STATISTICA, anno LXXI, n. 4, 2011

IMPROVING THE ESTIMATION OF MULTIPLE CORRELATED DIETARY EFFECTS ON COLON-RECTUM CANCER IN MULTICENTRIC STUDIES: A HIERARCHICAL BAYESIAN APPROACH

G. Roli, P. Monari

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 neighbourhoods, towns or countries (Leyland and Goldstein, 2001). More generally, the nested structure of data is a common phenomenon, especially in behavioural 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).

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 and 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 modelling (see, e.g., Greenland, 1992; Witte *et al.*, 1994; Diez-Roux, 2000 and 2004). Over the last 50 years, hierarchical modelling 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 (Greenland, 2000; Raudenbush and Bryk, 2002; Graham, 2008).

When the structure of the data is nested, hierarchical modelling allows to handle simultaneously multiple levels of information and dependencies (Hox, 1995; Snijders and Bosker, 1999; Leyland and Goldstein, 2001; Raudenbush and Bryk, 2002). 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; Diez-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 and 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 modelling for correlated effects and 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 and 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 from 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 (Gelman *et al.*, 2003). 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; Carlin and Louis, 1998). 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 and 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 exploit the available prior data for some (hyper-) parameters and some kinds of proper distributions for the others (Deeley and Lindley, 1981). 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 at extending 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 and into a single analysis. These additional data are second-stage covariates which can arise from specific features of the clusters, as well as information about the regressors. 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.

The method we propose is conceived basing on a real study carried out at European level to investigate the association of dietary exposures with the occurrence of colon-rectum cancer on individual data. Thus, a multilevel setting is involved, as individuals are enrolled from different countries and centres of Europe, and we are interested in partitioning the different effects of dietary exposures across these centres. Moreover, additional information on the nutrient compositions of each dietary item are arranged to model the correlation among the exposures. Then, comparing our results with those obtained by several conventional regressions allows us to measure the gains in the final estimates of the crucial parameters.

The paper develops as follows. We firstly introduce the study and data used to develop the hierarchical regression method we propose. The model based on the real data and corresponding assumptions under the Bayesian framework are described in section 3. In section 4, we compare the hierarchical Bayesian regression method with the conventional regression results with respect to the study application. The last section summarizes our findings and concludes.

2. DATA: THE EPIC STUDY

We consider data drawn from the European Prospective Investigation into Cancer and Nutrition (EPIC) study. EPIC is an ongoing multi-centre 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 centres in 10 European countries and followed for cancer incidence and cause-specific mortality for several decades. During the enrolment, 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; 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.

In this work, we consider a sample of 24,376 individuals nested in the 27 European centres of recruitment. Subjects who developed a colon-rectum cancer after the enrolment and until the last observed year (i.e., 2005) are included in the analysis. Then, 5% of controls are randomly selected by centre membership. 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 centre 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 enrolment 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 analysed according to the suggestions of nutritionists and epidemiologists working on the study (Table 2).

TABLE 1

Descriptive statistics

Centre	Women	Age	BMI	Cases	Controls	Total
	%	Mean (SD)	Mean (SD)			
North-East of France	100	56.4 (6.7)	23.5 (3.5)	83	1525	1608
North-West of France	100	54.5 (7.1)	22.9 (3.3)	35	538	573
South of France	100	56.1 (5.9)	23.1 (3.4)	47	853	900
South coast of France	100	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	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	60.4 (6.0)	25.7 (4.0)	135	783	918
Heidelberg	28	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 & East of Norway	100	51.8 (3.8)	24.6 (4.0)	29	970	999
North & West of Norway	100	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	55.57	49.021
Root Vegetables	27.35	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	21.77
Fruits	218.786	171.468
Nuts and Seeds	3.189	8.04
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	0.04	0.553
Chocolate + Confectionery + Syrup	13.902	20.901
Coffee	452.99	400.388
Sauces	22.872	22.101

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*-th dietary exposure. Such matrices are usually named tables of nutrient composition and may vary between countries and centres. As a result, they can be generally regarded as centre-specific information which can contribute to explain the variability in the dietary effects among the centres. According to the dietary items involved into the analysis, we select a list including the most considerable nutrients (Table 3).

TABLE 3

Nutrients and corresponding unit of measurement

Nutrients	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

3. METHODS: THE HIERARCHICAL MODEL FOR CORRELATED EXPOSURES AND NESTED DATA

Let's consider the multicentric case-control study introduced above, where the presence/absence of colon-rectum cancer is denoted by the disease indicator y_{ij} (y_{ij} = 1 for cases, y_{ij} = 0 for control units) for the *i*-th subject in centre *j* and the food intakes are summarized by the symbol x_{ijk} for each dietary exposure *k*.

A conventional analysis would use the method of Maximum Likelihood (ML) to estimate the effects β_{jk} of the dietary exposures in centre *j* according to the following logistic regression

$$\text{logit}[E(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{w}_{ij}] = \alpha_j + \sum_{k=1}^{K} \beta_{jk} x_{ijk} + \sum_{p=1}^{P} \gamma_{jp} w_{ijp} \tag{1}$$

where a set of potential confounders (i.e., 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) denoted by w_{ijp} are embedded in the model specification in order to control for their effects γ_{jp} (Rothman *et al.*, 2008).

When the data structure is hierarchical with subjects (level 1) nested in clusters/centres (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 (Diez-Roux, 2000). Therefore, unless some different statistical models are introduced, we should be forced to carry out several ordinary logistic regressions, one for each centre of enrolment *j*.

A proper method to manage correlations among the responses is represented by the hierarchical modelling for nested data or, more simply, multilevel methods (Hox, 1995; Snijders and Bosker, 1999; Leyland and Goldstein, 2001; Raudenbush and Bryk, 2002). This allows to unify the analysis across the centres, partition the variability at both levels and choose the parameters to be random among the groups. Under this perspective, the logistic regression (1) represents a level-1 model which is part of a unique analysis. In our application, we further assume that the intercepts α_j and the dietary coefficients β_{jk} may vary across the centres. Conversely, the effects of confounders can be reasonably assumed to be the same in all the groups, i.e.

$$\text{logit}[E(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{w}_{ij}] = \alpha_j + \sum_{k=1}^{K} \beta_{jk} x_{ijk} + \sum_{p=1}^{P} \gamma_p w_{ijp} \tag{2}$$

According to multilevel approach, at level 2 the random intercepts are modelled to control for the multilevel structure of the data (i.e., the within-centre dependence). Here, no additional covariates for the centres are available. Therefore, we consider an empty model for the intercepts which splits the random parameter into a common effect ψ_0 and a residual term u_j, yielding the differences among the centres:

$$\alpha_j = \psi_0 + u_j \tag{3}$$

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

As far as the random slopes, which represent the key objective of the analysis, are concerned, we need to control for interactions and collinearity among the large number of dietary items which are involved. Conversely, inference results can be invalidated (Morris, 1983; Greenland, 1992 and 1993; Greenland, 1997). Moreover, the sub-samples generated by the nested data structure can be too sparse and small to yield accurate estimates. A level-2 regression for slopes in the multilevel sense can only partially model the associations among the exposures. A more appropriate solution should include some information mediating the exposures in order to explain (part of) their associations. Therefore, the data on constituents for each food and centre (denoted by z_{jkq} for q-th nutrient) are used to develop a level-2 regression model for the dietary coefficients:

$$\beta_{jk} = \pi_0 + \sum_{q=1}^{Q} \pi_q z_{jkq} + \delta_{jk} \tag{4}$$

where we assume that the effects of the food exposures on the colon-rectum cancer are partially mediated by the effects of nutrients π_q (with q=1, ..., Q). In this application, we can suitably suppose that the values of the level-2 residual variances related to model (4) will be small for all the food effects. Indeed, once the centre-specific information on nutrients are considered, we believe that the variability of dietary effects among the centres would be entirely explained. As a result, the residuals δ_{jk} can be assumed to hold the simple hypothesis of independence and normal distribution with null means and constant variances, denoted by τ^2, and to be further independent on u_j.

3.1. *The Bayesian perspective*

The hierarchy of models is completed by adopting a Bayesian perspective, where the other parameters are further considered as random with their own prior distribution to be specified. This is more precisely a BEB approach, as in equation (4) the prior distribution of the dietary effects is assigned by regressing on observed data (i.e. nutrient compositions).

The Bayesian approach would ensure more credible results with respect to those obtained by the frequentist methods, especially regarding the estimation of level-2 variance τ^2. Indeed, the frequentist EB approach (Witte *et al.*, 1998 and 2000) often yield null estimates for τ^2 leading to an extreme shrinkage estimation of the dietary effects toward the estimated prior means. This seems more likely to reflect a marginal likelihood for τ^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^2. Indeed, it represents the uncertainty about the residuals δ_{jk} and therefore also about the estimation of β_{jk} after incorporating the level-2 information. In particular, if τ^2 tends to ∞ the hierarchical model and the conventional logistic regression come to the same results with respect to β_{jk}. On the contrary, if $\tau^2 = 0$, then the residuals δ_{jk} result to be null, meaning that we implicitly assume the absence of any effects of dietary items beyond those of nutrients.

Other works suggest the SB approach as a good and easy strategy to tackle the problem of null estimation of the level-2 variance by setting suitable values for τ^2 (Greenland, 1992; Witte *et al.*, 1998). Despite SB estimates appear to be better than EB ones when the sample sizes and the ratio of subjects to parameters are small, a great caution to overspecify these values is required, especially when either the sample size or number of parameters are large (Greenland, 1992).

Here, by adopting a Bayesian approach we are allowed to assign a fully reasonable distribution to the hyper parameter τ^2 by letting the data can contribute to its final estimation. We base on plausible ranges of variation for log normal random dietary effects. More specifically, in model (4) the log ORs (i.e. β_{jk}) are implicitly supposed to be normally distributed with means $\mu_{jk} = \pi_0 + \sum_{q=1}^{Q} \pi_q z_{jkq}$ and variance τ^2. Therefore, a priori, 90% of values of each β_{jk} are believed to lie in interval $\mu_{jk} \pm 1.645\tau$ and, as a consequence, $\beta_{jk}^{95\%}$-$\beta_{jk}^{5\%} = 2 \times 1.645 \times \tau = 3.29\tau$. We believe a 2-fold variation between the ORs for the upper and lower 5% of units is reasonable, that is $\beta_{jk}^{95\%}$-$\beta_{jk}^{5\%} = \log 2$. Hence, our prior guess on the standard deviation τ is $\log 2/3.29 \approx 0.21$, corresponding to a precision term τ^{-2} equal to 22.53.

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 5.63. Assuming both the hypothesis are consistent with the problem at hand, these are sufficient to 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, i.e.

$$\tau^{-2} \sim \text{Gamma}(5; 0.22) \tag{4}$$

For the other (hyper) parameters we assign proper and vague prior distributions (Gelman *et al.*, 2003).

The Bayesian analysis 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. 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.

4. RESULTS

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

We select some results, which are the most representative across the large number of estimates. Thus, in Tables 4 to 9, the ORs and their 95% Credibility (or simply Confidence for the ordinary regressions) Intervals (CI) are calculated according to food-specific values of unit increase which are the sample standard deviations reported 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 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 conversely identify the intakes of cabbages as a protective factor for colon-rectum cancer.

TABLE 4

Estimated ORs (95% CI) by centre: milk and milk beverages

Centre	Ordinary logistic regression	Hierarchical Bayesian model
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 & East 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

Estimated ORs (95% CI) by centre: fruits

Centre	Ordinary logistic regression	Hierarchical Bayesian model
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 & East 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

Estimated ORs (95% CI) by centre: processed meat

Centre	Ordinary logistic regression	Hierarchical Bayesian model
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 & East 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

Estimated ORs (95% CI) by centre: fish

Centre	Ordinary logistic regression	Hierarchical Bayesian model
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 & East 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

Estimated ORs (95% CI) by centre: legumes

Centre	Ordinary logistic regression	Hierarchical Bayesian model
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 & East of Norway	–	1.008 (0.876-1.159)
North & West of Norway	–	0.998 (0.863-1.151)

TABLE 9

Estimated ORs (95% CI) by centre: cabbages

Centre	Ordinary logistic regression	Hierarchical Bayesian model
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 & East of Norway	–	1.027 (0.903-1.162)
North & West of Norway	0.684 (0.161 2.910)	0.967 (0.844-1.108)

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 centres. 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 centres (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 centres. 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 centres as their levels of nutrients are more likely to be similar. Indeed, previous evidence (Roli, 2006) showed that 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.

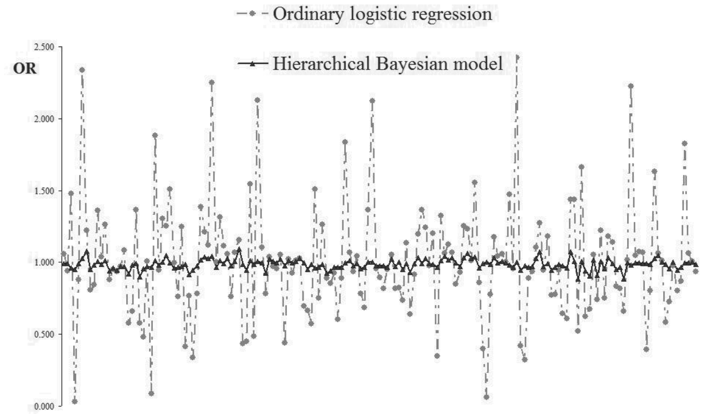

Figure 1 – Estimated ORs.

5. DISCUSSION AND CONCLUSIONS

We showed the advantages related to the use of hierarchical models under a BEB framework through the results from real data on a multicentric study aiming at evaluating multiple dietary effects on colon-rectum cancer. This method allows to exploit all the prior knowledge about the problem at hand, as well as making suitable assumptions which are arranged to form (hyper) prior probability distributions and can facilitate the task of estimation. As a result, the ordinary ML estimates of multiple dietary effects on colon-rectum cancer are improved, for each centre separately, 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 centres.

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 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. Moreover, the hierarchical approach provides the estimation of nutrients' effects, that may be alike useful from a nutritional point of view (e.g., for the formulation of a balanced diet).

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. However, the sample size limits the performance of BEB hierarchical model and the number of level-2 covariates to be embedded into the analysis. 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 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. Moreover, the hierarchical framework in multiple regression analysis can provide an alternative to conventional variable selection techniques (Gelman and Hill, 2007), allowing to retain all the variables in the analysis in order to be further evaluated whenever additional information would be available.

Department of Statistical Sciences
University of Bologna

GIULIA ROLI
PAOLA MONARI

ACKNOWLEDGEMENTS

This work is part of a PHD thesis, whose realization would never have been possible without the collaboration with the International Agency for Research on Cancer, Lyon, France. Therefore, we would like to express a deep acknowledgement to all the Nutrition and Hormones Group of IARC, who constantly invest their time and intellectual energy in maintenance and development of the EPIC project and supported this analysis with precious comments and suggestions.

REFERENCES

L. BERNARDINELLI, D. CLAYTON, C. PASCUTTO, C. MONTOMOLI, M. GHISLANDI, M. SONGINI, (1995), *Bayesian analysis of space-time variation in disease risk,* "Statistics in Medicine", 14, pp. 2433-2443.

J.F. JR BURGESS, C.L. CHRISTIANSEN, S.E. MICHALAK, C.N. MORRIS, (2000), *Medical profiling: improving standards and risk adjustments using hierarchical models*, "Journal of Health Economics", 19, pp. 291-309.

B. CARLIN, T. LOUIS, (1998), *Bayes and empirical Bayes methods for data analysis*, Chapman and Hall, New York.

C. CUBBIN, M.A. WINKLEBY, (2005), *Protective and harmful effects of neighborhood-level deprivation on individual-level health knowledge, behavior changes, and risk of coronary heart disease*, "American Journal of Epidemiology", 162, pp. 559-68.

J.J. DEELEY, D.V. LINDLEY, (1981), *Bayes Empirical Bayes*, "Journal of the American Statistical Association", 76, pp. 833-841.

A.V. DIEZ-ROUX, (2000), *Multilevel anlaysis in public health research*, "Annual Review of Public Health", 21, pp. 171-92.

A.V. DIEZ-ROUX, (2004), *The study of group-level factors in epidemiology: rethinking variables, study designs, and analytical approaches*, "Epidemiologic Reviews", 26, pp. 104-111.

A. GELMAN, J.B. CARLIN, H.S. STERN, D.B. RUBIN, (2003), *Bayesian Data Analysis*, 2nd edn., Chapman and Hall, New York.

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

H. GOLDSTEIN, (1999), *Multilevel statistical models*, John Wiley, New York.

P. GRAHAM, (2008), *Intelligent Smoothing Using Hierarchical Bayesian Models*, "Epidemiology", 19, pp. 493-495.

S. GREENLAND, (1992), *A semi-Bayes approach to the analysis of correlated multiple associations, with an application to an occipational cancer-mortality study*, "Statistics in Medicine", 11, pp. 219-230.

S. GREENLAND, (1993), *Methods for epidemiologic analysis of multiple exposures: a review and a comparative study of maximum-likelihood, preliminary testing and empirical Bayes regression*, "Statistics in Medicine", 12, pp. 717-736.

S. GREENLAND, (1997), *Second-stage least squares versus penalized quasi-likelihood for fitting hierarchical models in epidemiologic analysis*, "Statistics in Medicine", 16, pp. 515-526.

S. GREENLAND, (2000), *Principles of multilevel modelling,* "International Journal of Epidemiology", 29, pp. 158-167.

S. GREENLAND, (2006), *Bayesian perspectives for epidemiological research: I. Foundations and basic methods*, "International Journal of Epidemiology", 35, pp. 765-775.

S. GREENLAND, (2007), *Bayesian perspectives for epidemiological research: II. Regression analysis*, "International Journal of Epidemiology", 36, pp. 195-202.

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

A.B. LAWSON, (2001), *Disease map reconstruction,* "Statistics in Medicine", 20, pp. 2183-2204.

A. LEYLAND, H. GOLDSTEIN, (2001), *Multilevel modelling of health statistics*, John Wiley, New York.

R.F. MACLEHOSE, D.B. DUNSON, A.H. HERRING, J.A. HOPPIN, (2007), *Bayesian methods for highly correlated exposure data*, "Epidemiology", 18, pp. 199-207.

J. MARITZ, T. LWIN, (1989), *Empirical Bayes Methods*, Chapman and Hall, New York.

C. MORRIS, (1983), *Parametric empirical Bayes; theory and applcations* (with discussion), "Journal of the American Statistical Association", 178, pp. 47-65.

S.W. RAUDENBUSH, A.S. BRYK, (2002), *Hierarchical Linear Models - Application and data analysis methods,* Second edition, Sage Publications, London.

E. RIBOLI, R. KAAKS, (1997), *The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition*, "International Journal of Epidemiology", 26(1), pp. 6-14.

E. RIBOLI, K.J. HUNT, N. SLIMANI, *ET AL*., (2002), *European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection*, "Public Health Nutrition", 5(6B), pp. 1113-24.

G. ROLI, (2006), *Hierarchical logistic regression in a multicentric study of multiple dietary effects on a disease outcome: a fully Bayesian approach.* PHD thesis.

K.J. ROTHMAN, S. GREENLAND, T.L. LASH, (2008), *Modern epidemiology*, 3rd ed., Lippincott-Williams-Wilkins, Philadelphia.

T. SNIJDERS, R. BOSKER, (1999), *Multilevel analysis: an introduction to basic and advanced multilevel modeling*, Sage Publications, London.

D. SPIEGELHALTER, A. THOMAS, N. BEST, D. LUNN, (2003), *WinBUGS User Manual*, Version 1.4.

D.C. THOMAS, J. SIEMIATYCKI, R. DEWAR, J. ROBINS, M. GOLDBERG, B.G. ARMSRTONG, (1985), *The problem of multiple inference in studies designed to generate hypotheses*, "American Journal of Epidemiology", 122, pp. 1080-1095.

J. WITTE, S. GREENLAND, R. HAILE, C. BIRD, (1994), *Hierarchical regression analysis applied to a study of multiple dietray exposures and breast cancer*, "Epidemiology", 5 (6), pp. 612-621.

J. WITTE, S. GREENLAND, L.L. KIM, (1998), *Software for Hierarchical Modeling of Epidemiological Data*, "Epidemiology", 9(5), pp. 563-566.

J. WITTE, S. GREENLAND, L.L. KIM, L. ARAB, (2000), *Multilevel Modeling in Epidemiology with GLIMMIX*, "Epidemiology", 11(6), pp. 684-688.

SUMMARY

Improving the estimation of multiple correlated dietary effects on colon-rectum cancer in multicentric studies: a hierarchical Bayesian approach

The paper deals with the analysis of the effects of multiple exposures on the occurrence of a disease in observational case-control studies. We consider the case of multilevel data, with subjects nested in spatial clusters. As a result, we often face problems of small and sparse data, along with correlations among the exposures and the observations, which both invalidate the results from the ordinary analyses. A hierarchical Bayesian model is here proposed to manage the within-cluster dependence and the correlation among the exposures. We assign prior distributions on the crucial parameters by exploiting additional information at different levels and by making suitable assumptions according to the problem at hand. The model is conceived to be applied to a real multi-centric study aiming at investigating the association of dietary exposures with colon-rectum cancer occurrence. Compared with results obtained with conventional regressions, the hierarchical Bayesian model is shown to yield great gains in terms of more consistent and less biased estimates. Thanks to its flexibility, this approach represents a powerful statistical tool to be adopted in a wide range of applications. Moreover, the specification of more realistic priors may facilitate and extend the use of Bayesian solutions in the epidemiological field.