Estimation of multiple correlated effects on a disease outcome when the data are nested: a hierarchical Bayesian approach

Giulia Roli

Department of statistical sciences, University of Bologna, Italy.

Summary. The paper deals with the analysis of multiple exposures on the occurrence of a disease. We consider observational case-control data in a multilevel setting, with subjects nested in clusters. A hierarchical Bayesian model is proposed to tackle the within-cluster dependence and the correlation among the exposures, simultaneously. To do that, we assign prior distributions on the crucial parameters by exploiting additional information at different levels, as well as reasonable assumptions and previous knowledge. The model is applied to a multi-centric study aiming to investigate the association of dietary exposures with the colon-rectum cancer. Compared with results obtained with conventional regressions, our hierarchical Bayesian model yields great gains in terms of more consistent and less biased estimates.

Keywords: Bayesian analysis, case-control study, disease outcome, hierarchical regression, multiple exposures, prior assignment.

1. Introduction and background

When a case-control study aims to investigate the exposures which can be the cause of the occurrence of a disease, epidemiologists often deal with some complications that need to be somehow controlled during the analysis. In such cases, the use of the models conventionally employed becomes improper, yielding apparent associations between some exposures and the disease and unstable corresponding estimates.

We consider two kinds of such complications. The first one concerns the structure of the data and occurs whenever subjects are nested into higher level units involving their own variability and a dependence among the related observations. The commonest examples in epidemiology lie in patients admitted to different hospitals or wards, as well as subjects living in various neighborhoods, towns or countries (Leyland and Goldstein (2001)). More generally, the nested structure of data is a common phenomenon, especially in behavioral and social research, where the evaluation of the relationship between individuals and society is of crucial importance. In all these cases, the dependence of data is a focal interest of the research. Conversely, the hierarchy of data can be generated by the sampling design, such as in the multi-stage sampling, which is frequently employed in the traditional surveys to reduce the costs of data collection. As a result, the dependence is treated as a nuisance which requires further adjustments during the analysis. Whatever the dependence arises from, it is "neither accidental nor ignorable" (Goldstein (1999)). Indeed, the risks of drawing wrong conclusions are high if the clustering of the data is disregarded (Snijders and Bosker (1999)).

E-mail: g.roli@unibo.it

The joint analysis of multiple exposures gives rise to the second complication. Indeed, many epidemiologic studies involve a set of potential effects to be compared and, as a result, face problems of multiple inference (Thomas *et al.* (1985)). When a conventional analysis is carried out, these problems are revealed by failures in the convergence of the estimation process or by implausible large and unstable estimates, especially when the samples are small and sparse (Greenland (1992); Greenland (1993)). The main reason is that these effects are often correlated. Therefore, we need to take into account for a covariance structure among them to reduce the random errors in the estimates.

Both these complications have been tackled separately in various applications and simulations by using hierarchical modeling (see for example, Diex-Roux (2000), Diex-Roux (2004), Greenland (1992), Witte *et al.* (1994)). Over the last 50 years, hierarchical modeling has appeared in various forms to address many multiparameter problems, involving two or more levels of analysis and specifying various relationships among study variables and parameters. In epidemiological research, some noteworthy applications include disease mapping (see, e.g., Bernardinelli *et al.* (1995)), spatial and spatio-temporal analysis (Lawson (2001)), study of health-care programs and institutions (Burgess *et al.* (2000)). Moreover, the large increase in computing power over recent decades has strongly supported the spreading of this approach as a practical and powerful analysis tool (Graham (2008); Raudenbush and Bryk (2002); Greenland (2000)).

When the structure of the data is nested, hierarchical modeling allows to handle simultaneously multiple levels of information and dependencies (Raudenbush and Bryk (2002); Leyland and Goldstein (2001); Snijders and Bosker (1999); Hox (1995)). In this setting, we often refer to *multilevel* regression models. These can appropriately address different research aims: (i) improved estimation of the individual effects under investigation (i.e., all the available information at both levels are efficiently used in order to exploit both the group features and the relations existing in the overall sample); (ii) evaluation of the cross-level effects (e.g., how variables measured at one level affect relations occurring at another); and (iii) decomposition of the variance-covariance components at each level. Although it was firstly introduced and used in educational and social fields, during the past decade the multilevel approach has been increasingly employed also in epidemiologic analysis as a powerful strategy to explain the correlation between analytical units (see for example, Leyland and Goldstein (2001); Diex-Roux (2004); Cubbin and Winkleby (2005)).

As far as the multiple exposure issue is concerned, numerous authors have shown that empirical and semi-Bayes estimates from hierarchical models can improve standard regression estimation, allowing for correlated associations and showing to be less sensitive to sampling error and model misspecification (Morris (1983); Greenland (1992); Greenland (1993); Greenland (1997)). Indeed, relying on the presence of some additional information suitable to mediate the final effects of the exposures, they can be arranged in a second-stage regression to model similarities among the parameters of interest (Witte *et al.* (1994); Rothman *et al.* (2008)).

Although developed separately and for different purposes, hierarchical modeling for correlated effects an for nested data have important communalities, which can be strengthened especially when a Bayesian perspective is adopted. The use of Bayesian methods for epidemiological research is a relevant topic discussed by several authors (Greenland (2006); Greenland (2007); MacLehose *et al.* (2007); Graham (2008)). They all support the use of prior assumptions as they are more reasonable than those implicitly made by frequentist models and address the problems of sparse data, multiple comparisons, subgroup analysis and study bias. The main feature is that prior expectations on the parameters are embedded in a probability model with its own uncertainty to form a hierarchy of models and parameters. As a result, the corresponding posterior estimates are compromises between summaries of the sample data and such prior expectations.

In this framework, the assignment of prior judgements is of primary importance. In general, a reasonable Bayesian analysis needs a prior that reflects results form previous studies or review. A fully-Bayesian (FB) approach forces all the parameters in the model to be random and corresponding probability distributions to be assigned. When these prior distributions are in the form of prior data, we refer to *empirical prior*, arising from frequentist shrinkage-estimation or empirical-Bayes (EB) methods (Maritz and Lwin (1989)). Moreover, the increasing availability of data that can be easily linked each other by computer programs has strongly supported the use of the EB methods. Actually, both the hierarchical models described above for nested data and correlated effects involve the EB approach, as they employ additional information on the crucial parameters of interest arranged in a hierarchy of probability models.

Instead of assigning a full prior distribution, another method consists in fixing in advance a specific value for one or more parameters using background information. This strategy, called semi-Bayes (SB) approach, is commonly employed to avoid the drawback of absurd estimates of some (hyper-) parameters (Greenland (1992); Greenland (2000)).

Such criteria for the assignment of the priors can be jointly adopted to specify the probability distributions of different parameters. Indeed, the Bayes empirical-Bayes (BEB) methods (Deeley and Lindley (1981)) exploit the available prior data for some (hyper-) parameters and some kinds of proper distributions for the others. In the latter case, the specification can involve different levels of knowledge, as well as reasonable assumptions, to develop an informative prior. Otherwise, noninformative distributions can be specified.

In this paper, we aim to extend the hierarchical approach in a multilevel setting for the analysis of multiple exposures and highly correlated effects. We attempt to improve the ordinary estimates of such effects by using some descriptive information to develop a second-stage regression model mediating the effects of the exposure variables, separately by group membership but into a single analysis. Such additional data are second-stage covariates which can arise from specific features of the clusters, as well as information about the regressors. In addition, we adopt a BEB perspective and exploit the previous knowledge on the other (hyper-) parameters to specify prior distributions, which are suitable with respect to the problem at hand. The main purpose is to provide a flexible and powerful framework for the analysis of complex case-control data and to encourage the use of the Bayesian methods in epidemiology. In order to prove and measure the gains in the final estimates of the crucial parameters, we consider a notable application aiming to investigate the association of dietary exposures with the occurrence of colon-rectum cancer. In this study, a multilevel setting is involved, as individuals are enrolled from different countries and centers of Europe. Additional data on the nutrient compositions of each dietary item are arranged to model the correlation among the exposures.

The paper develops as follows. We firstly introduce the conventional analysis employed for the evaluation of multiple exposures on a disease and the corresponding hierarchical approach when the data are nested. In particular, the case of observational case-control studies is considered. Then, the extension to correlated multiple effects of the exposures in the Bayesian setting is described in sections 3 and 4. In section 5, we compare the hierarchical Bayesian regression method with the conventional maximum-likelihood results with respect to the study application. The last section summarizes our findings and concludes.

2. Hierarchical approaches for nested data (*multilevel models*)

We consider a case-control study, where the presence/absence of a disease is denoted by the individual indicator Y ($Y = 1$ for cases, $Y = 0$ for control units) and the information on K exposures are summarized by the matrix $\boldsymbol{X} = [x_{ik}]$. Under the independence assumption of the responses Y_i across the units, a conventional analysis would use the method of Maximum Likelihood (ML) to estimate the effects $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_K]$ of the exposures for $i = 1, ..., n$ individuals in the sample according to the following logistic regression

$$Pr[(y_i = 1|x_i; w_i] = \frac{exp(\alpha + \sum_{k=1}^{K} \beta_k x_{ik} + \sum_{p=1}^{P} \gamma_p w_{ip})}{1 + exp(\alpha + \sum_{k=1}^{K} \beta_k x_{ik} + \sum_{p=1}^{P} \gamma_p w_{ip})}$$

or, analogously,

$$logit[E(y_i|\boldsymbol{x}_i; \boldsymbol{w}_i]) = \alpha + \sum_{k=1}^{K} \beta_k x_{ik} + \sum_{p=1}^{P} \gamma_p w_{ip}. \tag{1}$$

Generally, the effects of a set of potential confounders (such as age and sex of individuals) need to be controlled for (Rothman *et al.* (2008)). In our formulation, they are embedded in the model specification as additional covariates in the matrix $\boldsymbol{W} = [w_{ip}]$ with corresponding vector of coefficients $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_P]$.

When the data structure is hierarchical with subjects at level 1 nested in clusters at level 2, the basic independence assumption across units is violated. If we ignore this within-cluster dependence, the conventional analysis yields incorrect standard errors and inefficient estimates (Diex-Roux (2000)). Therefore, unless some different statistical models are introduced, we should be forced to carry out separate ordinary logistic regressions, one for each cluster. Conversely, a proper method to manage correlations among the responses is represented by the hierarchical or multilevel modeling (Raudenbush and Bryk (2002); Leyland and Goldstein (2001); Snijders and Bosker (1999); Hox (1995)). This allows to unify the analysis across the clusters, partition the variability at both levels and choose the parameters to be random among the groups.

The simplest multilevel regression is represented by the random-intercept model, where among the set of parameters only the intercept varies across the clusters. If we can further suppose the effects of the exposures are random†, the resulting random-slopes model is

$$logit[E(y_{ij}|\boldsymbol{x}_{ij}; \boldsymbol{w}_{ij}]) = \alpha_j + \sum_{k=1}^{K} \beta_{kj} x_{ikj} + \sum_{p=1}^{P} \gamma_p w_{ipj} \tag{2}$$

where the n level-1 subjects in the J level-2 groups (with $i = 1, ..., n_j$, $n = \sum_j n_j$ and $j = 1, ..., J$) are characterized by the observed data, now denoted by the double index ij. The effects of potential confounders are estimated regardless of the nested structure of data. Conversely, the intercepts and the exposure effects are influenced by the data hierarchy and are modeled by a set of level-2 regressions, which can exploit some additional information.

If we can rely on the presence of a number R of level-2 data, $v_{j1}, \ldots, v_{jR}$, characterizing the group units, they can be employed as regressors in order to explain the cluster variability

†The effects of the potential confounders are reasonably assumed to be independent on the group membership.

and to shrink the estimation of the corresponding random coefficients toward each other when different groups have similar information. As an example, in a study of patients nested into hospitals a typical level-2 covariate should be the classification into private or public hospitals or the numbers of doctors working in each hospital. In such cases, the level-2 regressions for the intercepts and slopes can be formulated as

$$\alpha_j = \psi_{00} + \sum_{r=1}^{R} \psi_{0r} v_{jr} + u_{0j} \tag{3}$$

$$\beta_{kj} = \psi_{k0} + \sum_{r=1}^{R} \psi_{kr} v_{jr} + u_{kj} \tag{4}$$

where the vectors $\boldsymbol{\psi}_{k^*} = [\psi_{0k^*}, \psi_{1k^*}, \ldots, \psi_{Kk^*}]$ (with $k^* = 0, 1, \ldots, K$) include the level-2 intercepts and coefficients; and u_{k^*j} are the level-2 residuals, which are assumed to be normally distributed with null means. In the standard formulation, the dispersion of the level-2 random effects u_{k^*j} is represented by a variance-covariance matrix $\mathbf{\Phi}$, where the variance terms are denoted by $\phi^2_{k^*}$ and the covariances between slopes and intercepts by $\phi_{k^*k'}$ (with $k' = 0, 1, \ldots, K$ and $k^* \neq k'$). They are all defined as *conditional* or *residual* components. Conversely, whenever no level-2 predictors are included, we refer to *unconditional* variance-covariance components (Raudenbush and Bryk (2002)).

3. Hierarchical modeling for correlated exposures and nested data

Let consider the non-multilevel case in the conventional logistic regression (1), where it is crucial to estimate the effects of the exposures in vector $\boldsymbol{\beta}$. When a large number of exposures are involved, we often face several drawback due to interactions and collinearity among covariates or data which can be too sparse to yield accurate estimates (Thomas et al., 1985). As a result, a number of these standard ML estimates can show large absolute values, suggesting strong associations often implausible according to the relevant epidemiologic literature. A reasonable approach to address this problem consists in exploiting some information about the regressors $\boldsymbol{X}$, which can mediate their final effects (Morris (1983); Greenland (1992); Greenland (1993); Greenland (1997)). For instance, in a study where the vitamin intakes are related to the occurrence of a disease, we can exploit the distinction between water-soluble and fat-soluble vitamins; or if we aim to investigate the chlorinated hydrocarbon levels in tissues, we would know the degree of chlorination of measured compounds (Rothman *et al.* (2008)). In this setting, the coefficients β_k are supposed to be random parameters and modeled through these prior data in matrix $\boldsymbol{Z} = [z_{qk}]$ as follows

$$\beta_k = \pi_0 + \sum_{q=1}^{Q} \pi_q z_{qk} + \varepsilon_k \tag{5}$$

where the column vector $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_Q]$ includes the effects of such prior information on the exposures (and on the disease) common to all the exposures and ε_k are independent and normal distributed residuals with null mean and constant variance. This additional regression is shown to improve the final estimate of effects β_k by pulling the ML estimates from the conventional logistic model toward each other when corresponding exposures are alike in terms of prior information.

When the structure of the data is hierarchical and we aim to evaluate the effects of exposures with respect to the different clusters according to regression (2), the inference problems due to the joint analysis of multiple exposures can be strengthened by the sparseness of information in small groups. The level-2 regressions for slopes in the multilevel setting (4) can partially model the associations among the exposures thanks to the residual covariance terms included in matrix $\mathbf{\Phi}$. However, a more appropriate approach should include again some information about the regressors $\boldsymbol{X}$ in order to explain (part of) the covariance components.

With this aim, we can generalize the standard random-slopes model (4) to further control for interactions among the effects. We consider prior features $\boldsymbol{Z}$ on the exposures, which are supposed to be common to all the J clusters. The resulting model can be expressed as

$$\beta_{kj} = \psi_{k0} + \sum_{r=1}^{R} \psi_{kr} v_{jr} + \sum_{q=1}^{Q} \pi_q z_{qk} + \delta_{kj} \tag{6}$$

where δ_{kj} are the residuals which represent the effect not captured by the whole set of level-2 covariates for each exposure and cluster. They are now assumed to be independent, as well as normal distributed quantities having null means and variances τ_k^2. We further simplify the analysis by supposing the level-2 residuals u_{0j} and δ_{kj} are independent. As a result, the more the exposures share similar features in both the prior data in $\boldsymbol{Z}$ and group membership, the more the corresponding estimates will be alike.

The basic level-2 regression (6) for correlated effects and nested data offers the opportunity of a wide range of generalizations, as well as simpler submodels. Some examples lie into including cluster-specific prior data on the exposures ($\boldsymbol{Z}_j$) or constraining to a constant level-2 variance (in order to reduce the effort in the estimation process) or removing the group covariates (e.g., if they are not available). Anyway, the model specification depends on the problem at hand and needs to be supported by reasonable assumptions for the application.

4. The Bayesian perspective

Under the frequentist perspective, according to the model specification (2), (3) and (6) there are both fixed ($\boldsymbol{\gamma}$, $\boldsymbol{\psi}_{k^*}$, $\boldsymbol{\pi}$, ϕ_0^2, τ_k^2) and random coefficients (α_j, $\boldsymbol{\beta}_j$, u_{0j}, $\boldsymbol{\delta}_j$). This is clearly an EB approach, as for the random intercepts α_j and effects $\boldsymbol{\beta}_j$ specific prior distributions based on the additional data are fully assigned through the regressions in (3) and (6). In this framework, different methods for fitting the model have been proposed. In order to maximize the likelihood respect to the parameters to be estimated, the integral defining this function can be evaluated by Monte-Carlo techniques (Gelman *et al.* (2003)) or by approximation, such as penalized quasi-likelihood (Breslow and Clayton (1993)), pseudo-likelihood (Wolfinger and O'Connel (1993)) and other related methods, which in the simplest cases can be carried out with available procedures in ordinary statistical softwares (Witte *et al.* (1998); Witte *et al.* (2000)).

However, the frequentist EB method often yields null estimates for the level-2 variances τ_k^2 leading to an extreme shrinkage estimation of the target vectors $\boldsymbol{\beta}_j$ toward the estimated prior means. This seems more likely to reflect a marginal likelihood for τ_k^2 with peak at zero, rather than true under dispersion (Greenland (1992)). Moreover, a credible result would achieve a more reasonable positive value for τ_k^2. Indeed, it represents the uncertainty about the residuals δ_{kj} and therefore also about the estimation of β_{kj} after incorporating the level-2 information. In particular, if τ_k^2 tends to ∞, the hierarchical model and the conventional logistic regression come to the same results according to the estimates of β_{kj}. On the contrary, if $\tau_k^2 = 0$, then the residuals δ_{kj} results to be null, meaning that we implicitly assume the absence of any effects of first-stage covariates in $\boldsymbol{X}_j$ beyond those of second-stage regressors in $\boldsymbol{Z}_j$ and $\boldsymbol{V}_j$.

Previous works suggest the SB approach as a good and easy strategy to tackle the problem of null estimation of the level-2 variance parameters (Greenland (1992)) by setting specific suitable values for τ_k^2. In particular, SB estimates appear to be better than EB estimates when the sample sizes and the ratio of subjects to parameters are small. Moreover, they are proved to be robust to misspecification (Greenland (1993)). Besides being frequently employed for the point-assignment of the level-2 variances, such as τ_k^2, the SB approach can be also used to specify the level-2 intercepts. For instance, the intercepts ψ_{k0} in model (6), which reflect the knowledge about any residual effects of the exposures due to level-2 covariates not included in the analysis, can be differently specified for each exposure with a positive (corresponding to a causative residual effect on the disease), negative (preventive residual effect) or null (negligible or null residual effect) value. For more details see Witte *et al.* (1994). Anyway, a great caution to overspecify these values is required, especially when either the sample size or number of parameters are large.

Conversely, the BEB approach offers a more appropriate framework. Indeed, it gives the opportunity of assigning reasonable priors for the parameters τ_k^2 by letting the data contribute to their final estimation. As an example, we can suitably suppose that the value of each τ_k^2 will be small, on the grounds that most important level-2 covariates have been included in the analysis. As a consequence, we can identify reasonable values on the residual variation of the logarithm of the effects β_{kj} (i.e., the Odds Ratio (OR) of each exposure), reflecting both the prior guess and the corresponding uncertainty for τ_k^2. For instance, a prior guess $\tau_k^2 = 0.18$ implies a 95% *a priori* certainty that the residual OR for the effect of a given unit increase in the k-th exposure lies in a 4-fold range. In order to reflect our uncertainty in this prior guess, we can further believe that, e.g., 8-fold variation between the upper and the lower 5% of units is very unlikely (say, less than a 1% chance). These two assumptions are sufficient to fully specify a proper hyperprior for τ_k^2 (or for the precision τ_k^{-2}) (Gelman *et al.* (2003)).

Actually, under this Bayesian setting, all the parameters are random with their own prior distribution to be specified. With this aim, the conjugacy among the distributions of the parameters at different levels can be exploited to assign informative or, at least, noninformative priors (Gelman *et al.* (2003)).

If properly used, the BEB approach represents the best compromise between the SB analysis for small studies and the EB method in large samples. Moreover, the BEB modeling can be regarded as a natural generalization of the EB and SB approaches, involving an additional stage in the hierarchy of models which represents the hyperprior distributions on parameters defining the descriptive level-2 prior information.

Mainly thanks to the recent development of computational methods, such as Monte Carlo Markov Chain (MCMC) techniques together with Gibbs sampling or Metropolis-Hastings algorithm (Gelman *et al.* (2003); Carlin and Louis (1998)), BEB analysis are now more practical to be employed and can be entirely exploited in their potentials.

5. Application

We consider a sample of $24,376$ individuals nested in 27 European centers of recruitment drawn from the European Prospective Investigation into Cancer and Nutrition (EPIC) study ‡. Subjects who developed a colon-rectum cancer after the enrollment and until the last observed year (i.e., 2005) are included in the analysis. Then, a number of controls are randomly selected to be equal to 5% of the whole control units, separately by center. Some descriptive statistics about the sample data are reported in Table 1.

The main aim of the analysis is to evaluate the effect of multiple dietary exposures on the occurrence of colon-rectum cancer cases, separately by center membership. Indeed, empirical evidence shows significant differences among these groups with respect to the occurrence of the disease (Pearson chi-squared= 542.7; p-value= 0.000).

The dietary information collected during the enrollment refers to the internal EPIC-SOFT food classification system and the corresponding individual food intakes are expressed in grams-per-day (gm/d). A list of 30 food groups are selected to be analyzed according to the suggestions of nutritionists and epidemiologists working on the study (Table 2).

Additional dietary information on the nutrient compositions are further available. In detail, these concern the amounts of constituents for one gram of each food. These data are arranged in matrices where the generic k-th row refers to the amounts of food constituents for the k- dietary exposure. Such matrices are usually named tables of nutrient composition and may vary between countries and centers. As a result, they can be generally regarded as center-specific information which can further contribute to model the variability among the centers. According to the dietary items involved into the analysis, we select a list including the most considerable nutrients (Table 3).

We employ a hierarchical Bayesian model to analyze these data by controlling for both the multilevel structure of the data (i.e., the within-center dependence) and the correlation among the dietary exposures. At level 1, we consider the logistic regression (2), where both the intercepts α_j and the dietary coefficients $\boldsymbol{\beta}_j = [\beta_{1j}, \ldots, \beta_{Kj}]$ vary across the centers denoted by j. In this case, the disease outcome represents the individual indicator of presence ($y_{ij} = 1$) or absence ($y_{ij} = 0$) of colon-rectum cancer up to year 2005 and the food intakes are in the matrix $\boldsymbol{X}_j = [x_{ikj}]$. As potential confounders, arranged in matrix $\boldsymbol{W}_j = [w_{ipj}]$, we consider: age at recruitment, gender, body mass index (BMI), smoking status (smoker-never-former-unknown), physical activity at work (sedentary occupation-standing occupation-manual work-heavy manual work-non worker-unknown), alcohol intake. The corresponding effects are in vector $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_P]$, which are fixed among the centers.

At level 2, since no additional covariates for the centers are available, we consider an empty model for the intercepts which split the random parameter into a common effect, ψ_0, and a residual term, u_j, yielding the differences among the centers:

$$\alpha_j = \psi_0 + u_j \tag{7}$$

‡EPIC is an ongoing multi-center study designed to investigate the relationship between nutrition and cancer, with the potential for studying other diseases as well. Its participants have been enrolled from several centers in 10 European countries and followed for cancer incidence and cause-specific mortality for several decades. During the enrollment, which took place between 1992 and 2000, information was collected through a non-dietary questionnaire on lifestyle variables and through a dietary questionnaire (EPIC Large scale Intake Assessment) addressing usual diet (see Riboli and Kaaks (1997) and Riboli *et al.*(2002)). The EPIC study is coordinated by the Nutrition and Hormones Group of the International Agency for Research on Cancer (IARC) in Lyon, France.

where the u_j are assumed to be independent and normally distributed with null means and common variances ϕ^2.

Moreover, the data on constituents for each food are used to develop the level-2 model for the dietary coefficients. In detail, these are regressed on the nutrient covariates $\boldsymbol{Z}_j = [z_{qkj}]$ as follows:

$$\beta_{kj} = \pi_0 + \sum_{q=1}^{Q} \pi_q z_{qk} + \delta_{kj} \tag{8}$$

where we assume that the effects of the food exposures on the colon-rectum cancer are partially mediated by the effects of nutrients, $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_Q]$.

In this case, we can suitably suppose that the values of the level-2 residual variances related to model (8) will be small for all the food effects. Indeed, once the center-specific information on nutrients are considered, we believe that the variability of dietary effects among the centers would be entirely explained. As a results, the residuals δ_{kj} can be assumed to hold the simple hypothesis of independence and normal distribution with null means and constant variances, τ^2, and to be further independent on u_j.

In a BEB perspective, at level 3 the prior distributions for the other parameters are specified. We mainly focus our attention on the crucial level-2 variance τ^2§. As introduced above, we attempt to specify an informative hyperprior distribution for τ^2, based on plausible ranges of variation for log normal random dietary effects. More specifically, since at level 2 the log ORs (i.e., $\boldsymbol{\beta} = \{\beta_{kj}\}$) are supposed to be normally distributed with means $\mu(\boldsymbol{\beta}) = \boldsymbol{Z_j}\boldsymbol{\pi}$ and variance τ^2, the precision τ^{-2} can be expressed as

$$\tau^{-2} = (2 \times 1.645)^2/(\beta_{95\%} - \beta_{5\%})^2.$$

Then, we believe a 2-fold variation between the ORs for the upper and lower 5% of units is reasonable, that is $\beta_{95\%} - \beta_{5\%} = log2$. Hence our prior guess at τ^{-2} is $\tau^{-2} \approx 3.29^2/(log2)^2 \approx 22.53$, corresponding to a level 2 standard deviation τ equals to 0.21. To reflect our uncertainty in this prior guess, we believe that 4-fold variation between the upper and the lower 5% of units is very unlikely (say, less than a 1% chance). Thus, lower 1% quantile of our prior distribution for the precision τ^{-2} can be supposed to be $\tau_{1\%}^{-2} \approx 3.29^2/(log4)^2 \approx 5.63$.

In this application, the previous assumptions imply 95% a priori certainty that the residual OR for the effect of a given unit increase in the k-th dietary exposure lies in a 2-fold range and 99% certainty that it lies in a 4-fold range. For instance, supposing the prior mean of β_{kj} is -0.02 for an increase of 36.6 grams per day of leafy vegetables we are 95% certain that the corresponding residual OR lies in the range from 0.69 to 1.38, and 99% certain it lies from 0.49 to 1.96. Assuming both the hypothesis are consistent with the data and the previous knowledge, we specify an informative proper distribution for the hyperparameter τ^{-2}, that is a Gamma probability distribution of parameters 5 and 0.22 for shape and rate, respectively.

The Bayesian approach would ensure that inference about every parameter fully takes into account for the uncertainty about all other parameters. As a result, it provides the estimation of the joint posterior distribution for all the unknown parameters summarized into vector $\boldsymbol{\theta}$ as stated by the Bayes theorem:

$$p(\boldsymbol{\theta}|\boldsymbol{Y}) = \frac{f(\boldsymbol{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{h(\boldsymbol{Y})} \tag{9}$$

§For the other hyperparameters we assign proper non-informative priors.

where $f(\boldsymbol{Y}|\boldsymbol{\theta})$ is the likelihood function, $p(\boldsymbol{\theta})$ includes all the (hyper-) prior distributions on the parameters and

$$h(\boldsymbol{Y}) = \int f(\boldsymbol{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

is the marginal distribution of $\boldsymbol{Y}$, i.e. a kind of normalizing constant ensuring the joint posterior is a probability distribution.

Inferences about the parameters of crucial interest are ensued by averaging over auxiliary parameters. We particularly focus our attention on the vector of parameters $\boldsymbol{\beta}$, reflecting the effects of dietary exposures on colon-rectum cancer for each area of enrolment. Therefore, the corresponding posterior distribution is computed as

$$p(\boldsymbol{\beta}|\boldsymbol{Y}) = \int \dots \int p(\boldsymbol{\theta}|\boldsymbol{Y})\, d\boldsymbol{\gamma}\, d\psi_0\, d\boldsymbol{u}\, d\phi^2\, d\boldsymbol{\pi}\, d\boldsymbol{\delta}\, d\tau^2. \tag{10}$$

The need for numerical integration is avoided by taking repeated samples from the posterior distributions using the MCMC methods and Gibbs sampling. These procedures are implemented by using the software WinBUGS, version 1.4 (Spiegelhalter *et al.* (2003)). A total of $30,000$ iterations were run with a burn-in of $20,000$.

In order to measure the improvement in the estimates of dietary effects, we compare the results from this hierarchical Bayesian model with those obtained by carrying out several conventional analysis (1), separately by center of enrollment j.

Some results are showed in Tables 4 to 9, where the ORs and their 95% confidence intervals (CI) are calculated according to food-specific values of unit increase which are the sample standard deviations in Table 2.

The results from the conventional disease model are notably affected by problems of sparse data which preclude the full estimation of each dietary effect on the occurrence of colon-rectum cancer. In some cases, the ML estimation fails to converge because the predictors are highly correlated. Even when the convergence is achieved, a great number of estimates result with large and unstable absolute values, suggesting implausible strong associations according to the to the relevant diet and colon-rectum cancer literature. Moreover, when the results are compared across different areas, there are discordant values. As an example, let's consider the extremely large and unstable estimation of the cabbages effect in Turin (OR=5.503 and CI=$0.089 - 340.060$ for 37.9 grams of unit increase). This estimate appears to be strongly different from the most part of the corresponding results in other areas, which identify the intakes of cabbages as a protective factor for colon-rectum cancer.

When the hierarchical Bayesian model is fitted, formerly extreme and unstable estimates become more reasonable and less biased, even when the results on the same exposure are compared across different centers. For example, the excessive risk factor for additional 31.2 grams per day of processed meat in the south coast of France from the ordinary model (OR =1.901 and CI=0.777 – 4.654) becomes more realistic and stable (OR=1.076 and CI=$0.934 - 1.241$); and the estimate of the effect of milk and milk beverages in the north & west of Norway becomes consistent with the results in the other centers (OR from 2.275 to 0.964). On the other hand, stable conventional estimates remain much more the same (see, e.g., the estimates for legumes in San Sebastian or milk and milk beverages in Malmo). In these cases, great gains in term of standard errors are often reported. For instance, the effect of eating fish in Copenhagen shows similar estimates for both methods, but the improvement in the corresponding standard errors returns results which are significant.

The improvement on dietary estimation is mainly due to the shared food information on nutrients also across different centers. As a result, dietary estimates are pulled toward each other when they have similar compositions. Therefore, we expect this shrinkage especially occurs for the same exposures evaluated in different centers as their levels of nutrients are more likely to be similar. Indeed, previous evidence (Roli (2006)) showed that the the substantive improvements in the estimation of dietary effects are gained when a single multilevel analysis is carried out, while the inclusion of nutrient information alone for separate conventional regressions does not yield as good results.

The shrinkage of the estimates can be evaluated in practice by plotting the results from the conventional regressions and from the hierarchical Bayesian method, simultaneously (Figure 1). Indeed, for the former we can observe a great variability with peaks of extremely high and extremely low numbers. Conversely, the estimates from our model are closer to each other (i.e., to the prior means based on the nutrients) and are controlled for variations due to random occurrences in small samples.

6. Discussion and conclusions

Statistical theory, several simulation studies and a large number of applications all support the use of hierarchical modeling as a powerful method which allows to yield strong gains in the accuracy of predictions and effect estimates. The improvement is mainly due to the use of prior data arranged in an additional model. As a result, the ordinary estimates from the conventional level-1 model are pulled or 'shrunk' toward each other when they have similar levels of prior data.

In epidemiological field, hierarchical methods are strongly recommended to address estimation problems of multiple exposures, whose effects are often correlated, and to analyze small data set. Moreover, these complications commonly affect multicentric studies where the area-specific estimation of such effects is of crucial importance, but the independence among the units belonging to the centers is violated. In these cases, the hierarchical modeling implicitly assume the so-called exchangeability hypothesis, which states that the more the areas (or the exposures) have similar features (i.e., prior data), the more the corresponding parameters are likely to be close. In sparse samples, the exchangeability assumption is fundamental because it allows to mediate the poor or missing level-1 data of some groups by sharing the corresponding non-missing information with other groups having similar priors.

In multiple regression analysis, the hierarchical framework can further provide an alternative to conventional variable selection techniques (Gelman and Hill (2007)). These procedures begin with a maximal model including all the terms (such as in backward elimination) or a minimal model that has only the essential regressors, i.e. the confounders, (such as in forward and stepwise selection) and proceed with a model reduction based on some significance criteria to search for a final model. The hierarchical approach states the maximal model as the level-1 regression. Then, it specifies a level-2 model, where the corresponding values of the residual variances mark the degree of compromise between the extremes of putting each variable completely in or completely out of the model. Therefore, when these level-2 variances are null, then a minimal model holds; conversely, if they are large, the final model tends to be the maximal one. Moreover, the hierarchical approach does not make a definitively "all-or-nothing" choice for each term, but allows to retain all the variables in the analysis in order to be further evaluated whenever additional information would be available.

The use of the BEB perspective is proposed as a reasonable and flexible strategy to avoid prior restrictions regardless the sample data information, such as in the SB approach.

Furthermore, it is a natural completion of the hierarchical model structure which develops by specifying hyperprior probability distributions to be consistent with data and previous knowledge. The estimation of parameters of interest is supported by the computational powerful of recent softwares, such as WinBUGS (Spiegelhalter *et al.* (2003)), an interactive Windows version of the BUGS program for Bayesian analysis of complex statistical models implementing MCMC techniques and Gibbs sampling.

The advantages related to the use of hierarchical methods under a BEB framework are highlighted by the results of the empirical illustration, where for a multicentric study the ordinary ML estimates of multiple dietary effects are improved, for each center separately, by a hierarchy of models merging and exploiting all the prior knowledge about the problem at hand. The improvement is expressed in terms of more plausible estimates of dietary effects and lower mean-squared errors than traditional data summaries, thanks to a two-fold shrinkage action due to the similar nutrient compositions of dietary items between and within the centers.

If one is interested in the evaluation of effects of the level-2 covariates, a single level-1 conventional regression on nutrient intakes can be carried out. But in this case the unmeasured constituents and their interactions that might be responsible for some dietary item effects would be ignored. Conversely, the hierarchical model can offer a more realistic and generic representation of data allowing for the possibility that there are food effects beyond the nutrient contribution, as well as food interactions which are important to be investigated. Indeed, understanding dietary effects is crucial for development of public health recommendations and these effects are not captured by the effects of nutrients alone. Moreover, the hierarchical approach provides the food constituents to be estimated, that may be alike useful from a nutritional point of view (e.g., for the formulation of a balanced diet).

The sample size limits the performance of BEB hierarchical model and the number of level-2 covariates to be embedded into the analysis. As a result, some problems during the estimation process can be encountered when a large number of parameters have to be estimated. Therefore, only potentially relevant covariates, about which useful descriptive information are available, are recommended to be included in the level-2 model.

The hierarchical Bayesian model we propose can be further applied in many other epidemiologic contexts. For instance, in occupational studies, where more levels of information can be merged; or to perform polytomous logistic regressions of different causes of death on a set of exposures; or in disease mapping and spatial analysis, where the variations due to random occurrences need to be controlled by exploiting the spatial proximity and the consequent interaction of the geographical areas.

In all these examples, the use of hierarchical Bayesian modeling can be easily extended or raised thank to the substantial gains that it can be yield and its internal flexibility as regards to the prior assumptions. This paper is intended to encourage the use of Bayesian methods in epidemiology as a powerful statistical tool to address the problem of nested data and correlated effects. Indeed, in the simplest cases the implementation of such methods can be carried out by standard frequentist softwares (Witte *et al.* (1998); Greenland (2006); Greenland (2007)). Conversely, if the full flexibility of MCMC posterior sampling is required to analyze more complex model structures, some knowledge of Bayesian statistical theory and computation is needed. Anyway, the role of epidemiologists remains of primary importance and they should be closely involved into the crucial phase of model specification.

References

Bernardinelli L, Clayton D, Pascutto C, *et al.* (1995) Bayesian analysis of space-time variation in disease risk. *Statist. Med.* **14**:2433-2443.

Breslow N.E., Clayton D.G. (1993) Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass.*, **88**: 9-25.

Burgess JF Jr., Christiansen CL, Michalak SE, et al. (2000) Medical profiling: improving standards and risk adjustments using hierarchical models. *J. Health Econ.* **19**:291-309.

Carlin B., Louis T. (1998) *Bayes and empirical Bayes methods for data analysis.* Chapman and Hall/CRC.

Cubbin C, Winkleby M.A. (2005) Protective and harmful effects of neighborhood-level deprivation on individual-level health knowledge, behavior changes, and risk of coronary heart disease. *Am. J. Epidemiol.* **162**:559-68.

Diez-Roux A.V. (2000) Multilevel anlaysis in public health research. *Annu. Rev. Public Health*, **21**:171-92.

Diez-Roux A.V. (2004) The study of group-level factors in epidemiology: rethinking variables, study designs, and analytical approaches. *Epidemiol. Rev.*, **26**:104-111.

Deeley J.J., Lindley D.V. (1981) Bayes Empirical Bayes. *J. Am. Statist. Ass.*, **76**: 833–841.

Gelman A., Carlin J.B., Stern H.S., Rubin D.B. (2003) *Bayesian Data Analysis.* 2nd edn. New York: Chapman and Hall/CRC.

Gelman A., Hill J. (2007) *Data analysis using regression and multilevel/hierarchical models.* Cambridge University press.

Goldstein H. (1999) *Multilevel statistical models.* London: Institute of education, multilevel models project. (Available from http://www.arnoldpublishers.com/support/goldstein.htm).

Graham P. (2008) Intelligent Smoothing Using Hierarchical Bayesian Models. *Epidemiology*, **19**: 493-495.

Greenland S. (1992) A semi-Bayes approach to the analysis of correlated multiple associations, with an application to an occipational cancer-mortality study. *Statist. Med.*, **11**: 219–230.

Greenland S. (1993) Methods for epidemiologic analysis of multiple exposures: a review and a comparative study of maximum-likelihood, preliminary testing and empirical Bayes regression. *Statist. Med.*, **12**: 717–736.

Greenland S. (1997) Second-stage least squares versus penalized quasi-likelihood for fitting hierarchical models in epidemiologic analysis. *Statist. Med.*, **16**: 515-.526.

Greenland S. (2000) Principles of multilevel modelling. *Int. J. Epidemiol.*, **29**: 158–167.

Greenland S. (2006) Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *Int. J. Epidemiol.*, **35**: 765-775.

Greenland S.(2007) Bayesian perspectives for epidemiological research: II. Regression analysis. *Int. J. Epidemiol.*, **36**: 195-202.

Hox J.J. (1995) *Applied multilevel analysis.* TT-Pubblikaties, Amsterdam.

Lawson AB. (2001) Disease map reconstruction. *Statist. Med.* **20**:2183-2204.

Leyland A., Goldstein H. (2001) *Multilevel modelling of health statistics.* John Wiley.

Maritz J., Lwin T. (1989) *Empirical Bayes Methods.* Chapman and Hall/CRC.

MacLehose R.F., Dunson D.B., Herring A.H., Hoppin J.A. (2007) Bayesian methods for highly correlated exposure data. *Epidemiology*, **18**:199-207.

Morris C. (1983) Parametric empirical Bayes; theory and applcations (with discussion). *J. Am. Statist. Ass.*, **178**: 47–65.

Raudenbush S.W., Bryk A.S. (2002) *Hierarchical Linear Models - Application and data analysis methods.* Second edition. Sage.

Riboli E., Kaaks R. (1997) The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int. J. Epidemiol.*, **26**: Suppl 1:6-14.

Riboli E, Hunt KJ, Slimani N, *et al.* (2002) European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr.*, vol.5; **6B**:1113–24.

Roli G. (2006) Hierarchical logistic regression in a multicentric study of multiple dietary effects on a disease outcome: a fully Bayesian approach. PHD thesis. (Available from http://www2.stat.unibo.it/Dottorato/MSRS/TesiDottoratoMSRS/2006/RoliGiulia.pdf).

Rothman K.J., Greenland S., Lash T.L. (2008) *Modern epidemiology.* 3rd ed. Philadelphia: Lippincott-Williams-Wilkins.

Snijders T., Bosker R. (1999) *Multilevel analysis: an introduction to basic and advanced multilevel modeling.* Sage.

Spiegelhalter D., Thomas A., Best N., Lunn D. (2003) *WinBUGS User Manual*, Version 1.4.

Thomas D.C., Siemiatycki J., Dewar R., *et al.* (1985) The problem of multiple inference in studies designed to generate hypotheses. *Am. J. Epidemiol.* **122**:1080-1095.

Witte J., Greenland S., Haile R., Bird C. (1994) Hierarchical regression analysis applied to a study of multiple dietray exposures and breast cancer. *Epidemiology*, vol.5; **6**: 612–621.

Witte J., Greenland S., Kim L.L. (1998) Software for Hierarchical Modeling of Epidemiological Data. *Epidemiology*, vol.9; **5**: 563–566.

Witte J., Greenland S., Kim L.L., Arab L. (2000) Multilevel Modeling in Epidemiology with GLIMMIX. *Epidemiology*, vol.11; **6**: 684–688.

Wolfinger R., O'Connel M. (1993) Generalized linear mixed models: a pseudo-likelihood approach. *J. Statist. Comput. Simul.*. **48**: 223–243.

Figures and Tables

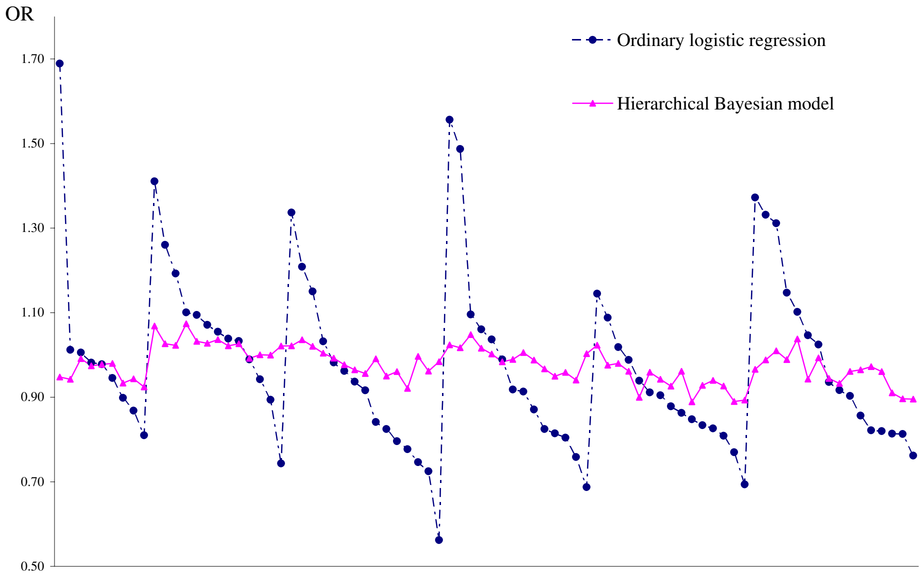

Fig. 1. Estimated ORs

Table 1. Descriptive statistics.

Center	Women	Age	BMI	Cases	Controls	Total
	(%)	mean (sd)	mean (sd)			
North-East of France	100.0	56.4 (6.7)	23.5 (3.5)	83	1525	1608
North-West of France	100.0	54.5 (7.1)	22.9 (3.3)	35	538	573
South of France	100.0	56.1 (5.9)	23.1 (3.4)	47	853	900
South coast of France	100.0	55.3 (5.4)	23.2 (3.0)	20	457	477
Florence	61.9	54.8 (6.1)	25.8 (3.5)	54	637	691
Varese	76.4	55.3 (7.4)	25.9 (4.3)	47	559	606
Ragusa	41.4	53.8 (5.8)	27.4 (3.9)	13	296	309
Turin	21.7	58.1 (3.8)	26.6 (3.7)	27	482	509
Naples	100.0	57.5 (7.8)	27.2 (4.8)	12	247	259
Asturias	51.3	54.1 (7.7)	28.3 (3.9)	22	413	435
Granada	58.6	54.7 (7.6)	30.6 (4.5)	18	378	396
Murcia	63.2	52.0 (8.9)	28.7 (4.6)	17	410	427
Navarra	39.3	55.4 (5.7)	29.4 (3.6)	28	388	416
San Sebastian	33.8	53.6 (7.4)	27.9 (3.7)	36	406	442
Cambridge	44.9	65.0 (7.8)	26.3 (3.8)	154	1112	1266
Oxford Health conscious	67.8	63.7 (13.0)	24.0 (3.7)	95	2297	2392
Oxford General population	61.5	56.7 (7.2)	26.0 (4.3)	28	335	363
Bilthoven	34.4	53.6 (6.4)	26.2 (3.7)	33	1079	1112
Utrecht	100.0	60.4 (6.0)	25.7 (4.0)	135	783	918
Heidelberg	28.0	56.9 (5.7)	27.1 (4.0)	82	1185	1267
Potsdam	42.2	57.0 (6.8)	27.2 (4.1)	90	1282	1372
Malmo	51.6	61.2 (6.6)	25.8 (3.9)	194	1206	1400
Umea	43.5	56.6 (5.1)	25.7 (3.9)	83	1212	1295
Aarhus	46.4	58.3 (4.4)	26.0 (3.9)	125	824	949
Copenhagen	44.4	58.5 (4.2)	26.2 (4.1)	286	1906	2192
South & Est of Norway	100.0	51.8 (3.8)	24.6 (4.0)	29	970	999
North & West of Norway	100.0	50.6 (3.5)	25.3 (3.6)	15	790	805
Total	58.5	58.4 (6.3)	25.9 (3.9)	1808	22568	24376

Table 2. Dietary items and corresponding average intakes and standard deviations (gm/d).

Dietary items	Mean	SD
Potatoes and Other Tubers	108.097	80.743
Leafy Vegetables	23.324	36.617
Fruiting Vegetables	55.570	49.021
Root Vegetables	27.350	32.656
Cabbages	25.992	37.925
Grain and Pod Vegetables	8.879	13.569
Stalk Vegetables, Sprouts	8.974	12.102
Mixed Salad, Mixed Vegetables	13.769	29.568
Legumes	11.463	21.770
Fruits	218.786	171.468
Nuts and Seeds	3.189	8.040
Mixed Fruits	3.881	12.097
Milk + Milk beverages	226.287	230.397
Yogurt	67.816	92.272
Fromage blanc, petit suisse + Cheeses	44.068	41.252
Pasta, rice, other grain	51.657	61.304
Crispbread, Rusks	8.593	15.972
Breakfast Cereals	22.014	55.757
Beef	19.651	20.464
Pork	18.989	19.819
Poultry	24.664	27.979
Processed meat	33.931	31.253
Fish	29.514	28.825
Eggs and Egg Product	18.833	18.136
Vegetable Oils	7.224	11.452
Margarines	15.506	17.607
Deep Frying Fat	0.040	0.553
Chocolate + Confectionery + Syrup	13.902	20.901
Coffee	452.990	400.388
Sauces	22.872	22.101

Table 3. Nutrients and corresponding unit of measurement.

Total proteins (g)
Saturated fatty acids (g)
Monosaturated fatty acids (g)
Polyunsaturated fatty acids (g)
Starch (g)
Sugar (g)
Fibre (g)
Calcium (mg)
Iron (mg)
Vitamin D (μg)
Vitamin E (mg)
Beta-carotene (μg)
Retinol (performed vitamin A) (μg)
Vitamin C (mg)

Table 4. Results: milk and milk beverages.

Center	Ordinary logistic regression OR (95% CI)	Hierarchical Bayesian model OR (95% CI)
North-East of France	0.892 (0.624-1.276)	0.923 (0.809-1.051)
North-West of France	-	0.948 (0.822-1.099)
South of France	-	0.904 (0.787-1.039)
South coast of France	0.079 (0.009-0.700)	0.896 (0.772-1.039)
Florence	0.883 (0.487-1.600)	0.925 (0.807-1.064)
Varese	-	0.944 (0.822-1.087)
Ragusa	-	0.945 (0.820-1.096)
Turin	0.466 (0.154-1.410)	0.917 (0.790-1.060)
Naples	-	0.942 (0.813-1.093)
Asturias	0.889 (0.401-1.968)	0.914 (0.792-1.053)
Granada	-	0.926 (0.802-1.073)
Murcia	-	0.914 (0.787-1.062)
Navarra	-	0.900 (0.779-1.033)
San Sebastian	1.120 (0.624-2.010)	0.923 (0.800-1.064)
Cambridge	0.989 (0.778-1.256)	0.953 (0.848-1.070)
Oxford Health conscious	0.851 (0.655-1.106)	0.921 (0.817-1.042)
Oxford General population	0.505 (0.258-0.989)	0.892 (0.770-1.024)
Bilthoven	0.616 (0.375-1.013)	0.895 (0.784-1.022)
Utrecht	1.070 (0.893-1.281)	0.989 (0.887-1.097)
Heidelberg	1.064 (0.808-1.400)	0.969 (0.853-1.099)
Potsdam	-	0.901 (0.787-1.027)
Malmo	0.999 (0.853-1.171)	0.962 (0.870-1.063)
Umea	1.187 (0.876-1.610)	0.982 (0.861-1.120)
Aarhus	1.026 (0.866-1.217)	0.969 (0.871-1.074)
Copenhagen	0.930 (0.838-1.031)	0.924 (0.857-0.997)
South & Est of Norway	-	0.948 (0.822-1.097)
North & West of Norway	2.275 (0.620-8.340)	0.964 (0.833-1.119)

Table 5. Results: fruits.

Center	Ordinary logistic regression OR (95% CI)	Hierarchical Bayesian model OR (95% CI)
North-East of France	1.436 (1.084-1.902)	1.074 (0.948-1.218)
North-West of France	-	0.990 (0.863-1.133)
South of France	-	1.024 (0.905-1.158)
South coast of France	1.434 (0.856-2.402)	1.008 (0.879-1.158)
Florence	0.523 (0.328-0.830)	0.883 (0.777-1.000)
Varese	-	0.960 (0.845-1.091)
Ragusa	-	0.936 (0.820-1.066)
Turin	1.664 (1.030-2.690)	1.008 (0.878-1.158)
Naples	-	0.973 (0.846-1.117)
Asturias	0.622 (0.375-1.031)	0.942 (0.818-1.076)
Granada	-	0.977 (0.848-1.124)
Murcia	-	0.971 (0.846-1.114)
Navarra	-	1.018 (0.894-1.165)
San Sebastian	0.673 (0.460-0.980)	0.903 (0.794-1.024)
Cambridge	1.054 (0.854-1.301)	1.019 (0.911-1.138)
Oxford Health conscious	0.742 (0.576-0.956)	0.909 (0.812-1.015)
Oxford General population	1.219 (0.688-2.160)	1.006 (0.875-1.158)
Bilthoven	0.751 (0.374-1.509)	0.958 (0.828-1.108)
Utrecht	1.180 (0.947-1.470)	1.033 (0.915-1.167)
Heidelberg	1.139 (0.681-1.904)	0.992 (0.862-1.142)
Potsdam	-	0.987 (0.863-1.135)
Malmo	0.833 (0.649-1.067)	0.948 (0.844-1.067)
Umea	0.820 (0.560-1.200)	0.964 (0.843-1.101)
Aarhus	0.660 (0.492-0.885)	0.883 (0.779-0.995)
Copenhagen	1.015 (0.867-1.189)	0.995 (0.902-1.098)
South & Est of Norway	-	1.015 (0.877-1.176)
North & West of Norway	2.224 (0.600-8.240)	0.981 (0.847-1.142)

Table 6. Results: processed meat.

Center	Ordinary logistic regression OR (95% CI)	Hierarchical Bayesian model OR (95% CI)
North-East of France	0.985 (0.684-1.419)	1.035 (0.912-1.175)
North-West of France	- -	1.019 (0.886-1.174)
South of France	- -	1.053 (0.918-1.203)
South coast of France	1.901 (0.777-4.654)	1.076 (0.934-1.241)
Florence	0.635 (0.336-1.200)	0.986 (0.859-1.130)
Varese	-	0.978 (0.853-1.118)
Ragusa	-	1.031 (0.895-1.192)
Turin	1.566 (0.503-4.870)	1.013 (0.875-1.174)
Naples	-	1.008 (0.873-1.167)
Asturias	1.111 (0.579-2.130)	1.014 (0.881-1.162)
Granada	-	0.999 (0.871-1.147)
Murcia	-	1.013 (0.888-1.149)
Navarra	-	1.030 (0.904-1.177)
San Sebastian	1.440 (1.062-1.950)	1.075 (0.947-1.222)
Cambridge	1.123 (0.857-1.472)	1.030 (0.915-1.162)
Oxford Health conscious	1.040 (0.733-1.476)	1.017 (0.899-1.152)
Oxford General population	0.720 (0.315-1.645)	1.005 (0.873-1.159)
Bilthoven	0.916 (0.608-1.379)	1.092 (0.942-1.266)
Utrecht	1.249 (0.933-1.670)	1.142 (0.990-1.311)
Heidelberg	1.023 (0.853-1.225)	1.036 (0.941-1.136)
Potsdam	-	1.033 (0.944-1.126)
Malmo	1.137 (0.989-1.306)	1.110 (1.011-1.213)
Umea	1.266 (0.921-1.740)	1.067 (0.934-1.218)
Aarhus	0.967 (0.697-1.341)	1.029 (0.903-1.175)
Copenhagen	1.044 (0.869-1.254)	1.068 (0.957-1.190)
South & Est of Norway	-	1.024 (0.885-1.181)
North & West of Norway	0.573 (0.115-2.860)	1.038 (0.900-1.199)

Table 7. Results: fish.

Center	Ordinary logistic regression OR (95% CI)	Hierarchical Bayesian model OR (95% CI)
North-East of France	0.809 (0.579-1.130)	0.927 (0.818-1.051)
North-West of France	-	0.976 (0.853-1.115)
South of France	-	1.029 (0.900-1.175)
South coast of France	0.310 (0.094-1.020)	0.957 (0.829-1.102)
Florence	0.904 (0.503-1.630)	0.943 (0.820-1.080)
Varese	-	0.941 (0.819-1.081)
Ragusa	-	0.969 (0.840-1.120)
Turin	0.911 (0.359-2.320)	0.959 (0.833-1.105)
Naples	-	0.941 (0.810-1.085)
Asturias	0.988 (0.607-1.610)	0.961 (0.845-1.095)
Granada	-	0.943 (0.828-1.074)
Murcia	-	0.984 (0.863-1.120)
Navarra	-	0.934 (0.826-1.056)
San Sebastian	0.834 (0.613-1.130)	0.928 (0.826-1.039)
Cambridge	1.145 (0.926-1.416)	1.023 (0.914-1.140)
Oxford Health conscious	0.879 (0.671-1.151)	0.926 (0.824-1.038)
Oxford General population	0.826 (0.396-1.725)	0.940 (0.817-1.077)
Bilthoven	0.442 (0.024-8.181)	0.920 (0.791-1.072)
Utrecht	0.939 (0.359-2.455)	0.900 (0.775-1.040)
Heidelberg	0.694 (0.419-1.149)	0.893 (0.769-1.036)
Potsdam	-	0.888 (0.762-1.030)
Malmo	1.019 (0.871-1.192)	0.980 (0.888-1.078)
Umea	0.770 (0.363-1.630)	0.890 (0.751-1.050)
Aarhus	1.088 (0.817-1.449)	0.976 (0.857-1.107)
Copenhagen	0.848 (0.693-1.037)	0.889 (0.799-0.989)
South & Est of Norway	-	0.999 (0.887-1.123)
North & West of Norway	0.863 (0.464-1.610)	0.961 (0.855-1.083)

Table 8. Results: legumes.

Center	Ordinary logistic regression OR (95% CI)	Hierarchical Bayesian model OR (95% CI)
North-East of France	0.958 (0.732-1.254)	0.967 (0.858-1.085)
North-West of France	-	0.945 (0.829-1.076)
South of France	-	1.002 (0.882-1.140)
South coast of France	2.424 (0.955-6.152)	1.003 (0.873-1.151)
Florence	0.422 (0.142-1.250)	0.947 (0.816-1.099)
Varese	-	0.980 (0.843-1.138)
Ragusa	-	0.975 (0.840-1.134)
Turin	0.324 (0.034-3.050)	0.977 (0.847-1.125)
Naples	- - -	0.978 (0.861-1.111)
Asturias	0.889 (0.645-1.225)	0.968 (0.869-1.077)
Granada	-	0.943 (0.826-1.075)
Murcia	-	0.971 (0.855-1.101)
Navarra	-	0.983 (0.885-1.091)
San Sebastian	0.938 (0.756-1.160)	0.974 (0.888-1.061)
Cambridge	1.105 (0.846-1.443)	1.029 (0.913-1.157)
Oxford Health conscious	1.273 (1.039-1.559)	1.075 (0.964-1.194)
Oxford General population	0.944 (0.436-2.041)	0.970 (0.846-1.115)
Bilthoven	1.181 (0.369-3.784)	0.992 (0.858-1.148)
Utrecht	0.770 (0.470-1.262)	0.945 (0.824-1.081)
Heidelberg	0.779 (0.400-1.519)	0.964 (0.841-1.110)
Potsdam	-	0.994 (0.865-1.144)
Malmo	0.944 (0.719-1.239)	0.986 (0.876-1.108)
Umea	0.645 (0.181-2.300)	0.977 (0.849-1.124)
Aarhus	0.609 (0.024-15.686)	0.962 (0.810-1.138)
Copenhagen	3.169 (0.589-17.057)	0.992 (0.843-1.173)
South & Est of Norway	-	1.008 (0.876-1.159)
North & West of Norway	-	0.998 (0.863-1.151)

Table 9. Results: cabbages.

Center	Ordinary logistic regression OR (95% CI)	Hierarchical Bayesian model OR (95% CI)
North-East of France	0.667 (0.338-1.315)	0.950 (0.833-1.087)
North-West of France	-	0.985 (0.855-1.135)
South of France	-	0.983 (0.856-1.129)
South coast of France	0.574 (0.056-5.911)	0.985 (0.858-1.138)
Florence	1.508 (0.133-17.060)	0.960 (0.826-1.112)
Varese	-	0.962 (0.825-1.116)
Ragusa	-	0.975 (0.841-1.135)
Turin	5.503 (0.089-340.060)	0.969 (0.836-1.128)
Naples	-	0.972 (0.826-1.137)
Asturias	0.751 (0.290-1.943)	0.968 (0.845-1.104)
Granada	-	0.987 (0.854-1.139)
Murcia	-	0.989 (0.858-1.139)
Navarra	-	1.035 (0.897-1.192)
San Sebastian	1.261 (0.538-2.950)	0.985 (0.860-1.132)
Cambridge	0.887 (0.763-1.032)	0.927 (0.849-1.011)
Oxford Health conscious	0.853 (0.718-1.013)	0.943 (0.860-1.031)
Oxford General population	0.925 (0.548-1.561)	0.967 (0.853-1.090)
Bilthoven	0.604 (0.197-1.858)	0.967 (0.838-1.113)
Utrecht	0.887 (0.566-1.393)	0.967 (0.845-1.106)
Heidelberg	1.833 (0.840-3.999)	0.998 (0.867-1.147)
Potsdam	-	0.956 (0.836-1.093)
Malmo	1.069 (0.805-1.420)	1.013 (0.897-1.147)
Umea	0.936 (0.419-2.090)	0.977 (0.853-1.124)
Aarhus	1.041 (0.513-2.114)	0.990 (0.863-1.137)
Copenhagen	0.784 (0.481-1.277)	0.955 (0.837-1.084)
South & Est of Norway	-	1.027 (0.903-1.162)
North & West of Norway	0.684 (0.161 2.910)	0.967 (0.844-1.108)