal of obesity*, *28*(4), 583.

Judd, S. E., Letter, A. J., Shikany, J. M., Roth, D. L., & Newby, P. K. (2015). Dietary patterns derived using exploratory and confirmatory factor analysis are stable and generalizable across race, region, and gender subgroups in the REGARDS study. *Frontiers in nutrition*, *1*, 29.

Castelló, A., et al. Reproducibility of data-driven dietary patterns in two groups of adult Spanish women from different studies. *British Journal of Nutrition* 116.4 (2016): 734-742.

Edefonti, V., 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. *Annals of Oncology* 23.7 (2011): 1869-1880.

De Vito, R., et al. Shared and study-specific dietary patterns and head and neck cancer risk in an international consortium. *Epidemiology* (2018) Epub ahead of print.

De Vito, R., Bellio, R., Trippa, L., & Parmigiani, G. (2016). Multi-study factor analysis. *arXiv preprint arXiv:1611.06350*.

Hashibe, M., 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. *Journal of the National Cancer Institute* 99.10 (2007): 777-789.

Meng, X. L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, *80*(2), 267-278.

Conway, D. I., et al. Enhancing epidemiologic research on head and neck cancer: INHANCE–The international head and neck cancer epidemiology consortium. *Oral oncology* 45.9 (2009): 743-746.

Bravi, F., et al. Foods, nutrients and the risk of oral and pharyngeal cancer. *British journal of cancer* 109.11 (2013): 2904.

Schantz, S. P., Zhang, Z. F., Spitz, M. S., Sun, M., & Hsu, T. C. (1997). Genetic susceptibility to head and neck cancer: interaction between nutrition and mutagen sensitivity. *The laryngoscope*, *107*(6), 765-781.

Levi, F., Pasche, C., La Vecchia, C., Lucchini, F., Franceschi, S., & Monnier, P. (1998). Food groups and risk of oral and pharyngeal cancer. *International journal of cancer*, *77*(5), 705-709.

Bosetti, C., et al. Influence of the Mediterranean diet on the risk of cancers of the upper aerodigestive tract. *Cancer Epidemiology and Prevention Biomarkers* 12.10 (2003): 1091-1094.

Peters, E. S., McClean, M. D., Liu, M., Eisen, E. A., Mueller, N., & Kelsey, K. T. (2005). The ADH1C polymorphism modifies the risk of squamous cell carcinoma of the head and neck associated with alcohol and tobacco use. *Cancer Epidemiology and Prevention Biomarkers*, *14*(2), 476-482.

Cui, Y., et al. Polymorphism of Xeroderma Pigmentosum group G and the risk of lung cancer and squamous cell carcinomas of the oropharynx, larynx and esophagus. *International journal of cancer* 118.3 (2006): 714-720.

Hashibe, M., et al. Marijuana use and the risk of lung and upper aerodigestive tract cancers: results of a population-based case-control study. *Cancer Epidemiology and Prevention Biomarkers* 15.10 (2006): 1829-1834.

Divaris, K., et al. Oral health and risk for head and neck squamous cell carcinoma: the Carolina Head and Neck Cancer Study. *Cancer Causes & Control* 21.4 (2010): 567-575.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38.

Mulaik, S. A. (2009). *Foundations of factor analysis*. Chapman and Hall/CRC.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716-723.

Härdle, W., & Simar, L. (2007). *Applied multivariate statistical analysis* (Vol. 22007, pp. 1051-8215). Berlin: Springer.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*(20), 1-11.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Pinheiro, J. C., & Bates, D. M. (2011). Mixed-effects Models in S and S-PLUS.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Sage.

Edefonti, V., et al. Nutrient-based dietary patterns and laryngeal cancer: evidence from an exploratory factor analysis. *Cancer Epidemiology and Prevention Biomarkers*19.1 (2010): 18-27.