Using Hierarchical Modeling in Genetic Association Studies with Multiple Markers: Application to a Case-Control Study of Bladder Cancer

Rayjean J. Hung,[1,2] Paul Brennan,[1] Christian Malaveille,[1] Stefano Porru,[3] Francesco Donato,[4] Paolo Boffetta,[1,5] and John S. Witte[1,6]

[1]International Agency for Research on Cancer, Lyon, France; [2]Department of Epidemiology, University of California at Los Angeles, Los Angeles, California; Institutes of [3]Occupational Health and [4]Hygiene, University of Brescia, Brescia, Italy; [5]German Cancer Research Center, Heidelberg, Germany; and [6]Department of Epidemiology and Biostatistics, University of California at San Francisco, San Francisco, California

Abstract

Background: Genetic association studies are generating much information, usually in the form of single nucleotide polymorphisms in candidate genes. Analyzing such data is challenging, and raises issues of multiple comparisons and potential false-positive associations. Using data from a case-control study of bladder cancer, we showed how to use hierarchical modeling in genetic epidemiologic studies with multiple markers to control overestimation of effects and potential false-positive associations. Methods: The data were first analyzed with the conventional approach of estimating each main effect individually. We subsequently employed hierarchical modeling by adding a second stage (prior) model that incorporated information on the potential function of the genes. We used an empirical-Bayes approach, estimating the residual effects of the genes from the data. When the residual effect was set to zero, we instead used a semi-Bayes approach, in which they were pre-specified. We also explored the impact of using different second-stage design matrices. Finally, we used two approaches for assessing gene-environment interactions. The first approach added product terms into the first-stage model. The second approach used three indicators for subjects exposed to gene-only, environment-only, and both genetic and environmental factors. Results: By pre-specifying the prior second-stage covariates, the estimates were shrunk to the mean of each pathway. The conventional model detected a number of positive associations, which were reduced with the hierarchical model. For example, the odds ratio for myeloperoxidase *(G/G, G/A)* genotype changed from 3.17 [95% confidence interval (CI), 1.32-7.59] to 1.64 (95% CI, 0.81-3.34). A similar phenomenon was observed for the gene-environment interactions. The odds ratio for the gene-environment interaction between tobacco smoking and *N*-acetyltransferase 1 fast genotype was 2.74 (95% CI, 0.68-11.0) from the conventional analysis and 1.24 (95% CI, 0.80-1.93) from the hierarchical model. Conclusion: Adding a second-stage hierarchical modeling can reduce the likelihood of false positive via shrinkage toward the prior mean, improve the risk estimation by increasing the precision, and, therefore, represents an alternative to conventional methods for genetic association studies. (Cancer Epidemiol Biomarkers Prev 2004;13(6):1013–21)

Background

Cancer is a complex disease involving multiple genetic and environmental risk factors. To clarify the contribution of genetic factors and decipher the relationship between genes, environment, and cancer, association studies are generating much genetic information. This has often taken the form of studying single nucleotide polymorphisms in different genes. Analyzing such data is challenging, and raises the issues of multiple comparisons and potential false-positive associations (1).

The common conventional analytic approaches currently in practice, such as (1) treating all exposures independently (i.e., estimating the parameters of interest one at a time), (2) including all exposures in a single model, and (3) using algebraic selection procedures, such as stepwise regression, each have limitations (2). The first approach ignores the biological relevance among the exposures. The second approach often results in overparameterized models, and the third approach selects the parameters arbitrarily based on ''statistical significance'' and can result in biased point and variance estimates (3). Furthermore, none of these approaches properly address the issues of multiple comparisons and potential false-positive associations. In contrast, hierarchical modeling may provide an avenue for addressing these issues in genetic association studies. In particular, such studies generally collect much information on interrelated genetic variables, which can be incorporated into a hierarchical model.

Received 11/4/03; revised 2/3/04; accepted 2/9/04.

Grant support: This work was supported by the NCI R01 grant CA 092039-01A2.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: R. Hung worked on this study under the tenure of a Special Training Award from the International Agency for Research on Cancer.

Requests for reprints: Rayjean J. Hung, Internal Agency for Research on Cancer Epidemiology, 150 cours Albert-Thomas, 69008 Lyon, France. Fax: 33-4-72738320. E-mail: hung@iarc.fr

A false-positive association is suspected to occur when the point estimate is far away from other estimates, and unstable (i.e., with a large SE; ref. 4). This problem can be reduced by ''correcting'' the overestimation of the observed variance with the goal of estimating the ''true'' variance. The true variance (V_T) can be approximated by subtracting the mean of the variance of the individual estimates (V_M) from the observed sample variance of the log relative risk estimate (V_O; ref. 4). Because the variance should be nonnegative, the estimate of V_T is the maximum of $V_O - V_M$ and zero.

$$V_T \cong V_O - V_M; \hat{V}_T = \max(\hat{V}_O - \hat{V}_M, 0) \qquad (A)$$

Hierarchical methods can be used to adjust the observed relative risk estimates by the estimated V_T. Typically this adjustment pulls the outlying relative risk estimates toward the geometric mean of the ensemble and leads to a narrower confidence interval (CI; because V_T is expected to be smaller than V_O; ref. 4). If the mean of the estimates are reasonably close to the mean of the true values, one should expect that the shrinkage to the mean of the estimates would decrease the total estimation errors [i.e., expected squared error (ESE) = bias^2 + SD^2].

There are three assumptions when using this adjustment. First, there is no systematic bias in the conventional estimates that threatens the validity of the ''shrinkage to the mean'' (i.e., if the systematic bias is present, the estimates would be shrunk toward the ''wrong mean'' after adjustment). Second, the true values and random errors in each ensemble are both roughly normally distributed. And third, the true values of the relative risk within each ensemble are exchangeable. To satisfy the last assumption, the parameters must be grouped in a manner that minimizes the ESE (4). For example, consider a study of smoking, genetic polymorphisms, and lung cancer. If smoking (a strong risk factor) is included in one ensemble with multiple low penetrance genes, large but unstable estimates for a few genetic polymorphisms (which are more likely to be produced by sampling error) would not be pulled down, as would happen *without* smoking in the same ensemble.

The adjustment by shrinkage estimation can be undertaken with hierarchical modeling. This approach unifies the seemly separate concepts of frequentist and Bayesian analysis, and represents a valid alternative to a conventional analysis (5). It adds in two or more levels in the model to specify the relations among study variables. General properties of hierarchical modeling have been reviewed elsewhere (6, 7).

While numerous authors have advocated the advantages of using hierarchical modeling, they remain little used in the health sciences (7-10). Few other authors used hierarchical modeling to analyze gene-gene or gene-environment interactions (11, 12). Here, we further demonstrate how hierarchical modeling can be applied to genetic epidemiologic studies, including several low penetrance genes and addressing gene-environment interactions. The present study is different from the previous studies in that we employed a data set with multiple markers in different genes, and also provided a possible construction of the prior multivariate second-stage model for such a data set. To evaluate the performance of our model, we contrasted the results based on hierarchical modeling to those from conventional analyses. We used the data from a case-control study of bladder cancer, which is representative of current genetic epidemiologic studies, containing a modest sample size and multiple markers.

Materials and Methods

The Bladder Cancer Case-Control Study. A hospital-based case control study of bladder cancer was conducted in Brescia, Northern Italy. Two hundred one male incident cases and 214 male controls with benign urological conditions were recruited from 1997 to 2000, frequency matched on age, period of recruitment, and hospital. Lifestyle and occupational information was collected by interview with a questionnaire. Occupational exposures to polycyclic aromatic hydrocarbons and aromatic amines were blindly coded by occupational physicians. Genotyping of glutathione *S*-transferase M1 (*GSTM1*) null, *GSTT1* null, *GSTP1 I105V*, *N*-acetyltransferase 1 (*NAT1*) fast, *NAT2* slow, cytochrome P450 1B1 (*CYP1B1*) *V432L*, sulfotransferase 1A1 (*SULT1A1*) *R213H*, myeloperoxidase (*MPO*) *G-463A*, catechol-*O*-methyltransferase (*COMT*) *V108M*, manganese superoxide dismutase (*MnSOD*) *A-9V*, NAD(P)H:quinone oxidoreductase (*NQO1*) *P187S*, X-ray repair cross-complementing group 1 (*XRCC1*) *R399Q*, *XRCC3 T241M*, and xeroderma pigmentosum complementation group (*XPD*) *K751Q* polymorphisms was done with PCR-RFLP methods. Detailed results from this study have been reported elsewhere (13-15).

Hierarchical Modeling Overview. Hierarchical modeling extends and complements conventional analyses. In a two-stage model, one can use a conventional logistic regression as the first stage:

$$ln\left(\frac{P}{1-P}\right) = \alpha + X_j\beta_j + W\gamma \qquad j = 1, \ldots, n \qquad (B)$$

The coefficients β_j represent the effects of X_j, the exposures of interest. In our example, each coefficient represents the effect of particular genetic marker on cancer risk on a log scale. W is the matrix of the covariates (e.g., age) and γ is the vector of covariate coefficients.

Hierarchical modeling adds a second stage to the conventional model to try and improve the estimation of the β_j. For example, one could use the following linear model

$$\beta_j = z_j\pi + \delta_j \qquad \beta = Z\pi + \delta \qquad (C)$$

$$\delta \sim N(0, \tau^2) \qquad (D)$$

where z_j is a row vector of p prior covariates, and π is a column vector of p prior coefficients. Z is the n-by-p matrix of second-stage covariates, with rows z_j. The Z matrix is a key component of the hierarchical approach: it is defined by the investigators to reflect the similarities between the first-stage factors. For example, in a

case-control study of multiple dietary exposures and breast cancer risk (8), the Z matrix was based on the nutritional contents of the food items. In an analysis of genetic factors, it was based on physical distance between polymorphisms (16). In a study of gene-environment interactions in polyp formation, the Z matrix incorporated the conversion rates for heterocyclic amines (12).

In Eq. C, π is the column vector of linear effects of the second-stage covariates (Z) on β_j, and δ_j are the residual effects of item j (e.g., specific genetic factors). Residual effects can arise from interactions among second-stage covariates, or from the covariates not included in the Z matrix, and are assumed to have mean of zero and variance of τ^2. The τ can either be estimated from the data [empirical-Bayes (EB) approach] using a variance model based on the concept outlined above (Eq. A; refs. 2, 4, 17), or pre-specified by the investigators using background information [semi-Bayes (SB) approach] (7, 17). τ^2 reflects the range of the potential residual effect that remain after accounting for all first- and second-level covariates, and can be viewed as an approximation of V_T.

Substituting Eq. C into B gives a mixed model (Eq. E), which contains fixed coefficients (π and γ) and random coefficients δ.

$$ln\left(\frac{P}{1-P}\right) = \alpha + X(Z\pi + \delta) + W\gamma$$
$$= \alpha + XZ\pi + X\delta + W\gamma \qquad \text{(E)}$$

Application of Hierarchical Modeling to the Case-Control Study. For comparison's sake, the data were initially analyzed with the conventional approaches using the following two logistic regression models: treating all the genetic factors independently and constructing one-at-a-time models; and including all genetic factors in one model.

Our first hierarchical model was constructed under the assumption that all genes are exchangeable, with the EB approach (allowing τ^2 to be estimated from the data). Because the assumption of full exchangeability is unlikely to reflect the underlying biological relations among genes, we further estimated the main effects of genetic polymorphisms by using a Z matrix to specify the similarity between genetic factors in terms of their function.

Construction of *Z* Matrix. The genetic factors were broadly classified into four main pathways related to bladder carcinogenesis, based on the biological functions of the genes: (1) reduced carcinogen detoxification; (2) enhanced procarcinogen activation; (3) increased oxidative stress; and (4) decreased DNA repair capacity. The risk genotype of each gene is designated based on the functional significance of the polymorphism. For each pathway, a score was assigned to each genetic polymorphism according to biological function of the gene and the functional significance of its polymorphism. A score of 0 was assigned to the genetic polymorphism if the gene is not involved in such pathway (the coefficient of the factor was expected to be zero). For example, MPO was assigned a score of 0 in the pathway of reduced detoxification and DNA repair, because it is not expected to have roles in such pathways; whereas a score of 1 was assigned if the gene is involved in such pathway, and the functional significances of the risk genotypes correspond to the pathway that contribute to the risk of cancer (the factor was expected to have a positive coefficient). Take XRCC1, for example, it has an important role in repair of single strand break. *399Gln* allele was associated with higher levels of aflatoxin B1-DNA adduct and higher bleomycin sensitivity; therefore, the designated risk genotype of *XRCC1* is *Arg/Gln*, *Gln/Gln*, and a score of 1 was assigned in the pathway of reduced DNA repair (18, 19).

The genes assigned to the same pathway were assumed to be exchangeable, that is, for their effects to arise from a common distribution. The four pathways were not mutually exclusive. For example, MPO activates procarcinogens in tobacco smoke, such as benzo[*a*]pyrene, through the release of reactive oxygen species (21-23). *−463A* allele is associated with reduced mRNA expression and its transcription activity is about 25 times lower than *G* allele *in vitro* (23). A score of 1 was assigned to *MPO G/G*, *G/A* in the pathway of enhanced activation and increased oxidative stress due to MPO's role in both pathways. Another example, *NQO1 Pro187Ser* variant genotypes (*Pro/Ser, Ser/Ser*) was given a score of 1 in both pathways of reduced carcinogen detoxification and enhanced procarcinogen activation, because NQO1 converts the highly genotoxic benzoquinine to the less toxic hydroxy metabolites (24), whereas it is also involved in bioactivation of some procarcinogens, such as 4-nitroquinoline-1-oxide (25).

The basic Z matrix is reported in Table 1. In summary, the biological implication of this Z matrix is that *GSTs*, *NAT2*, *SULT1A1*, and *NQO1* play roles in carcinogen detoxification, and their designated risk genotypes reduce the detoxification; *NAT1*, *SULT1A1*, *CYP1B1*, *MPO*, and *NQO1* are involved in procarcinogen activation, and their designated risk genotypes enhance the activation; *CYP1B1*, *MPO*, *COMT*, *MnSOD*, and *NQO1* modulate the oxidative stress, and the risk genotypes reduce the oxidative stress; *XRCC1*, *XRCC3*, and *XPD* repair DNA via different mechanisms, and their risk genotypes decreased the DNA repair capacity.

We conducted sensitivity analyses by altering the Z matrix, based on the function of genes with possible mechanisms that are still open for discussion. Two variations of the Z matrices are also shown in Table 1. In Z-matrix variant 1, we added *COMT Val108Met* and *MnSOD Ala-9Val* in the pathway of enhanced procarcinogen activation, and removed *CYP1B1* from the pathway of increased oxidative stress. These changes are based on function of genes with possible mechanisms that are still open for discussion. For example, COMT and MnSOD are known to protect the cell from oxidative stress (26, 27). Their genetic variants are designated as risk genotypes because they are known to have lower enzyme activities or predicted to be less effective (28-30), and, therefore, contribute to increased oxidative stress (as in basic Z matrix). Because reactive oxygen species are often involved in the activation of the procarcinogens, genetic polymorphisms that increase oxidative stress might indirectly attribute to the enhanced activation of procarcinogens. We, therefore, added these two genetic polymorphisms into the pathway of enhanced activation in Z-matrix variant 1.

Table 1. *Z* matrices used in the hierarchical model

Gene	Single nucleotide polymorphism	High-risk genotype	Basic *Z* matrix			
			Reduced detoxification	Enhanced activation	Increased oxidative stress	Decreased DNA repair
GSTM1	—	null	1	0	0	0
GSTT1	—	null	1	0	0	0
*GSTP1**	*Ile105Val Ala114val*	≥one allele*C*	1	0	0	0
NAT2†	**5A*, **6A*, **7A*	Slow	1	0	0	0
NAT1‡	**10* and **11*	Fast	0	1	0	0
SULT1A1	*Arg213His*	*Arg/Arg*	1	1	0	0
CYP1B1	*Val432Leu*	*Val/Val*	0	1	1	0
MPO	*G-463A*	*G/G, G/A*	0	1	1	0
COMT	*Val108Met*	*Val/Met*, *Met/Met*	0	0	1	0
MnSOD	*Ala-9Val*	*Val/Val*	0	0	1	0
NQO1	*Pro187Ser*	*Pro/Ser*, *Ser/Ser*	1	1	1	0
XRCC1	*Arg399Gln*	*Arg/Gln*, *Gln/Gln*	0	0	0	1
XRCC3	*Thr241Met*	*Thr/Thr*	0	0	0	1
XPD	*Lys751Gln*	*Lys/Gln*, *Gln/Gln*	0	0	0	1

**GSTP1* allele which has both *105Val* and *114Val*.
†**5A*: *T341C*, *C481T*; **6A*: *C282T*, *G590A*; **7A*: *G857A*; Slow: at least two slow acetylator alleles.
‡**10*: *T1088A*, *C1095A*; **11*: *C-344T*, *A-40T*, *G445A*, *G459A*, *T640G*, △9 between nt 1095 and 1090, *C1095A*; Fast: **10* or **11* homo-heterozygous.

In *Z*-matrix variant 2, scores were allowed to range from −1 to 1. Genetic polymorphisms were assigned a score of −1 when hypothesizing that the prior risk genotypes have opposite effect in the given pathway (e.g., increased detoxification, instead of reduced detoxification). In other words, the factor was expected to have a negative coefficient. For example, NAT1 have been shown to be involved in both the activation and detoxification of aromatic amines. In general, NAT1 has a major role in the *O*-acetylation of *N*-hydroxy aromatic amines, and lead to the activation of aromatic amines (31). *NAT1 *10* and **11* are associated with faster enzyme activity and were, therefore, assigned a score of 1 in the pathway of enhanced activation. On the other hand, because NAT1 is also involved in detoxification process, it is possible that a higher enzyme activity conferred by **10* and **11* might be associated with an increased detoxification via *N*-acetylation, although the affinity for which is less. We, therefore, assigned a score of −1 in the pathway of reduced detoxification to indicate the opposite effect via same pathway.

We assumed that the basic *Z* matrix was the most comprehensive prior for our purposes, and the analyses of gene-environment interaction were based on the basic *Z* matrix. The details of *Z* matrix applied to gene-environment interaction is described in the following section.

We used an EB approach, which estimated τ^2 from the data. When this approach estimated τ^2 as being less than or equal to zero, we used a SB approach setting τ^2 equal to 0.05, a value similar to that estimated for main effects from the data in the EB approach. This assumes that the true odds ratio (OR) of each parameter would fall within a 2.4-fold range, that is, $\exp(\sqrt{0.05} \times 3.92) \cong 2.40$. To see the how sensitive the results were to τ^2, we increased the pre-specified value to 0.35, which assumes with 95% probability that the true ORs fall within a 10-fold range [i.e., $(\ln(10)/3.92)^2 \approx 0.35$].

Assessment of Gene-Environment Interactions. Three environmental agents were included in the analysis: tobacco smoking and occupational exposure to polycyclic aromatic hydrocarbons and to aromatic amines. These exposures are known to cause of bladder cancer (32). Two approaches were used for assessing the gene-environment interactions. The first approach added product terms to the first-stage model (see Eq. F).

$$\text{logit}(R) = \alpha + G\beta_g + E\beta_e + (G \times E)\beta_{ge} + W\gamma \qquad (F)$$

Attempting to include one environmental factor, with all the genetic factors and their product terms in a single model (a total of 29 parameters), resulted in the over-parameterization of the model, which could not be fitted. Therefore, we grouped genetic polymorphisms based on the biological functions of the genes (four groups: *GSTs*, oxidative stress, DNA repair, and others). The basic *Z* matrix was then divided into four corresponding groups. One column was added to the *Z* matrix to indicate the environmental exposure. All gene-environment product terms were assigned a score of 1 in the relevant genetic pathways and environmental exposure. The *Z* matrices used in gene-environment interaction analyses are available from the authors on request.

The second approach used three indicators for subjects exposed to high-risk genotype only, environmental factors only, and to both factors (see Eq. G below). Because there were three indicators for each pair of gene-environment interactions, analyzing the interaction by the group also resulted in an over-parameterized model. Therefore, gene-environment interactions were analyzed one pair at the time while using this approach. The *Z* matrix applied in this approach was also simplified to a three-by-two matrix with two columns indicating genetic and environmental factors. Consistent with the notation

Table 1. *Z* matrices used in the hierarchical model (Cont'd)

Gene	*Z*-matrix variant 1				*Z*-matrix variant 2			
	Reduced detoxification	Enhanced activation	Increased oxidative stress	Decreased DNA repair	Reduced detoxification	Enhanced activation	Increased oxidative stress	Decreased DNA repair
GSTM1	1	0	0	0	1	0	0	0
GSTT1	1	0	0	0	1	0	0	0
*GSTP1**	1	0	0	0	1	0	0	0
NAT2†	1	0	0	0	1	0	0	0
NAT1‡	0	1	0	0	−1	1	0	0
SULT1A1	1	1	0	0	−1	1	0	0
CYP1B1	0	1	0	0	0	1	1	0
MPO	0	1	1	0	0	1	1	0
COMT	0	1	1	0	0	0	1	0
MnSOD	0	1	1	0	0	0	1	0
NQO1	0	1	1	0	1	−1	1	0
XRCC1	0	0	0	1	0	0	0	1
XRCC3	0	0	0	1	0	0	0	1
XPD	0	0	0	1	0	0	0	1

in Eq. G: D_1 was assigned a score of 1 in genetic column, D_2 was assigned a score of 1 in environmental column, and D_3 was assigned a score of 1 in both columns.

$$\text{log}it(R) = \alpha + D_1\beta_{g-only} + D_2\beta_{e-only} + D_3\beta_{g\&e} + W\gamma \quad (G)$$

Consistent with the analyses for the genetic main effects, in gene-environment interaction analyses, we first used an EB approach to estimate τ^2 from the data. When this approach estimated τ^2 as being less than or equal to zero, we set τ^2 equal to 0.05, a value similar to that estimated from the data in the EB approach. To see how sensitive the results were to τ^2, we increased the pre-specified value to 0.35. All statistical analyses were conducted with SAS IML code in conjunction with GLIMMIX macro. An example program is available at URL http://darwin.cwru.edu/~witte/hm.html (17).

Results

Table 2 shows the results of the analyses of genetic main effects, comparing the conventional and the hierarchical analyses. In the conventional analysis, 5 of the 14 high-risk alleles included in the analysis were associated with a significantly increased risk of bladder cancer (one risk estimate was no longer significant in the analysis including all parameters at once). When assuming that all genetic factors are exchangeable (hierarchical model 1), all risk estimates were shrunk to a single mean. By pre-specifying the prior second-stage covariates (hierarchical model 2), the estimates were shrunk to the mean of each pathway. Most of the time, the CIs were narrower from hierarchical modeling compared with the conventional estimates, that is, the precision of the risk estimates was increased. This feature was observed across all the results. Extreme but unstable values experienced the greatest shrinkage. For example, the OR of *MPO G/G*, *G/A* genotypes was reduced from 3.17 (95% CI, 1.32-7.59) to 1.64 (95% CI, 0.81-3.34). Note that using a hierarchical model does not always shrink the estimates toward the null; this depends on the *Z* matrix. For example, the OR for *NQO1 Pro187Ser* polymorphism increased from 1.33 (95% CI, 0.89-1.97, conventional analysis) to 1.41 (95% CI, 0.98-2.03, hierarchical model 2).

Slight alterations of the *Z* matrix did not have major influence on the risk estimates (results based on hierarchical models 3 and 4 in Table 2). When τ^2 was pre-specified, the larger the pre-specified value, the lesser the "shrinkage" (results based on hierarchical models 5 and 6 in Table 2, and hierarchical models 1 and 2 in Tables 3 and 4).

Table 3 shows the results of gene-environment interaction analyses based on the product-term approach. For focus and brevity, we do not present the entire set of 29 × 3 = 87 estimates, but highlight only the most illustrative findings. The interaction OR for smoking and *NAT1* fast genotype was 2.74 (95% CI, 0.68-11.0) from the conventional analysis, and 1.24 (95% CI, 0.80-1.93) from the hierarchical modeling. The point estimate of the product term of between *GSTM1* null genotype and smoking did not change much: however, the CI narrowed considerably (from 0.43-5.14 to 0.74-3.83). Similar results were observed in the interaction analyses between genetic polymorphisms and exposure to aromatic amines. For example, in multiplicative terms, the hierarchical CI of the *NAT2*-aromatic amines product term was almost 10-fold narrower than that of the conventional analysis (i.e., 2.88/0.90 = 3.20 versus 10.1/0.34 = 29.7), indicating a dramatic increase in precision.

Table 2. OR and 95% CI of main effects of genetic polymorphisms from conventional analyses and hierarchical modeling

Gene	Allele at risk	Case no.	Control no.	Conventional (a)		Conventional (b)	
				OR	(95% CI)	OR	(95% CI)
GSTM1	null	132	112	1.74	(1.17-2.58)	1.61	(1.06-2.45)
GSTT1	null	43	33	1.52	(0.92-2.51)	1.57	(0.92-2.68)
GSTP1 105-114	allele with both variants	21	24	0.92	(0.49-1.71)	1.04	(0.54-2.01)
NAT2	slow	123	111	1.48	(1.00-2.19)	1.28	(0.82-1.97)
NAT1	fast	80	97	0.79	(0.53-1.17)	0.92	(0.59-1.43)
SULT1A1	*Arg/Arg*	121	116	1.29	(0.87-1.91)	1.27	(0.84-1.91)
CYP1B1	*Val/Val*	31	36	0.90	(0.53-1.52)	0.95	(0.55-1.66)
MPO	*G/G, G/A*	194	192	3.17	(1.32-7.59)	3.02	(1.21-7.53)
COMT	*Val/Met, Met/Met*	139	157	0.80	(0.52-1.23)	0.86	(0.55-1.36)
MnSOD	*Val/Val*	68	45	1.91	(1.23-2.96)	1.69	(1.06-2.70)
NQO1	*Pro/Ser, Ser/Ser*	88	79	1.33	(0.89-1.97)	1.36	(0.90-2.06)
XRCC1	*Arg/Gln, Gln/Gln*	108	122	0.87	(0.59-1.28)	0.90	(0.59-1.35)
XRCC3	*Thr/Thr*	89	71	1.60	(1.07-2.38)	1.68	(1.10-2.55)
XPD	*Lys/Gln, Gln/Gln*	122	134	0.92	(0.62-1.37)	1.04	(0.68-1.59)
Estimated τ^2							

NOTE: (a) Model parameters of interest included one at a time; (b) all parameters included in one model. (1) No prior Z matrix and no prior τ^2; (2) basic Z matrix and no prior τ^2; (3) Z matrix variant 1 and no prior τ^2; (4) Z-matrix variant 2 and no prior τ^2; (5) basic Z matrix and prior $\tau^2 = 0.05$; (6) basic Z matrix and prior $\tau^2 = 0.35$.
Abbreviation: OR, odds ratio adjusted for age.

Table 4 shows the noteworthy results of gene-environment interaction based on the indicator-term approach. Similar to the results in Table 3, the hierarchical estimates were more stable, with the amount of shrinkage depending on the value of τ^2. For example, when looking at the interaction between the *GSTT1* null genotype and exposure to aromatic amines, the conventional analysis resulted in an unstable OR that was likely overestimated and suggested the presence of interaction. The hierarchical modeling, on the other hand, gave more stable estimates and suggested no interaction.

Discussion

Genetic association studies with multiple genetic markers are susceptible to false-positive associations. We showed that hierarchical modeling may improve the analyses of genetic epidemiologic data by increasing the precision of the risk estimates (e.g., as evidenced by narrower CIs), and reducing the likelihood of false positive via shrinkage toward the prior mean. For example, when looking at *GSTM1* and smoking, the width of the CI, in multiplicative terms, was less than half of the width of conventional method (3.83/0.74 = 5.18, versus 5.14/0.43 = 11.95). This reduction in CI width is akin to increasing the number of cases in the present study 4-fold (33). Moreover, hierarchical modeling gives estimates that are closer to the prior expectation. On the basis of the literature and our biological understanding of carcinogenesis, one would expect that common genetic polymorphisms confer a modest effect on cancer risk. For example, one would expect a common polymorphism as MPO G-870A to confer a relative risk of less than 2 in the present study. The results from conventional analyses exceeded the prior, and the results from the hierarchical modeling were effectively adjusted toward the prior. As a consequence, hierarchical modeling reduces the likelihood of false-positive results. Finally, this approach allows for residual effects that are not accounted for in the fixed first- and second-stage coefficients, and avoids numerous independent single-inference testing (7). In our study, these benefits of hierarchical modeling were apparent for both the main effects and interaction analyses.

Potential false-positive results are characterized by increased imprecision in effect estimation, which results in less weight in the hierarchical models, and are, therefore, more liable to the ''shrinkage'' effect of hierarchical modeling than true positive results (4). Although hierarchical modeling may also increase the potential for false-negative findings, it should be noted that even initial positive findings that are later replicated in subsequent studies are also usually overestimated (34). Thus, the shrinkage effect typically seen in hierarchical modeling is also relevant for true positive effects.

Other authors had presented the use of hierarchical models for looking at gene-gene or gene-environment interactions. Aragaki et al. (12) studied the interaction between diet and a single candidate gene *NAT2*, using a different data structure than the one presented here. We had similar data as in De Roos et al. (11), though their second-stage model assumed that all effects (i.e., main genetic, environmental, and when modeled, their interactions) should be shrunk toward a single prior mean. However, this may not hold in practice. Therefore, in the present report, we extended the previous work by using a second-stage (Z) design matrix that reflects the similarity of different genetic factors in a more detailed manner. Furthermore, in our interaction analysis, we provided results from both indicator-term and product-term approaches. In the product-term approach, we went beyond looking at one interaction at a time. In addition, we undertook a sensitivity analysis of the design of the Z matrix and the pre-specified τ^2.

There are a number of assumptions underlying the use of hierarchical models; when these are violated, the resulting estimates can be worse than obtained with a conventional analysis (35). While hierarchical modeling

Table 2. OR and 95% CI of main effects of genetic polymorphisms from conventional analyses and hierarchical modeling (Cont'd)

Gene	Hierarchical (1)		Hierarchical (2)		Hierarchical (3)		Hierarchical (4)		Hierarchical (5)		Hierarchical (6)	
	OR	(95% CI)	OR	(95% CI)	OR	(95% CI)	OR	(95% CI)	OR	(95% CI)	OR	(95% CI)
GSTM1	1.43	(1.02-1.99)	1.49	(1.05-2.13)	1.51	(1.06-2.14)	1.51	(1.05-2.16)	1.47	(1.06-2.03)	1.57	(1.06-2.33)
GSTT1	1.35	(0.92-1.99)	1.43	(0.94-2.17)	1.45	(0.96-2.20)	1.44	(0.94-2.20)	1.39	(0.96-2.02)	1.50	(0.92-2.46)
GSTP1 105-114	1.13	(0.74-1.70)	1.16	(0.72-1.86)	1.20	(0.75-1.94)	1.18	(0.72-1.92)	1.19	(0.79-1.79)	1.07	(0.60-1.92)
NAT2	1.29	(0.93-1.77)	1.30	(0.92-1.85)	1.33	(0.93-1.90)	1.31	(0.92-1.88)	1.31	(0.94-1.81)	1.29	(0.85-1.94)
NAT1	1.02	(0.73-1.43)	0.90	(0.62-1.32)	0.94	(0.66-1.36)	0.91	(0.64-1.31)	0.90	(0.63-1.29)	0.91	(0.60-1.39)
SULT1A1	1.24	(0.90-1.69)	1.24	(0.86-1.80)	1.15	(0.80-1.66)	1.14	(0.79-1.65)	1.24	(0.87-1.77)	1.26	(0.85-1.88)
CYP1B1	1.08	(0.74-1.60)	1.05	(0.67-1.65)	0.98	(0.64-1.51)	1.20	(0.72-1.99)	1.08	(0.72-1.62)	1.01	(0.60-1.68)
MPO	1.58	(0.87-2.88)	1.64	(0.81-3.34)	1.80	(0.94-3.45)	1.97	(0.99-3.92)	1.41	(0.86-2.31)	2.19	(1.03-4.69)
COMT	1.01	(0.71-1.44)	1.00	(0.66-1.52)	1.07	(0.70-1.64)	0.98	(0.66-1.46)	1.04	(0.72-1.51)	0.91	(0.59-1.40)
MnSOD	1.47	(1.02-2.10)	1.54	(1.02-2.32)	1.58	(1.07-2.33)	1.51	(1.00-2.26)	1.50	(1.03-2.20)	1.67	(1.07-2.60)
NQO1	1.28	(0.93-1.77)	1.41	(0.98-2.03)	1.36	(0.96-1.93)	1.30	(0.89-1.90)	1.42	(1.00-2.01)	1.38	(0.93-2.06)
XRCC1	1.02	(0.73-1.42)	0.96	(0.66-1.38)	0.97	(0.67-1.40)	0.96	(0.66-1.39)	0.97	(0.69-1.38)	0.91	(0.61-1.36)
XRCC3	1.43	(1.01-2.02)	1.41	(0.94-2.12)	1.39	(0.93-2.09)	1.43	(0.95-2.14)	1.35	(0.95-1.92)	1.59	(1.06-2.37)
XPD	1.08	(0.79-1.49)	1.03	(0.71-1.49)	1.04	(0.72-1.50)	1.03	(0.71-1.50)	1.03	(0.72-1.47)	1.03	(0.68-1.55)
Estimated τ^2	0.038		0.05		0.04		0.05					

attempts to improve estimation accuracy, it may sometimes do so at the expense of increasing the bias (6). Recall that estimation accuracy reflects both validity (systematic error) and precision (random error). In the presence of systematic bias, and when the mean of the estimates is not reasonably close to the true mean, the hierarchical estimates would be pulled to the ''wrong'' mean. Furthermore, in the case when the second-stage

Table 3. Selected OR and 95% CI of gene-environment interaction with product-term approach from conventional analyses and hierarchical modeling

	Conventional		Hierarchical (1)			Hierarchical (2)	
	OR	(95% CI)	OR	(95% CI)	Estimated τ^2	OR	(95% CI)
GSTM1 null	1.28	(0.40-4.05)	1.08	(0.49-2.37)	0.03	1.05	(0.54-2.06)
GSTT1 null	2.14	(0.60-7.70)	1.30	(0.52-3.22)		1.18	(0.56-2.50)
*GSTP1**C	1.73	(0.28-10.8)	1.11	(0.33-3.75)		0.85	(0.36-2.03)
Ever tobacco smoking	3.39	(1.15-9.99)	2.42	(1.10-5.36)		2.18	(1.14-4.16)
*GSTM1**smoking	1.49	(0.43-5.14)	1.68	(0.74-3.83)		1.81	(0.91-3.63)
*GSTT1**smoking	0.69	(0.17-2.79)	1.17	(0.44-3.10)		1.30	(0.58-3.92)
*GSTP1**smoking	0.49	(0.07-3.47)	0.97	(0.26-3.57)		1.11	(0.45-2.74)
NAT2 slow	1.49	(0.42-5.28)	1.11	(0.75-1.65)	0.05*	1.16	(0.57-2.37)
NAT1 fast	0.37	(0.10-1.37)	0.83	(0.55-1.23)		0.62	(0.29-1.31)
SULT1A1 Arg/Arg	1.76	(0.52-5.88)	1.02	(0.67-1.54)		1.10	(0.55-2.21)
CYP1B1 Val/Val	0.88	(0.20-3.84)	1.30	(0.39-4.36)		1.11	(0.29-4.17)
Ever tobacco smoking	3.50	(0.48-25.5)	1.38	(0.79-2.44)		1.79	(0.66-4.84)
*NAT2**smoking	0.99	(0.26-3.78)	1.41	(0.92-2.16)		1.31	(0.62-4.84)
*NAT1**smoking	2.74	(0.68-11.0)	1.24	(0.80-1.93)		1.63	(0.75-3.55)
*SULT1A1**smoking	0.73	(0.20-2.63)	1.29	(0.87-1.93)		1.20	(0.58-2.46)
*CYP1B1**smoking	0.89	(0.18-4.37)	1.55	(0.53-4.55)		1.21	(0.32-4.65)
NAT2 slow	1.38	(0.90-2.13)	1.41	(0.99-1.99)	0.05*	1.39	(0.93-2.08)
NAT1 fast	0.95	(0.61-1.46)	0.89	(0.62-1.26)		0.91	(0.61-1.36)
SULT1A1 Arg/Arg	1.23	(0.82-1.84)	1.25	(0.87-1.79)		1.25	(0.85-1.83)
CYP1B1 Val/Val	0.90	(0.51-1.58)	0.92	(0.53-1.59)		0.91	(0.52-1.59)
Ever exposure to AA	1.16	(0.12-11.6)	1.09	(0.68-1.77)		1.09	(0.41-2.89)
*NAT2**AA	1.86	(0.34-10.1)	1.61	(0.90-2.88)		1.78	(0.64-4.97)
*NAT1**AA	0.42	(0.08-2.18)	0.89	(0.50-1.61)		0.69	(0.24-1.97)
*SULT1A1**AA	2.10	(0.38-11.5)	1.40	(0.81-2.44)		1.56	(0.56-4.37)
*CYP1B1**AA	1.02	(0.15-7.17)	1.01	(0.50-2.05)		1.03	(0.32-3.33)

NOTE: Hierarchical (1), τ^2 was estimated from the data, or pre-specified as 0.05. Hierarchical (2), τ^2 was pre-specified as 0.35.
Abbreviation: AA, aromatic amines.
*τ^2 was pre-specified as 0.05, because the estimated one was set to 0 or negative value.

Table 4. Selected OR and 95% CI of gene-environment interaction with indicator-term approach from conventional analyses and hierarchical modeling

Allele	Environmental exposure	Case no.	Control no.	Conventional		Hierarchical (1)		Hierarchical (2)	
				OR	(95% CI)	OR	(95% CI)	OR	(95% CI)
COMT	Smoking								
Val/Val	Never	3	14	1	(ref)	1	(ref)	1	(ref)
Val/Val	Ever	59	43	6.23	(1.68-23.1)	3.90	(1.76-8.67)	5.07	(1.66-15.4)
Val/Met, Met/Met	Never	14	39	1.61	(0.40-6.50)	0.95	(0.45-2.02)	1.28	(0.40-4.06)
Val/Met, Met/Met	Ever	125	118	4.78	(1.33-17.1)	3.18	(1.30-7.78)	3.99	(1.30-12.3)
NQO1	Smoking								
Pro/Pro	Never	6	36	1	(ref)	1	(ref)	1	(ref)
Pro/Pro	Ever	107	99	6.46	(2.61-16.0)	4.00	(1.97-8.10)	5.93	(2.34-12.5)
Pro/Ser, Ser/Ser	Never	11	17	3.86	(1.22-12.1)	1.79	(0.89-3.58)	2.89	(1.06-7.89)
Pro/Ser, Ser/Ser	Ever	77	62	7.40	(2.93-18.7)	5.11	(2.26-11.6)	6.43	(2.65-15.6)
MPO	PAH exposure								
A/A	Never	2	13	1	(ref)	1	(ref)	1	(ref)
A/A	Ever	5	9	3.45	(0.54-22.0)	1.44	(0.66-3.16)	2.14	(0.54-8.45)
G/G, G/A	Never	122	128	6.03	(1.33-27.3)	3.37	(1.22-9.30)	4.38	(1.24-15.5)
G/G, G/A	Ever	72	64	7.16	(1.55-33.0)	4.19	(1.36-12.9)	5.34	(1.42-20.0)
GSTT1	AA exposure								
present	Never	143	170	1	(ref)	1	(ref)	1	(ref)
present	Ever	15	11	1.63	(0.73-3.67)	1.71	(0.79-3.70)	1.69	(0.77-3.69)
null	Never	40	32	1.51	(0.90-2.53)	1.54	(0.93-2.55)	1.53	(0.92-2.54)
null	Ever	3	1	4.02	(0.41-39.6)	2.75	(0.89-8.47)	3.13	(0.62-15.9)

NOTE: (1) Pre-specified $\tau^2 = 0.05$;(2) pre-specified $\tau^2 = 0.35$.
Abbreviations: OR, odds ratio adjusted for age; PAH, polycyclic aromatic hydrocarbons; AA, aromatic amines.

model does not provide a reasonable presentation of the biological phenomenon, the estimates from the hierarchical model would also be biased. For example, if one assumes the effect of *GSTM1* null (risk allele) and *MPO A/A* genotype (protective allele) come from the same mean, the estimate of *GSTM1* null may be biased downward and the estimate of *MPO A/A* genotype may be biased upward (5). However, under a reasonably specified prior, the hierarchical models are expected to provide more precise estimates of risks.

Slight alterations of the Z matrix did not have a major influence on the risk estimates. This suggests that the model is relatively robust, instead of implying that one particular Z matrix is ''correct.'' One should aim to construct a Z matrix that reasonably reflects the underlying biological phenomena; however, it would not be feasible to estimate how well the Z matrix reflects ''biological truth'' by a statistical method, including sensitivity analyses.

Note that we used values of 0 or 1 in our second-stage design (Z) matrix. While such a crude design matrix has been shown to improve over conventional estimates (9), it is unlikely to fully capture the true differences between genetic factors. This is more likely to be on a continuous scale, reflecting aspects such as enzyme kinetics, the rate or the level of the adduct formation, etc. A study of *NAT2* genotype-specific dietary effects on adenomatous polyps provided a fine example, in which Z matrix is the calculated conversion rates for heterocyclic amines, derived from the estimated concentrations of heterocyclic amines in the selected foods, and Michaelis-Menten constant estimates for *NAT*s genotypes (12). Our knowledge on the effects of the products of the genes included in our study is far from complete. As this information becomes available, one can further refine the Z matrix. In the context of genetic polymorphisms, the main obstacle for researchers to apply hierarchical modeling may be the construction of prior (Z matrix). Our results suggest that the model is relatively robust, and that a Z matrix which reasonably represents the biological mechanisms can be helpful in improving the risk estimates.

One can estimate τ^2 from the data (EB approach) when the information of the residual effects is unknown, because when using a SB approach, τ^2 is pre-specified and hence potentially subject to criticism. However, previous work has shown that SB generally often outperforms EB, which may estimate τ^2 as equaling zero (2). Regardless, one should consider varying τ^2 to see how sensitive results are to this value, as in our example. Simulation studies have shown that the sensitivity of SB estimates to the pre-specified τ^2 increases with increasing study size (2).

We used two different approaches to investigate gene-environment interactions (i.e., product terms versus indicator variables). When considering more than one pair of interaction, the product-term approach requires fewer terms in one's model, and so allowed us to consider more markers simultaneously than the indicator variable approach. In particular, when looking at the interaction analysis based on the product-term approach, the number of models to test, including genetic main effects and gene-environment interactions was reduced from 56 (considering polymorphisms of 14 genes and 3 environmental exposures) to 13 [1 for genetic main effect, and 12 (4×3) models for gene-environment interactions]. However, the indicator-term approach may more directly exhibit the joint effect of the genetic and environmental factors.

In summary, association studies are commonly using high-throughput genotyping techniques, which generate observations on several hundreds or thousands of genetic markers in large populations. One must be very cautious about multiple comparison and potential false-positive associations when analyzing and interpreting

such data. Hierarchical modeling can help address these issues, provide more plausible and stable estimates, and, therefore, represents an alternative to conventional methods for genetic association studies.

Acknowledgments

The authors thank Umberto Gelatti, Donatella Placidi, and Antonio Scotto di Carlo for their contribution to the bladder cancer case-control study in Brescia.

References

1. Greenland S, Rothman KJ. Fundamental of epidemiologic data analysis. In: Rothman KJ, Greenland S, editors. Modern epidemiology. Philadelphia: Lippincott-Raven; 1998. p. 201-29.
2. Greenland S. Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. Stat Med 1993;12:717-36.
3. Greenland S. Modeling and variable selection in epidemiologic analysis. Am J Public Health 1989;79:340-9.
4. Greenland S, Poole C. Empirical-Bayes and semi-Bayes approaches to occupational and environmental hazard surveillance. Arch Environ Health 1994;49:9-16.
5. Witte JS. Genetic analysis with hierarchical models. Genet Epidemiol 1997;14:1137-42.
6. Greenland S. Principles of multilevel modeling. Int J Epidemiol 2000;29:158-67.
7. Greenland S. Introduction to regression modeling. In: Rothman KJ, Greenland S, editors. Modern epidemiology. 2nd ed. Philadelphia: Lippincott-Raven; 1998. p. 401-32.
8. Witte JS, Greenland S, Haile RW, Bird CL. Hierarchical regression analysis applied to a study of multiple dietary exposures and breast cancer. Epidemiology 1994;5:612-21.
9. Witte JS, Greenland S. Simulation study of hierarchical regression. Stat Med 1996;15:1161-70.
10. Kim LL, Fijal BA, Witte JS. Hierarchical modeling of the relation between sequence variants and a quantitative trait: addressing multiple comparison and population stratification issues. Genet Epidemiol 2001;21 Suppl 1:S668-73.
11. De Roos AJ, Rothman N, Inskip PD, et al. Genetic polymorphisms in GSTM1, -P1, -T1, and CYP2E1 and the risk of adult brain tumors. Cancer Epidemiol Biomarkers & Prev 2003;12:14-22.
12. Aragaki CC, Greenland S, Probst-Hensch N, Haile RW. Hierarchical modeling of gene-environment interactions: estimating NAT2 genotype-specific dietary effects on adenomatous polyps. Cancer Epidemiol Biomarkers & Prev 1997;6:307-14.
13. Shen M, Hung RJ, Brennan P, et al. Polymorphisms of the DNA repair genes *XRCC1, XRCC3*, *XPD*, interaction with environmental exposures, and bladder cancer risk in a case-control study in northern Italy. Cancer Epidemiol Biomarkers & Prev 2003;12:1234-40.
14. Hung RJ, Boffetta P, Brennan P, et al. Genetic polymorphisms of *MPO, COMT*, *Mnsod* and *NQO1*, interactions with environmental exposures and bladder cancer risk. Carcinogenesis. In press 2004.
15. Hung RJ, Boffetta P, Brennan P, et al. GSTs, NATs, SULT1A1, CYP1B1 genetic polymorphisms, interactions with environmental exposures and bladder cancer risk in a high-risk population. Int J Cancer 2004;110:598-604.
16. Conti DV, Witte JS. Hierarchical modeling of linkage disequilibrium: genetic structure and spatial relations. Am J Hum Genet 2003;72:351-63.
17. Witte JS, Greenland S, Kim LL. Software for hierarchical modeling of epidemiologic data. Epidemiology 1998;9:563-6.
18. Lunn RM, Langlois RG, Hsieh LL, Thompson CL, Bell DA. XRCC1 polymorphisms: effects on aflatoxin B1-DNA adducts and glycophorin A variant frequency. Cancer Res 1999;59:2557-61.
19. Wang Y, Spitz MR, Zhu Y, Dong Q, Shete S, Wu X. From genotype to phenotype: correlating XRCC1 polymorphisms with mutagen sensitivity. DNA Repair (Amst) 2003;2:901-8.
20. Petruska JM, Mosebrook DR, Jakab GJ, Trush MA. Myeloperoxidase-enhanced formation of (+-)-*trans*-7,8-dihydroxy-7,8-dihydrobenzo[*a*]-pyrene-DNA adducts in lung tissue *in vitro*: a role of pulmonary inflammation in the bioactivation of a procarcinogen. Carcinogenesis 1992;13:1075-81.
21. Mallet WG, Mosebrook DR, Trush MA. Activation of (+-)-*trans*-7, 8-dihydroxy-7,8-dihydrobenzo[*a*]pyrene to diolepoxides by human polymorphonuclear leukocytes or myeloperoxidase. Carcinogenesis 1991;12:521-4.
22. Kadlubar FF, Butler MA, Kaderlik KR, Chou HC, Lang NP. Polymorphisms for aromatic amine metabolism in humans: relevance for human carcinogenesis. Environ Health Perspect 1992;98:69-74.
23. Piedrafita FJ, Molander RB, Vansant G, Orlova EA, Pfahl M, Reynolds WF. An Alu element in the myeloperoxidase promoter contains a composite SP1-thyroid hormone-retinoic acid response element. J Biol Chem 1996;271:14412-20.
24. Ross D, Traver RD, Siegel D, Kuehl BL, Misra V, Rauth AM. A polymorphism in NAD(P)H:quinone oxidoreductase (NQO1): relationship of a homozygous mutation at position 609 of the NQO1 cDNA to NQO1 activity. Br J Cancer 1996;74:995-6.
25. Ross D, Kepa JK, Winski SL, Beall HD, Anwar A, Siegel D. NAD(P)H:quinone oxidoreductase 1 (NQO1): chemoprotection, bioactivation, gene regulation and genetic polymorphisms. Chem-Biol Interact 2000;129:77-97.
26. Zhu BT, Ezell EL, Liehr JG. Catechol-*O*-methyltransferase-catalyzed rapid *O*-methylation of mutagenic flavonoids. Metabolic inactivation as a possible reason for their lack of carcinogenicity *in vivo*. J Biol Chem 1994;269:292-9.
27. McCord JM. Superoxide dismutase in aging and disease: an overview. Methods Enzymol 2002;349:331-41.
28. Lotta T, Vidgren J, Tilgmann C, et al. Kinetics of human soluble and membrane-bound catechol *O*-methyltransferase: a revised mechanism and description of the thermolabile variant of the enzyme. Biochemistry 1995;34:4202-10.
29. Shimoda-Matsubayashi S, Matsumine H, Kobayashi T, Nakagawa-Hattori Y, Shimizu Y, Mizuno Y. Structural dimorphism in the mitochondrial targeting sequence in the human manganese superoxide dismutase gene. A predictive evidence for conformational change to influence mitochondrial transport and a study of allelic association in Parkinson's disease. Biochem Biophys Res Commun 1996;226:561-5.
30. Rosenblum JS, Gilula NB, Lerner RA. On signal sequence polymorphisms and diseases of distribution. Proc Natl Acad Sci USA 1996;93:4471-3.
31. Hein DW, Doll MA, Fretland AJ, et al. Molecular genetics and epidemiology of the NAT1 and NAT2 acetylation polymorphisms. Cancer Epidemiol, Biomarkers & Prev 2000;9:29-42.
32. Kogevinas M, Trichopoulos D. Urinary bladder cancer. In: Adami HO, Hunter D, Trichopoulos D, editors. Text book of cancer epidemiology. New York: Oxford University Press, Inc.; 2002. p. 446-66.
33. dos Santos Silva I. Size of a study. In: dos Santos Silva I, editor. Cancer epidemiology: principles and methods. Barcelona: IARC Press; 1999. p. 333-53.
34. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 2003;33:177-82.
35. Greenland S, Robins JM. Empirical-Bayes adjustments for multiple comparisons are sometimes useful. Epidemiology 1991;2:244-51.