

The use of hierarchical models for estimating relative risks of individual genetic variants: An application to a study of melanoma

Marinela Capanu^{1,*}, Irene Orlow¹, Marianne Berwick², Amanda J. Hummer¹,
Duncan C. Thomas³ and Colin B. Begg¹

¹Memorial Sloan-Kettering Cancer Center, 307 E 63rd St., 3rd Floor, New York, NY 10021, U.S.A.

²University of New Mexico, MSC 10 5550, 1, Albuquerque, NM 87131, U.S.A.

³Keck School of Medicine of University of Southern California, CHP 220, 9011,
Los Angeles, CA 90089, U.S.A.

SUMMARY

For major genes known to influence the risk of cancer, an important task is to determine the risks conferred by individual variants, so that one can appropriately counsel carriers of these mutations. This is a challenging task, since new mutations are continually being identified, and there is typically relatively little empirical evidence available about each individual mutation. Hierarchical modeling offers a natural strategy to leverage the collective evidence from these rare variants with sparse data. This can be accomplished when there are available higher-level covariates that characterize the variants in terms of attributes that could distinguish their association with disease. In this article, we explore the use of hierarchical modeling for this purpose using data from a large population-based study of the risks of melanoma conferred by variants in the *CDKN2A* gene. We employ both a pseudo-likelihood approach and a Bayesian approach using Gibbs sampling. The results indicate that relative risk estimates tend to be primarily influenced by the individual case–control frequencies when several cases and/or controls are observed with the variant under study, but that relative risk estimates for variants with very sparse data are more influenced by the higher-level covariate values, as one would expect. The analysis offers encouragement that we can draw strength from the aggregating power of hierarchical models to provide guidance to medical geneticists when they offer counseling to patients with rare or even hitherto unobserved variants. However, further research is needed to validate the application of asymptotic methods to such sparse data. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: hierarchical models; Bayesian methods; pseudo-likelihood; genetic risk

*Correspondence to: Marinela Capanu, Memorial Sloan-Kettering Cancer Center, 307 E 63rd St., 3rd Floor, New York, NY 10021, U.S.A.

†E-mail: capanum@mskcc.org

Contract/grant sponsor: National Cancer Institute; contract/grant numbers: CA83180, CA131010, CA46592, CA16086

1. INTRODUCTION

Recent technological progress has led to rapid identification and sequencing of large numbers of genetic variants. Studying the associations between these variants and a particular disease is of great importance to epidemiologists in their quest to decipher the disease etiology. The challenge lies in determining the genetic variants that are most likely to contribute to the occurrence of the disease, thereby allowing genetic counselors to provide informed choices for interventions to reduce the risk of the disease. Hierarchical modeling is a technique that has been shown to provide more accurate and stable estimates of individual variants, by incorporating external information through the higher levels of the multilevel model, and its use in this context is becoming increasingly popular (see, e.g. [1–4], among others). However, these previous applications have been directed at studies of single nucleotide polymorphic variants (SNPs), for the most part chosen to have high minor allele frequency. Indeed previous investigators have often excluded those SNPs that occurred relatively infrequently in their studies. In this article, we focus our attention on the application of hierarchical modeling to a setting where the data are inevitably sparse, a setting we believe to be the one in which this technique has the potential for providing the greatest insights for cancer epidemiologists.

Our investigation involves the examination of a gene known to be a risk factor for cancer. We study the impact of the tumor suppressor gene *CDKN2A* on the incidence of melanoma. However, the methodology would be applicable to similar studies of the many genes now known to be implicated in cancer incidence, including *BRCA1* and *BRCA2* for breast and ovarian cancers, the *APC* gene for colorectal cancer, genes that greatly elevate the risks of many cancers such as *RB* and *TP53*, and others. In this setting there is no doubt that the gene plays an important role. However, many different types of variants are observed to occur, including single nucleotide changes at specific loci, insertions, deletions, and changes both inside and outside of the coding exons, and all of these variants are typically rare. It is of great importance to know which of these individual variants confer risk and which ones are harmless. Our challenge is to be able to utilize the individually sparse empirical evidence we have from the case–control frequencies for each of these variants, combined with the aggregating power conferred by features that are shared by groups of variants, in order to provide reliable evidence about the risk status of each variant or indeed of future variants in the gene which have not heretofore been observed.

The first level of our proposed hierarchical model contains the parameters of fundamental interest, namely the relative risks conferred by the individual genetic variants, and patient-level confounding factors. The higher level links these parameters by a model that draws strength from the fact that the parameters (i.e. the relative risks of the individual variants) may share important characteristics. The maximum likelihood least-squares approach (which performs maximum likelihood at the first stage followed by weighted-least-squares regression at the second stage) and the joint iteratively reweighted-least-squares approach (which uses iterative reweighted least squares at both stages) are computationally the simplest estimation methods for hierarchical models. These methods require initial estimates for all of the first-stage model parameters. When the data are sparse, one must use techniques that do not have this requirement, such as the pseudo-likelihood approach [5, 6] or Markov chain Monte Carlo methods such as Gibbs sampling [7].

The pseudo-likelihood approach of Wolfinger and O'Connell [5] uses Taylor expansions to approximate a generalized linear mixed model (GLMM) with a linear mixed pseudo-model, and the parameters can be estimated using restricted maximum likelihood estimation. The penalized quasi-likelihood (PQL) approach of Breslow and Clayton [6] is a special case of the

pseudo-likelihood technique and involves approximations based on Laplace's method. These techniques allow one to model a large number of random effects and complex covariance structures. Pseudo-likelihood methods have been previously used to study the interaction between diet and the *NAT2* gene [2], and to model the risk of breast cancer as a function of different dietary items [8]. Greenland [9] provides a review of the weighted-least-squares and PQL approaches and concludes that these methods yield very similar results for large samples. However, the PQL method has small sample performance superior to the basic maximum likelihood least-squares approach.

Thomas *et al.* [10] used Gibbs sampling to examine associations between large numbers of candidate human leukocyte antigen genes and insulin-dependent diabetes mellitus. Conti *et al.* [11] employed Bayesian model averaging via the Gibbs sampler to make inferences about the effects of various metabolic genes and their interaction with two exposure variables on the risk of colorectal polyps. However, one limitation of both these approaches is the assumption that the log-relative risk parameters are identically distributed (or exchangeable), which can be unrealistic in many applications [12], such as those in which some genotypes or variants are expected to have much larger relative risks than others.

To the best of our knowledge, none of these techniques have been applied to estimate individual relative risks for a large number of rare genetic variants under more general assumptions. In this article, we show how pseudo-likelihood and Gibbs sampling methods can be employed in hierarchical modeling regression to address this issue by borrowing strength from covariates introduced in the second- or higher-level models. As an application, we used these models to analyze the data from the Genes Environment and Melanoma (GEM) study, described in Section 2. Section 3 describes our methods. In Section 4 we apply the methods to the GEM data, and we conclude with a discussion in Section 6.

2. GEM STUDY

The GEM data were collected through an international multi-center population-based case-control study of melanoma (the GEM study). The study subjects were individuals diagnosed either with a single first invasive primary melanoma (SPM), who served as the 'control' group, or with a second or higher invasive or *in situ* primary melanoma (multiple primary melanoma, MPM) who served as the 'case' group. Details of the rationale for the study design, recruitment of patients, and participation rates of the subjects are provided in Begg *et al.* [13]. *CDKN2A* has been identified as a major melanoma susceptibility gene based on the presence of germline mutations in high-risk melanoma families. In the GEM study, after losses due to non-participation and unsuccessful amplification of the DNA, a total of 1189 individuals with MPM and 2424 with SPM were available for screening of *CDKN2A* mutations. Sequencing identified 43 different nucleotide changes in either the coding region (exons) or the surrounding genetic regions.

In our analyses we characterized the variants on the basis of three higher-level covariates because these are characteristics that are plausibly associated with risk of melanoma at the outset and because they are observed to be associated with risk in the data. The first of these, which we denote by the term 'functional', distinguishes rare variants that can be expected to alter directly the protein product. Thus, a variant is defined as functional if it occurs in one of the coding exons and is either a missense single nucleotide substitution or an insertion or a deletion. A variant is defined as 'non-functional' if it is a (silent) synonymous single nucleotide substitution or if it occurs outside

the coding exons. Note that there are three common polymorphic variants. These were presumed at the outset to be unlikely to be related strongly with risk and are included as (lower-level) fixed-effects covariates in our analyses. These common variants are, in effect, potential confounders as they may or may not co-occur with any given rare variant. Including them as fixed effects allows us to estimate their effects directly without allowing them to dominate the random effects analysis of the rare variants.

The second higher-level covariate included in the analysis is a binary variable which indicates for each variant whether or not the variant has been observed in families with three or more melanoma patients based on data reported by the International Melanoma Genetics Consortium (GenoMEL) [14]. The GenoMEL registry comprises major familial melanoma research groups from North America, Europe, and Australia. Of the 43 variants observed in the GEM study, 11 have been observed in families with three or more melanoma patients in GenoMEL. We note that this covariate is ‘dynamic’, in that the status of a variant may change in the future as new melanoma families are reported worldwide. The final covariate is a bioinformatic predictor known as Polyphen (Polymorphism Phenotyping), a tool that uses protein structure, sequence, and phylogenetic information to predict the impact on the 3-D structure and the consequences of amino acid changes on protein function [15–17]. A Polyphen score is available only for missense variants, and Polyphen scores were also unavailable for a few missense variants located in positions with no homology with the family blocks. Another bioinformatic measure available is Sort Intolerant From Tolerant (SIFT) [18, 19], but we have excluded this measure since it demonstrated no association with case/control status in our data. Further details on how these scores were obtained can be found in Orlov *et al.* [20]. We note that SIFT and Polyphen are also dynamic in that they depend on available knowledge from bioinformatic resources that are being updated constantly.

Of the 43 mutations, 31 were defined as functional. Functional variants were observed in 28 controls and 27 cases, and non-functional rare variants were observed in 38 controls and 21 cases. Frequency data for the individual (rare) variants are provided in Table I, along with the higher-level covariates. Most variants occurred in just one or a few study participants. It is this feature of the data which prevents us from using conventional maximum likelihood methods to estimate the individual effects of these variants on the risk of developing melanoma and which offers the promise of considerable efficiency gains from the use of hierarchical modeling.

3. OVERVIEW OF HIERARCHICAL MODELS

The hierarchical model that we adopt involves the use of a logistic regression model in the first stage and of a linear model in the second stage. The first stage model can be expressed as

$$\text{logit } P(Y|X, W) = X\beta + W\gamma \quad (1)$$

where in our context the $N \times 1$ vector Y represents the vector of indicators of case versus control status for each of the N patients in the study. X represents the $N \times n$ design matrix that indicates which of the n variants are possessed by each patient, and W represents the $N \times m$ matrix of information about the patient factors known to affect the risk of melanoma (i.e. confounders). Thus, the $n \times 1$ vector β contains the random effects of interest which distinguish the risks of the n variants, and the $m \times 1$ vector γ contains the relative risks of the confounders. These could

Table I. Position, functional characteristics, and prevalence of the variants identified in GEM.

<i>CDKN2A</i> nucleotide changes	Controls* (SPM [†])	Cases* (MPM [‡])	Functional	GenoMEL [§]	Polyphen
<i>5'UTR</i>					
(-) 70 G/A	0	1	No	0	—
(-) 34 G>T	2	7	No	1	—
(-) 33 G>C	9	6	No	0	—
(-) 25 C>T	2	0	No	0	—
(-) 14 C>T	2	1	No	0	—
<i>Exon 1α</i>					
c.8_9ins	6	0	Yes	0	—
c.8_33del	0	2	Yes	0	—
c.47 T>G	0	2	Yes	1	2.02
c.67 G>A	0	1	Yes	0	1.82
c.67 G>C	1	0	Yes	1	2.27
c.71 G>C	2	0	Yes	1	1.72
c.87_89delG	2	0	Yes	1	—
c.95 T>C	0	2	Yes	1	2.32
c.123 G>A	4	1	No	0	—
c.131_132insA	0	1	Yes	0	—
c.132 C>A	0	1	Yes	0	—
c.136 C>A	1	0	No	0	—
c.146 T>C	1	1	Yes	0	1.75
c.146 T>G	0	1	Yes	1	1.98
c.149 A>C	1	1	Yes	0	2.47
<i>Intron 1</i>					
c.50+21 G/A	1	0	No	0	—
c.50+37 G/C	15	6	No	0	—
c.50+54 G/A	2	0	No	0	—
c.51-5 G>A	1	0	No	0	—
<i>Exon 2</i>					
c.159 G>A	1	2	Yes	1	2.52
c.159 G>C	1	1	Yes	1	2.52
c.170 C>T	0	2	Yes	0	0.28
c.173 G>A	1	0	Yes	0	0.06
c.174 A>C	1	0	Yes	0	—
c.179 C>T	1	0	Yes	0	0.54
c.247 C>T	0	1	Yes	0	2.67
c.249 C>A	1	0	Yes	0	3.12
c.295 C>T	0	1	Yes	0	2.05
c.301 G>T	1	5	Yes	1	2.29
c.304 G>A	0	1	Yes	0	1.89
c.306 G>A	1	0	Yes	0	—
c.318 G>A	2	1	Yes	0	—
c.322 G>A	2	0	Yes	1	0.96
c.334 C>G	0	2	Yes	1	1.99
c.370 C>T	1	0	Yes	0	0.17
c.373 G>C	3	0	Yes	0	1.99
c.384 G>A	1	0	No	0	—
c.427 G>A	1	0	Yes	0	0.04
c.442 G>A	164	72	No	—	1.41

Table I. *Continued.*

<i>CDKN2A</i> nucleotide changes	Controls* (SPM [†])	Cases* (MPM [‡])	Functional	GenoMEL [§]	Polyphen [¶]
<i>3'UTR</i>					
*29 CC	1769 (73%)	876 (74%)	No	—	—
*29 CG	628 (26%)	294 (25%)	No	—	—
*29 GG	26 (1.07%)	20 (1.68%)	No	—	—
*69 CC	1957 (81%)	950 (80%)	No	—	—
*69 CT	425 (18%)	212 (18%)	No	—	—
*69 TT	29 (1.20%)	22 (1.86%)	No	—	—

*Interested reader should note that the frequencies presented here are cumulative over all patients and thus are different from those in Orlov *et al.* [20] where mutually disjoint combinations are reported.

[†]Single primary melanoma: $n = 2424$.

[‡]Multiple primary melanoma: $n = 1189$.

[§]Variants denoted '1' have been observed in at least one family with three or more melanomas in the GenoMEL registry of melanoma families.

[¶]The Polyphen score is not available for all variants (see text for explanation).

^{||}The common polymorphic variants at 442 G>A and in the 3'UTR are included as confounders in our analysis.

include age, sex, geographical site, and various phenotypic factors such as mole count, hair color, and others. These data are not displayed in Table I, as they are patient specific rather than variant specific, and readers are referred to Berwick *et al.* [21] for further details.

The second-stage model relates the individual risks of the variants to additional, hierarchical information about the genetic variants captured in an $n \times p$ matrix Z , where

$$\beta = Z\pi + \delta \quad (2)$$

with

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

In this model, π is a $p \times 1$ column vector of the second-stage parameters, δ is a vector of residual effects characterizing the imprecision of the log-relative risk parameters in β as a function of the second-stage model covariates, assumed (for convenience) to be statistically independent, and I_n is the $n \times n$ identity matrix. The information in Z that we use for the GEM study is displayed in the last three columns of Table I.

Note that we can combine the first- and second-level models into a mixed-effects logistic model by substituting (2) into (1)

$$\begin{aligned} \text{logit } P(Y|X, W) &= X(Z\pi + \delta) + W\gamma \\ &= XZ\pi + X\delta + W\gamma \end{aligned} \quad (3)$$

$$= U\alpha + X\delta \quad (4)$$

where $U = (XZ, W)$ and $\alpha^T = (\pi^T, \gamma^T)$.

3.1. Pseudo-likelihood method

Consider a GLMM represented as follows:

$$\begin{aligned} Y &= \mu + e \quad \text{such that } g(\mu) = U\alpha + X\delta = \eta \\ \text{cov}(\delta) &= G \\ E(e|\mu) &= 0, \quad \text{cov}(e|\mu) = R_\mu^{1/2} R R_\mu^{1/2} \end{aligned} \quad (5)$$

where α is a vector of fixed effects, δ is a vector of random effects normally distributed with mean 0 and variance matrix G , and $g(\cdot)$ is a differentiable monotonic link function with inverse g^{-1} (in our case, the logit link). Here e is a vector of unobserved residuals, R_μ is a diagonal matrix that contains the variance functions of the model evaluated at μ , and R is unknown.

The model parameters of GLMMs can be estimated using a linearization technique that employs Taylor expansions to approximate the model with a model based on pseudo-data with fewer nonlinear components. One such approach is pseudo-likelihood estimation, which uses an expansion around the common estimate of the best linear unbiased predictors (BLUPs) of the random effects. The linearization method first constructs a pseudo-model and pseudo-data. Following Wolfinger and O'Connell [5], if $\hat{\alpha}$ and $\hat{\delta}$ are known estimates of α and δ , a first-order Taylor series of μ about $\hat{\alpha}$ and $\hat{\delta}$ yields

$$g^{-1}(\eta) \doteq g^{-1}(\hat{\eta}) + \hat{\Delta}U(\alpha - \hat{\alpha}) + \hat{\Delta}X(\delta - \hat{\delta})$$

where

$$\hat{\Delta} = \left(\frac{\partial g^{-1}(\eta)}{\partial \eta} \right)_{\hat{\alpha}, \hat{\delta}}$$

is a diagonal matrix of first-order derivatives of g^{-1} evaluated at $\hat{\eta}$. This can be expressed as

$$\hat{\Delta}^{-1}(\mu - g^{-1}(\hat{\eta})) + U\hat{\alpha} + X\hat{\delta} \doteq U\alpha + X\delta \quad (6)$$

If we denote

$$\hat{\Delta}^{-1}(Y - g^{-1}(\hat{\eta})) + U\hat{\alpha} + X\hat{\delta} \equiv P$$

then note that

$$E(P|\alpha, \delta) = \hat{\Delta}^{-1}(\mu - g^{-1}(\hat{\eta})) + U\hat{\alpha} + X\hat{\delta}$$

which is exactly the left-hand side of (6), and that

$$\text{cov}(P|\alpha, \delta) = \hat{\Delta}^{-1} R_\mu^{1/2} R R_\mu^{1/2} \hat{\Delta}^{-1}$$

One can thus approximate the initial GLMM by the linear mixed model, $P = U\alpha + X\delta + \varepsilon$, based on current values of the parameter estimates. This normal linear mixed model with pseudo-response P , mean η , and $\text{cov}(\varepsilon|\alpha, \delta) = \text{cov}(P|\alpha, \delta)$ can be estimated using either maximum likelihood or restricted maximum likelihood, based on whether the method is called either maximum pseudo-likelihood or restricted pseudo-likelihood, respectively. To initiate the process of computing the pseudo-response, one determines starting values for the covariance parameters, which are computed as minimum variance quadratic unbiased estimates (with 0 priors, MIVQUE0, [22]). The next step

is then to obtain estimates for the fixed and random effects using the standard formulas in the GLIMMIX documentation (see [23]).

Fitting the resulting linear mixed model is itself an iterative process, which upon convergence leads to new parameter estimates that are then used to update the linearization. The predictors $\hat{\delta}$ are the estimated BLUPs in the approximated linear mixed model. For more details on this algorithm and additional formulas, the interested reader is referred to Wolfinger and O'Connell [5] and Schabenberger [23]. Employing this algorithm to the GLMM in (3), one obtains estimates $\hat{\pi}$, $\hat{\delta}$, $\hat{\gamma}$, and their corresponding covariance matrices. From (2), the coefficients for the individual variants can be estimated as

$$\hat{\beta} = Z\hat{\pi} + \hat{\delta}$$

The covariance matrix estimate for $\hat{\beta}$ follows immediately [8]:

$$\text{cov}(\hat{\beta}) = Z \text{cov}(\hat{\pi})Z' + \text{cov}(\hat{\delta}) + Z \text{cov}(\hat{\pi}, \hat{\delta}) + \text{cov}(\hat{\delta}, \hat{\pi})Z' \quad (7)$$

The standard errors of $\hat{\beta}$ can then be obtained as the square roots of the diagonal elements of $\text{cov}(\hat{\beta})$, which then can be used to compute 95 per cent confidence intervals.

This technique also produces the estimate \hat{G} [5]. Since the random coefficients δ are assumed to have a common variance, G is restricted to $\tau^2 I_n$, and thus the procedure reduces this task to the estimation of a single parameter τ^2 .

Instead of estimating τ^2 empirically, one can also use a semi-Bayes (SB) approach that pre-specifies τ^2 based on one's expectation that 95 per cent of the true odds ratios would fall within an interval width $\exp(2 \cdot 1.96\sqrt{\tau^2})$. Greenland [24] compares the performance of the empirical Bayes (EB) approach with that of the SB approach and concludes that both lead to similar results for large studies when SB correctly specifies τ^2 , but the SB approach is sensitive to misspecification of the variance. However, for small studies, the SB estimation can be superior to EB and also appears robust to variance misspecification. Nevertheless, if the SB approach is adopted, one needs to perform sensitivity analyses to check how the results vary with different pre-specified τ^2 values. For more discussion and details on the estimation of τ^2 , see Greenland [9, 24] and SAS Institute Inc. [25].

3.2. Bayesian analysis using the Gibbs sampling

In what follows, we present the joint posterior distribution of the unknown parameters arising for a GLMM for canonical one-parameter exponential families [26, 27]. Assume that conditional on a random effect δ , the responses y_i are independent and follow an exponential family with conditional mean related to the linear predictor through $g(\mu) = U\alpha + X\delta = \eta$, with $\delta \sim N(0, G)$. A Bayesian model is completed by the specification of prior distributions for α and G . If we assume the prior for α follows a $N(0, S)$ distribution independent of G , then the joint posterior distribution is

$$p(\alpha, \delta, G | Y, X, U) \propto \exp \left\{ -\frac{1}{2} \alpha^T S^{-1} \alpha + \sum_{i=1}^N \frac{y_i \theta_i - b(\theta_i)}{\phi} \right\} |G|^{-1} \exp \left\{ -\frac{1}{2} \delta^T G^{-1} \delta \right\} p(G) \quad (8)$$

where $p(G)$ is the prior distribution for G , $\theta_i = \eta_i$ is the canonical parameter, and ϕ is the dispersion parameter assumed to be known (for the logistic case, $b(\theta) = \log(1 + \exp(\theta))$ and $\phi = 1$).

In the case of GLMMs such as the one in (5), the Gibbs sampling generates random draws from the joint posterior distribution of the unknown parameters, conditional on the observed data, namely from $[\alpha, \delta, G|Y, X, U]$, based on the following full conditioned distributions:

$$[\alpha|\delta, G, Y, X, U], \quad [\delta|\alpha, G, Y, X, U] \quad \text{and} \quad [G|\alpha, \delta, Y, X, U] \quad (9)$$

Note that when the prior on the variance components is inverse-Gamma or inverse-Wishart (which are ‘conditionally conjugate’ for this model), then the last full conditional distribution in (9) has a standard form. The first full conditionals do not have a closed form and cannot be sampled directly. However, when G is a diagonal matrix (as it is our case: $G = \tau^2 I_n$ (see Section 3.1)), it has been shown that the full conditional distributions of the fixed and random effects are log-concave, and thus one can use the adaptive rejection sampling of Gilks [28] to sample from these quantities (this is the default sampling method for log-concave distributions in WinBUGS).

For our application we used an inverse-Gamma prior distribution for the variance τ^2 , which leads to a closed form of its full conditional distribution, from which we can simulate directly.

Details on the use of Gibbs sampler and on the full conditional distributions in the case of GLMMs can be found in Zeger and Karim [26], Gamerman [29], and Natarajan and Kass [27]. WinBUGS [30] software was used to generate the samples and derive inferences on the parameters of interest and the WinBUGS code is supplied in Appendix A.

4. APPLICATION

We fit hierarchical models to the GEM data using both methods. The first-stage model contained 43 indicators of silent or functionally relevant mutations and the potential confounder variables age, sex, and age-by-sex, interaction, along with three common polymorphisms: 442 G>A, *nt* 500 (GG versus GC or CC) and *nt* 540 (TT versus TC or TT). We adjust for age and sex because both these factors are known from cancer incidence registries to be strongly associated with melanoma risk, and we adjust for age by sex interaction because it is known that the age incidence curve is much steeper for men. We elected not to adjust for phenotypic risk factors of likely genetic origin such as hair color, mole count, and others, and for the major environmental risk factor, sun exposure. The coefficients β represent the relative risks of individual genetic variants for melanoma on a log scale, and their estimation is the main focus of the analysis. Variants for which the Polyphen score was not available were assigned a 0 score, and an indicator that takes value 1 if the score is available and 0 otherwise was included in the model to account for this missing information.

The pseudo-likelihood method was accomplished using the GLIMMIX macro following the code provided by Witte *et al.* [8]. The major distinction between Witte’s program and ours is that rather than using an SB approach which requires specification of τ^2 , we made use of the options in the GLIMMIX macro that allowed us to estimate empirically the common variance τ^2 . For purposes of comparison, we also included results obtained from the SB methods with prespecification of τ^2 to a set of selected values. Other differences between our program and Witte’s is that all our second-stage models contained an intercept (as compared with none in Witte’s model) and that the standard errors of our estimates were adjusted using the Kenward–Roger adjustment [31], an inflation factor that takes into account the uncertainty in estimating the parameters τ and σ . The existing MIXED and GLIMMIX procedures implement this adjustment only for the matrix of covariances of the fixed effects and for the standard errors of the

random effects. However, for the purposes of our application, the full-adjusted matrix of covariances of the fixed and random effects is needed to compute the variances of the individual log-relative risk estimates in (7). The SAS Help and Documentation manual for these procedures incorrectly states that these covariances are adjusted for the Kenward–Roger inflation factor. We therefore used a macro that implements the Kenward–Roger adjustment for all the desired estimates. Details on this implementation along with the macro that was used are provided in the Appendix.

We implemented the Gibbs sampling algorithm to fit the GLMM in (3) and assumed non-informative priors for the hyperparameters (an inverse-Gamma distribution with both shape and scale parameters set at 0.01 for τ^2 and normal distributions with mean 0 and variance 100 for the other hyperparameters). These models were estimated using the WinBUGS software based on 8500 burn-ins followed by 10 000 iterations.

For purposes of benchmarking, we also employed ordinary logistic regression, i.e. with no second-level modeling. However, this method produces maximum likelihood estimates (MLEs) only for the relatively few variants that occurred in at least one case and at least one control.

5. RESULTS

Table II presents estimates of the odds ratios and corresponding 95 per cent confidence and credible intervals for the first-stage covariates for the pseudo-likelihood method and the Bayesian method, respectively (columns 1 and 2). It also includes results obtained from the pseudo-likelihood method under the assumption that the random effects β are exchangeable (corresponding to no second-stage covariates) and the MLEs obtained from an ordinary logistic regression (first-stage model with

Table II. Odds ratios and corresponding 95 per cent confidence intervals and credible intervals (Bayesian) for the first-stage covariates under Gibbs sampling, pseudo-likelihood, and maximum likelihood methods.

Variable	Odds ratios (95 per cent CI)				
	Bayesian	Pseudo-likelihood	Exchangeable*	Maximum likelihood	
Age	< 45	1.0	1.0	1.0	1.0
	45–59	1.8 (1.3–2.5)	1.8 (1.3–2.5)	1.8 (1.3–2.5)	1.8 (1.3–2.5)
	60–74	3.7 (2.7–5.0)	3.6 (2.6–5.0)	3.6 (2.6–4.9)	3.6 (2.6–5.0)
	75+	3.7 (2.5–5.5)	3.7 (2.5–5.4)	3.7 (2.5–5.4)	3.7 (2.5–5.5)
Sex	Female	1.0	1.0	1.0	1.0
	Male	1.0 (0.7–1.5)	1.0 (0.7–1.6)	1.0 (0.7–1.6)	1.0 (0.7–1.6)
Age × sex	45–59, male	1.2 (0.7–2.0)	1.1 (0.7–1.9)	1.1 (0.7–1.9)	1.2 (0.7–2.0)
	60–74, male	1.5 (0.9–2.4)	1.4 (0.9–2.4)	1.4 (0.9–2.4)	1.5 (0.9–2.4)
	75+, male	1.9 (1.1–3.4)	1.9 (1.1–3.3)	1.9 (1.1–3.3)	1.9 (1.1–3.4)
Polymorphism	442GA	1.0 (0.7–1.3)	1.0 (0.7–1.3)	1.0 (0.7–1.3)	1.0 (0.7–1.3)
	ex3 nt500 [†]	2.0 (1.1–3.7)	2.0 (1.1–3.7)	2.0 (1.1–3.7)	1.9 (1.0–3.6)
	ex3 nt540 [‡]	1.5 (0.9–2.7)	1.5 (0.9–2.8)	1.5 (0.9–2.8)	1.5 (0.8–2.7)

*This model was fit by the GLIMMIX macro assuming exchangeability of the random effects.

[†]GG versus CC or CG.

[‡]TT versus CC or CT.

fixed effects only). Estimates for the confounder variables included in the first-stage model are very similar for all of these methods and are not sensitive to different prespecified values of τ^2 for the SB pseudo-likelihood approach (data not shown), suggesting that with sufficient sample size, the methods accurately estimate these parameters.

In subsequent tables we present analyses of the individual variants. We include in these tables all variants observed on at least three patients and a representative selection of the remaining variants, which are each observed only once or twice in the data set. In Tables III and IV, we present the results of four analyses based on the pseudo-likelihood method using different combinations of the higher-level covariates. Models 1–4 all include an intercept and, respectively, no higher-level covariates, functional status alone, functional status plus history of familial aggregation (GenoMEL), and finally both these in addition to the Polyphen score and the indicator for missing Polyphen scores. The coefficients of the higher-level covariates are given in Table III. All three higher-level covariates appear to contain important information for predicting the risks of the individual variants, although their individual rate ratios are attenuated when they are included together in the model, due to modest collinearity between these factors.

In Table IV we see that the relative risks of the individual variants seem to be well differentiated by the second-stage covariates and by the empirical case–control relative frequencies. The results for the six most frequent variants at the top seem to show estimators that are influenced by the observed case–control (MPM/SPM) ratios of counts. That is, high odds ratios (both statistically significant) are estimated for the two variants that have a strong preponderance of MPM cases, namely –34 G>T, despite the fact that this variant is non-functional, and c.301 G>T, although this variant is both functional and has been observed in melanoma families also. Null or negative associations are estimated for the other four of these relatively frequent variants (despite the fact that c.8_9 is a ‘functional’ variant). The remaining variants with three or fewer occurrences seem to be strongly influenced by the hierarchical covariates. For example, c.159 G>C has a baseline relative risk estimate of 2.2, but this rises to 2.7 on the basis of its classification as functional (Model 2), to 4.5 on the basis of its additional classification as a ‘familial’ variant, and to 6.5 on the basis of its high Polyphen score. Conversely, c.170 C>T has a relatively high relative risk of 4.7 based on its classification as

Table III. Odds ratios and corresponding 95 per cent confidence intervals for the higher-level covariates under pseudo-likelihood methods.

Second stage	Model 1 ^{*,†}	Model 2 [‡]	Model 3 [§]	Model 4 [¶]
Functional	—	2.8 (1.4–5.9)	1.6 (0.7–3.9)	1.0 (0.3–4.2)
GenoMEL	—	—	3.6 (1.0–12.5)	2.3 (0.6–9.1)
Polyphen	—	—	—	2.2 (0.7–6.5)
Polyphen exist	—	—	—	0.6 (0.04–8.1)
$\hat{\tau}$	1.16	1.19	0.94	0.98

*A ‘—’ indicates that the corresponding variable was not included in the model.

[†]Model 1 assumes exchangeability of the random effects β .

[‡]Model 2 contains an intercept and functional status as second-stage covariates.

[§]Model 3 contains as second-stage covariates an intercept, the functional status, and an indicator whether the variant was observed in GenoMEL [14].

[¶]Model 4 contains the following second-stage covariates: intercept, functional status, an indicator whether the variant was observed in GenoMEL, the Polyphen score, and an indicator whether the Polyphen score is missing or not.

Table IV. Odds ratios and corresponding 95 per cent confidence intervals for selected variants using maximum likelihood and pseudo-likelihood methods under different second-stage model scenarios.

Variant	Odds ratios (95 per cent CI)										
	SPM	MPM	Functional	GenoMEL	Poly*	MLE†	Model 1‡	Model 2§	Model 3¶	Model 4	
c:50+37 G/C	15	6	No	0	0	0.8 (0.3-2.1)	0.9 (0.4-2.3)	0.8 (0.3-2.1)	0.8 (0.3-2.1)	0.8 (0.3-2.1)	0.8 (0.3-2.1)
(-)33G>C	9	6	No	0	0	1.2 (0.4-3.6)	1.3 (0.5-3.7)	1.2 (0.4-3.3)	1.1 (0.4-3.3)	1.1 (0.4-3.3)	1.1 (0.4-3.3)
(-)34G>T	2	7	No	1	1	13.5 (2.7-67.3)	7.6 (2.1-27.5)	6.4 (1.8-23.3)	7.5 (2.0-28.6)	6.8 (1.8-25.3)	6.8 (1.8-25.3)
c:301G>T	1	5	Yes	1	2.29	14.7 (1.6-132.3)	6.3 (1.4-29)	7.3 (1.5-35.4)	8.8 (1.8-44.2)	10.8 (2.0-57.8)	10.8 (2.0-57.8)
c:8:9ins	6	0	Yes	0	0		0.6 (0.1-3.5)	0.7 (0.1-3.9)	0.7 (0.1-3.7)	0.5 (0.1-3.3)	0.5 (0.1-3.3)
c:123G>A	4	1	No	0	0	0.4 (0.0-5.0)	0.9 (0.2-4.5)	0.6 (0.1-3.6)	0.7 (0.1-3.7)	0.6 (0.1-3.7)	0.6 (0.1-3.7)
(-)14C>T	2	1	No	0	0	1.2 (0.1-14.6)	1.6 (0.2-14.3)	0.8 (0.1-10.1)	0.8 (0.1-6.7)	0.8 (0.1-7.5)	0.8 (0.1-7.5)
(-)25C>T	2	0	No	0	0		1.1 (0.1-8.9)	0.7 (0.1-6.8)	0.7 (0.1-5.4)	0.7 (0.1-5.9)	0.7 (0.1-5.9)
c:373G>C	3	0	Yes	0	1.99		0.9 (0.1-6.4)	1.1 (0.2-7.6)	1.0 (0.1-6.2)	1.3 (0.2-9.0)	1.3 (0.2-9.0)
c:87_89delG	2	0	Yes	1	0		0.9 (0.1-6.3)	1.0 (0.1-7.6)	2.2 (0.3-15.1)	1.1 (0.1-10.0)	1.1 (0.1-10.0)
c:318G>A	2	1	Yes	0	0	2.0 (0.2-23.1)	2.0 (0.3-12.5)	2.4 (0.4-15.0)	1.8 (0.3-10.7)	1.3 (0.2-9.5)	1.3 (0.2-9.5)
c:159G>A	1	2	Yes	1	2.52	2.3 (0.2-25.5)	2.2 (0.4-13.0)	2.6 (0.4-16.4)	4.2 (0.6-28.1)	6.1 (0.8-46.3)	6.1 (0.8-46.3)
c:159G>C	1	1	Yes	1	2.52	2.5 (0.1-41.4)	2.2 (0.3-15.2)	2.7 (0.4-19.3)	4.5 (0.7-31.0)	6.5 (0.8-50.4)	6.5 (0.8-50.4)
c:334C>G	0	2	Yes	1	1.99		6.3 (0.9-43.8)	8.0 (1.1-59.6)	10.2 (1.5-70.7)	11.5 (1.6-85.5)	11.5 (1.6-85.5)
c:170C>T	0	2	Yes	0	0.28		3.6 (0.5-28.3)	4.7 (0.5-40.5)	2.7 (0.4-18.1)	1.6 (0.2-14.0)	1.6 (0.2-14.0)
c:95T>C	0	2	Yes	1	2.32		7.1 (1.0-48.4)	9.0 (1.2-65.3)	11.0 (1.6-75.0)	15.0 (2.0-111.9)	15.0 (2.0-111.9)
(-)70G/A	0	1	No	0	0		3.3 (0.4-28.0)	2.0 (0.2-19.6)	1.4 (0.2-11.3)	1.6 (0.2-12.9)	1.6 (0.2-12.9)
c:146T>G	0	1	Yes	1	1.98		3.7 (0.4-31.6)	4.9 (0.5-44.9)	7.5 (1.0-58.9)	8.5 (1.0-72.2)	8.5 (1.0-72.2)
c:247C>T	0	1	Yes	0	2.67		4.5 (0.5-37.8)	5.9 (0.7-53.2)	3.0 (0.4-22.1)	7.6 (0.8-74.0)	7.6 (0.8-74.0)
c:67G>A	0	1	Yes	0	1.82		3.9 (0.5-32.5)	5.1 (0.6-46.1)	2.7 (0.4-19.8)	4.0 (0.5-32.8)	4.0 (0.5-32.8)
c:249C>A	1	0	Yes	0	3.12		1.3 (0.2-11.7)	1.7 (0.2-15.5)	1.3 (0.2-9.4)	4.0 (0.3-46.6)	4.0 (0.3-46.6)
c:370C>T	1	0	Yes	0	0.17		1.6 (0.2-14.1)	2.0 (0.2-19.1)	1.4 (0.2-10.5)	0.6 (0.0-9.5)	0.6 (0.0-9.5)
c:384G>A	1	0	No	0	0		1.3 (0.2-11.7)	0.8 (0.1-8.3)	0.7 (0.1-6.1)	0.8 (0.1-6.7)	0.8 (0.1-6.7)
c:67G>C	1	0	Yes	1	2.27		1.6 (0.2-14.4)	2.0 (0.2-19.5)	4.1 (0.5-31.8)	5.3 (0.6-44.3)	5.3 (0.6-44.3)

* A ' - ' indicates that the Polyphen score could not be computed.

† Variants without at least one case and one control do not have a finite MLE.

‡ Model 1 assumes exchangeability of the random effects β .

§ Model 2 contains an intercept and the functional status as a second-stage covariate.

¶ Model 3 contains as second-stage covariates an intercept, the functional status, and an indicator whether the variant was observed in GenoMEL [14].

|| Model 4 contains the following second-stage covariates: intercept, functional status, an indicator whether the variant was observed in GenoMEL, and the Polyphen score.

functional, but absence of evidence of prior family clustering and a low Polyphen score reduces this estimate to 1.6. Confidence intervals are generally wide for these rarer mutations. However, it is clearly possible to achieve individual statistical significance, an example being c.334 C>G, where the designation of functionality, the fact that this variant has been observed in melanoma families, and the fact that both carriers are cases are sufficient, after drawing strength from the aggregating powers of the model, to conclude with statistical significance that this variant confers a high risk.

It is of interest to compare the results of ordinary logistic regression (MLE) with those from Model 1, which is in essence a random-effects model with no higher-level covariates. It is noticeable that the more extreme odds ratio estimates from the logistic regression are attenuated substantially in the random-effects setting (Table IV), notably those at -34 G>T, c.301 G>T, and c.123 G>A, a familiar shrinkage phenomenon of EB techniques.

Table V contrasts the results from the Bayesian (Gibbs sampling) and pseudo-likelihood approaches. The results from the two methods are generally similar, although the Bayesian approach seems to lead to more extreme estimates in several cases and generally wider intervals (note that the Bayesian method produces a larger estimate for τ than the pseudo-likelihood method: 1.9 versus 1.2). One limitation of running the Gibbs sampler in this scenario is that due to the sparseness of the data the procedure is computationally intensive. For the GEM data, on an Intel Pentium 4 CPU 3.06 GHz workstation it takes 4 min in WinBUGS (1.35 min in OpenBUGS on Linux) per each additional iteration, thus needing almost 67 (23)h per 1000 iterations. To ensure convergence of the Markov chain, one would typically allow it to run for a large number of iterations, a task that may be infeasible depending on the data set. In contrast, the pseudo-likelihood method implemented by the GLIMMIX macro converges and produces results in a few seconds.

To assess convergence we used the graphical tools provided by the WinBUGS menu (trace, history, and quantile plots) and also some more formal convergence diagnostics such as the Gelman–Rubin convergence statistic [32] and the convergence tests proposed by Heidelberger and Welch [33]. On the basis of the three chains, the scale reduction factors of Gelman and Rubin [32] and Brooks and Gelman [34] were all close to 1 suggesting satisfactory convergence. After 8500 burn-in iterations followed by 10 000 iterations, all the parameters in the model passed the stationarity test of Heidelberger and Welch [33]. However, not all of these passed the half-width test, where the accuracy of the test was $\varepsilon=0.1$ (which is the CODA default). These results can be seen in Table V for our selected set of the investigated variants.

Recognizing that hierarchical analyses can be sensitive to the choice of prior [35], we conducted a small sensitivity analysis in which we varied the scale and shape parameters of the inverse-Gamma prior for τ from 0.01 to 0.0001. A Uniform(0, 100) distribution was also assumed for comparison. The choice of the scale and shape parameters of the inverse-Gamma distribution did not change the results substantially, although the relative risk estimates of the variants were slightly more variable in the case of the inverse-Gamma(0.0001, 0.0001). The choice of a uniform prior led to somewhat more extreme estimates for the relative risks of some of the variants and generally wider credible intervals. The odds ratios and credible intervals for the first- and second-stage covariates were very similar regardless of the prior distribution for τ .

Finally, we examined the SB approach by comparing the odds ratio estimates of the variants for varying pre-specified values of τ . The results show that the estimates diverge monotonically, suggesting that pre-specification of τ is not a good strategy in the absence of strong prior information about τ as it can lead to radically different estimates. As an example, c.301 G>T has an odds ratio

Table V. Odds ratios and corresponding 95 per cent confidence intervals and credible intervals (Bayesian) for selected variants and for second-stage covariates under Gibbs sampling, pseudo-likelihood methods, and diagnostic tests for the Gibbs sampler (p=passed, f=failed).

Level	Variant	Controls	Cases	Functional	Odds ratios (95 per cent CI)		Heidelberger tests	
					Pseudo-likelihood	Bayesian	Stationary	Halfwidth
	c.50+37 G/C	15	6	No	0.8 (0.3–2.1)	0.8 (0.3–2)	p	p
	(–) 33 G>C	9	6	No	1.2 (0.4–3.3)	1.2 (0.4–3.3)	p	p
	(–) 34 G>T	2	7	No	6.4 (1.8–23.3)	8.9 (2–50.7)	p	p
	c.301 G>T	1	5	Yes	7.3 (1.5–35.4)	10.0 (1.8–92.9)	p	p
	c.8_9ins	6	0	Yes	0.7 (0.1–3.9)	0.3 (0.0–2.4)	p	p
	c.123 G>A	4	1	No	0.6 (0.1–3.6)	0.5 (0.0–3.1)	p	p
	(–) 14 C>T	2	1	No	0.8 (0.1–10.1)	1.1 (0.1–9.2)	p	f
	(–) 25 C>T	2	0	No	0.7 (0.1–6.8)	0.3 (0.0–5.7)	p	f
	c.373 G>C	3	0	Yes	1.1 (0.2–7.6)	0.4 (0.0–4.9)	p	f
	c.87_89delG	2	0	Yes	1.0 (0.1–7.6)	0.4 (0.0–5.5)	p	f
	c.318 G>A	2	1	Yes	2.4 (0.4–15.0)	2.0 (0.2–15.0)	p	p
	c.159 G>A	1	2	Yes	2.6 (0.4–16.4)	2.6 (0.3–26.5)	p	p
	c.159 G>C	1	1	Yes	2.7 (0.4–19.3)	2.4 (0.2–28.0)	p	p
	c.334 C>G	0	2	Yes	8.0 (1.1–59.6)	16.9 (1.3–1072.8)	p	p
	c.170 C>T	0	2	Yes	4.7 (0.5–40.5)	8.4 (0.6–507.2)	p	p
	c.95 T>C	0	2	Yes	9.0 (1.2–65.3)	19.4 (1.5–1126.6)	p	p
	(–) 70 G/A	0	1	No	2.0 (0.2–19.6)	4.1 (0.2–245.7)	p	p
	c.146 T>G	0	1	Yes	4.9 (0.5–44.9)	8.5 (0.5–641.6)	p	p
	c.247 C>T	0	1	Yes	5.9 (0.7–53.2)	10.6 (0.6–769.7)	p	p
	c.67 G>A	0	1	Yes	5.1 (0.6–46.1)	5.2 (0.3–370.2)	p	p
	c.249 C>A	1	0	Yes	1.7 (0.2–15.5)	0.8 (0.0–15.0)	p	f
	c.370 C>T	1	0	Yes	2.0 (0.2–19.1)	1.0 (0.0–17.7)	p	f
	c.384 G>A	1	0	No	0.8 (0.1–8.3)	0.4 (0.0–9.4)	p	f
	c.67 G>C	1	0	Yes	2.0 (0.2–19.5)	1.0 (0.0–20.5)	p	f
Second stage				Functional	2.8 (1.4–5.9)	2.4 (0.7–6.8)		
				$\hat{\tau}$	1.2	1.9		

of 3.5 for $\tau=0.1$, 4.6 for $\tau=0.5$, 6.0 for $\tau=0.8$, 7.0 for $\tau=1$, 7.3 for $\tau=1.2$, and 10.0 for $\tau=1.9$. Other variants exhibited similar trends.

6. DISCUSSION

Bayesian analysis using Gibbs sampling and the pseudo-likelihood method emerge as being the only feasible approaches possibly applicable to the problem of estimating the individual risks conferred by rare genetic variants [8, 9]. The novelty of this article lies in the application of these methods to estimate the individual effects of very sparse variants (even for variants with only a single carrier observed in the data set). To the best of our knowledge, this problem has not been

previously addressed; yet it is likely to be of great importance to epidemiologists and genetic counselors as more detailed information emerges from epidemiologic studies of the effects of major cancer genes. Other authors have proposed multi-level or Bayesian models to address related problems. For example, Parmigiani *et al.* [36] used a Bayesian hierarchical approach to model the probability that a woman is a carrier of *any* mutation in the *BRCA1* and *BRCA2* genes. Goldgar *et al.* [37] developed a multifactorial likelihood-ratio approach to estimate the odds of causality of a specific genetic variant in the setting of a familial segregation analysis, using information on amino acid conservation, severity of amino acid change, and functional data. Neither of these approaches was designed to improve the risk estimation for individual variants. By contrast, the hierarchical modeling methods we have used draw upon the information in important higher-level covariates to augment the meager information from the case and control frequencies of subjects carrying the variant. They provide a viable analytic strategy, which offers the promise that we will be able to distinguish variants that can increase risk from those that are harmless, thus facilitating genetic counseling.

To effectively apply these methods, one needs important second-stage covariates that are strongly associated with the disease. That is, if a higher-level covariate successfully categorizes the variants into high- and low-risk categories on the basis of their aggregated case-control ratios, then the risks conferred by individual variants with minimal data can be predicted more accurately on the basis of this higher-level classification. In the context of our example, we see that 'functional' variants in aggregate have a greatly elevated risk, whereas non-functional variants appear to confer little or no increased risk. Conversely, if a variant is known to be 'functional', then we can infer that it probably confers increased risk in the absence of contradictory evidence from case-control frequencies. Our other two higher-level covariates also provide substantial capacity to sort the variants into high- and low-risk groups.

Much current research is directed at the problem of determining which genetic variants within known genes confer risk. These bioinformatic tools are generally deductive, in that they are based on assumptions about the relevance of genomic features of the variants that are postulated to predict functional relevance. For example, SIFT is based on the premise that changes in amino acids that are evolutionarily conserved within a family of proteins or across species are more likely to have an important impact on function [18, 19]. PolyPhen uses protein structure, sequence, and phylogenetic information to predict the impact on the 3D-structure and the consequences of amino acid changes on protein function [15–17]. Recent trends in cancer epidemiologic research have led to a climate in which investigators are discouraged from publishing case-control associations of candidate genetic variants based on purely empirical findings, in the absence of evidence that the variant under study is likely to affect the gene [38]. Although this issue is controversial [39–43], it is abundantly clear that information on the likely impact of the variant can be highly relevant. With improvement of bioinformatic tools to predict effect, our opinion is that the best way to determine the relevance of individual variants is to utilize well-designed epidemiologic studies in which primacy is given to the data, but where the data analytic inferences can draw strength from the bioinformatic predictors to the extent that these predictors are observed empirically to be associated with disease risk. Hierarchical modeling provides a statistical approach that combines in a natural way the evidence relating the individual variants with risk, and the evidence linking the bioinformatic predictors with risk.

Our data example is limited by the fact that there are relatively few variants, and data are sparse for most of them. However, as more data are assembled, the precision of the estimates of the hierarchical predictors should increase, and this in turn should improve our ability to predict the

risk conferred by the individual rare variants. Similarly, as the tools for bioinformatic prediction are refined, this too can only improve the capability of these methods.

Model fit is not a focus of this paper and the models presented for the GEM data are not a result of a model selection procedure but rather an illustration of how information from higher-level covariates can be used to improve our ability to predict the risks conferred by individual rare variants. We recognize that there are available Bayesian criteria for model comparison [44]. However, as the Gibbs sampling is computationally intensive in this setting, running multiple models for the purpose of model selection is practically unfeasible. To the best of our knowledge, there is no available method to compare different models under the pseudo-likelihood method as this procedure uses the log-likelihood of a linearized model, and thus it employs a pseudo-log-likelihood that cannot be compared formally across different models even if the models are nested.

Finally, we recognize a major limitation with our approach. The confidence intervals obtained with the pseudo-likelihood method are based on large sample approximations that may have poor coverage properties in the sparse data configurations under consideration. Indeed, for the sparse variants, the disparities between the confidence intervals obtained with the pseudo-likelihood method and the credible intervals obtained with the Bayesian approach that we used as a benchmark give cause for caution. Further research is needed to investigate the validity of the method as a function of the sample sizes available for each variant. The disparities in the estimates of the variance τ of the second-stage model between the pseudo-likelihood and Bayesian methods also need to be explored further.

APPENDIX A

A.1. SAS code for the pseudo-likelihood method and the Kenward–Roger adjustment

We attach below the SAS code that was used to fit the pseudo-likelihood method with τ estimated from the data. The macros `%getXZ` and `%get_oput` can be downloaded from John Witte's webpage (<http://darwin.cwru.edu/~witte/glimmix.htm>), whereas the `%glimmix` macro can be downloaded from: <http://support.sas.com/ctx/samples/index.jsp?sid=536>.

To implement the Kenward–Roger adjustment, we made use of the fact that the standard errors of the estimates obtained with the ESTIMATE statement in the `%glimmix` macro are properly adjusted with the Kenward–Roger adjustment. To obtain the full-adjusted matrices of covariances between fixed and random effects, we wrote a macro (`%estimate`) that takes as arguments the number of fixed and random effects in the model and outputs ESTIMATE statements for each of the differences between the fixed and the random effects. Using the formula $\text{var}(a - b) = \text{var}(a) + \text{var}(b) - 2\text{cov}(a, b)$, we derived the covariances between the fixed and random effects in terms of the variances of the fixed effects, variances of the random effects, and the variances of the differences between each fixed and each random effect, quantities that are all directly adjusted for the Kenward–Roger adjustment when the option `ddfm=kewardroger` is used in the `%glimmix` macro. Once the full-adjusted matrix of covariances of the fixed and random effects was obtained, equation (7) was then used to compute the adjusted standard errors of the log-relative risk estimates $\hat{\beta}$. We expanded `%get_oput` macro of Witte to incorporate these computations that were implemented in the macro `%get_oput_kenward` not presented here due to space limitations, but available upon request from the authors.

```

%glimmix(data=dataset, procopt=mmeqsol,
stmts=%str(
model Y1 = W1 W2 W3 W4 W5 W6 W7 W4*W7 W5*W7 W6*W7
coll col2/ solution covb DDFM=KENWARDROGER;
random X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15 X16 X17
X18 X19 X20 X21 X22 X23 X24 X25 X26 X27 X28 X29 X30 X31 X32
X33 X34 X35 X36 X37 X38 X39 X40 X41 X42 X43 / g solution type=toep(1);
%estimate(2,43)
%str(make 'mmeqsol' out=mmeqsol;make 'g' out=g;make 'solutionr' out=solutionr;
make 'Estimates' out=Estimates;make 'covb' out=covb;),
options=mixprintlast, error=binomial
)
run;

%get_oput_kenward(dir=G:/GEM, Xdata=X_dat, Zdata=Z_dat,
nvarX=43, nvarZ = 2, nvarW=10, nmsol=57);

```

A.2. The Gibbs sampler

WinBUGS code: We attach below the WinBUGS code that was used to implement the Gibbs sampling method. Note that in WinBUGS the Normal distribution is defined as $N(\text{mean}, \text{precision})$ where $\text{precision} = 1/\text{variance}$; for this application we used normal priors with mean 0 and variance 100. The precision of the normal distribution of the random effects was given as prior a $\text{Gamma}(0.01, 0.01)$ that has mean 1 and variance 100.

```

model {
for (n in 1:3613)
{ status[n] ~ dbern(pr[n]); sumG[n] ← inprod(b[], v[n,])
sumW[n] ← alpha1*age45_59[n] + alpha2*age60_74[n] + alpha3*agegt75[n] + alpha4*male[n]
+ alpha11*age45_59[n]*male[n] + alpha21*age60_74[n]*male[n] + alpha31*agegt75[n]*male[n]
+ alpha5*c442GA[n] + alpha6*ex3_nt500[n] + alpha7*ex3_nt540[n]
logit(pr[n]) ← alpha0 + sumG[n] + sumW[n]
}
for (s in 1:43) { b[s] ~ dnorm(beta[s], sigma)
beta[s] ~ pi0+pi1*functional[s]}
alpha0 ~ dnorm(0,.001); alpha1 ~ dnorm(0,.001); alpha2 ~ dnorm(0,.001);
alpha3 ~ dnorm(0,.001); alpha4 ~ dnorm(0,.001); alpha5 ~ dnorm(0,.001);
alpha6 ~ dnorm(0,.001); alpha7 ~ dnorm(0,.001); alpha11 ~ dnorm(0,.001);
alpha21 ~ dnorm(0,.001); alpha31 ~ dnorm(0,.001); pi0 ~ dnorm(0,.001);
pi1 ~ dnorm(0,.001); sigma ~ dgamma(.01, .01); SDpi ~ 1/sqrt(sigma); }

```

ACKNOWLEDGEMENTS

The study was conducted by the GEM Study Group: Coordinating Center, Memorial Sloan-Kettering Cancer Center, New York, NY, U.S.A.: Marianne Berwick (PI, currently at the University of New

Mexico), Colin Begg (Co-PI), Irene Orlow (Co-Investigator), Urvi Mujumdar (Project Coordinator), Amanda Hummer (Biostatistician), Nandita Mitra (Biostatistician), Klaus Busam (Dermatopathologist), Pampa Roy (Laboratory Technician), Rebecca Canchola (Laboratory Technician), Brian Clas (Laboratory Technician), Javier Cotignola (Laboratory Technician), and Yvette Monroe (Interviewer).

Study Centers: The University of Sydney and The Cancer Council New South Wales, Sydney (Australia): Bruce Armstrong (PI), Anne Krickler (co-PI), and Melisa Litchfield (Study Coordinator). Menzies Research Institute, University of Tasmania, Hobart (Australia): Terence Dwyer (PI), Paul Tucker (Dermatopathologist), and Nicola Stephens (Study Coordinator). British Columbia Cancer Agency, Vancouver (Canada): Richard Gallagher (PI) and Teresa Switzer (Coordinator). Cancer Care Ontario, Toronto (Canada): Loraine Marrett (PI), Elizabeth Theis (Co-Investigator), Lynn From (Dermatopathologist), Noori Chowdhury (Coordinator), Louise Vanasse (Coordinator), and Mark Purdue (Research Officer). David Northrup (Manager for CATI). Centro per la Prevenzione Oncologia Torino, Piemonte (Italy): Roberto Zanetti (PI), Stefano Rosso (Data Manager), and Carlotta Sacerdote (Coordinator). University of California, Irvine (USA): Hoda Anton-Culver (PI), Nancy Leighton (Coordinator), and Maureen Gildea (Data Manager). University of Michigan, Ann Arbor (U.S.A.): Stephen Gruber (PI), Joe Bonner (Data Manager), and Joanne Jeter (Coordinator). New Jersey Department of Health and Senior Services, Trenton (USA): Judith Klotz (PI), Homer Wilcox (Co-PI), and Helen Weiss (Coordinator). University of North Carolina, Chapel Hill (USA): Robert Millikan (PI), Nancy Thomas (Co-Investigator), Dianne Mattingly (Coordinator), Jon Player (Laboratory Technician), and Chiu-Kit Tse (Data Analyst). University of Pennsylvania, Philadelphia, PA, U.S.A.: Timothy Rebbeck (PI), Peter Kanetsky (Co-Investigator), Amy Walker (Laboratory Technician), and Saarene Panossian (Laboratory Technician). *Consultants:* Harvey Mohrenweiser, University of California, Irvine, Irvine, CA, U.S.A.; Richard Setlow, Brookhaven National Laboratory, Upton, NY, U.S.A. Supported by the National Cancer Institute, Awards CA83180, CA131010, CA46592 and CA16086.

REFERENCES

- Hung RJ, Brennan P, Malaveille C, Porru S, Donato F, Boffeta P, Witte JS. Using hierarchical modeling in genetic association studies with multiple markers: application to a case-control study of bladder cancer. *Cancer Epidemiology, Biomarkers and Prevention* 2004; **13**(6):1013–1021.
- Aragaki C, Greenland S, Probst-Hensch N, Haile R. Hierarchical modeling of gene-environment interaction: estimating NAT2* genotype specific dietary effects on adenomatous polyps. *Cancer Epidemiology, Biomarkers and Prevention* 1997; **6**:307–314.
- Conti DV, Witte JS. Hierarchical modeling of linkage disequilibrium: genetic structure and spacial relations. *American Journal of Human Genetics* 2003; **72**:351–363.
- De Roos AJ, Rothman N, Inskip PD, Linet MS, Shapiro WR, Selker RG, Fine HA, Black PM, Pittman GS, Bell DA. Genetic polymorphisms in GSTM1, -P1, -T1, and CYP2E1 and the risk of adult brain tumors. *Cancer Epidemiology, Biomarkers and Prevention* 2003; **12**:14–23.
- Wolfinger R, O'Connell M. Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* 1993; **48**:233–243.
- Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993; **88**(421):9–25.
- Gelfand AE, Smith AFM. Sampling-based approaches to calculate marginal densities. *Journal of the American Statistical Association* 1990; **85**:398–409.
- Witte JS, Greenland S, Kim L, Arab L. Multilevel modeling in epidemiology with GLIMMIX. *Epidemiology* 2000; **11**(6):684–688.
- Greenland S. Second-stage least squares versus penalized quasi-likelihood for fitting hierarchical models for epidemiologic analyses. *Statistics in Medicine* 1997; **16**:515–526.
- Thomas DC, Langholz B, Clayton D, Pitkaniemi J, Tuomilehto-Wolf E, Tuomilehto J. Empirical Bayes methods for testing associations with large numbers of candidate genes in the presence of environmental risk factors, with applications to HLA associations in IDDM. *Annals of Medicine* 1992; **24**:387–391.
- Conti DV, Cortessis V, Molitor J, Thomas D. Bayesian modeling of complex metabolic pathways. *Human Heredity* 2003; **56**:83–93.
- Greenland S. A semi-Bayes approach to the analysis of correlated multiple associations, with an application to an occupational cancer-mortality study. *Statistics in Medicine* 1992; **11**:219–230.

13. Begg CB, Hummer AJ, Mujumdar U, Armstrong BK, Kricger A, Marrett LD, Millikan RC, Gruber SB, Anton-Culver H, Zanetti R, Gallagher RP, Dwyer T, Rebbeck TR, Busam K, From L, Berwick M, for the GEM Study Group. A design for cancer case-control studies using only incident cases: experience with the GEM study of melanoma. *International Journal of Epidemiology* 2006; **35**:756–764.
14. Goldstein AM, Chan M, Harland M, Gillanders EM, Hayward NK, Avril MF, Azizi E, Bianchi-Scarra G, Bishop DT, Bressac-de Paillerets B, Bruno W, Calista D, Albright LAC, Demenais F, Elder DE, Ghiorzo P, Gruis NA, Hansson J, Hogg D, Holland EA, Kanetsky PA, Kefford RF, Landi MT, Lang J, Leachman SA, MacKie RM, Magnusson V, Mann GJ, Niendorf K, Bishop JN, Palmer JM, Puig S, Puig-Butille JA, de Snoo FA, Stark M, Tsao H, Tucker MA, Whitaker L, Yakobson E, the Lund Melanoma Study Group, the Melanoma Genetics Consortium (GenoMEL). High risk melanoma susceptibility genes and pancreatic cancer, neural system tumors, and uveal melanoma across GenoMEL. *Cancer Research* 2006; **66**(20):9818–9828.
15. Sunyaev S, Ramensky V, Bork P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends in Genetics* 2000; **16**:198–200.
16. Sunyaev S, Ramensky V, Koch I, Lathe III W, Kondrashov AS, Bork P. Prediction of deleterious human alleles. *Human Molecular Genetics* 2001; **10**:591–597.
17. Ramensky V, Bork P, Sunyaev S. Human non-synonymous snps: server and survey. *Nucleic Acids Research* 2002; **30**:3894–3900.
18. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Research* 2001; **11**:863–874.
19. Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. *Genome Research* 2002; **12**:436–446.
20. Orlow I, Begg CB, Cotignola J, Roy P, Hummer AJ, Clas BA, Mujumdar U, Canchola R, Armstrong BK, Kricger A, Marrett LD, Millikan RC, Gruber SB, Anton-Culver H, Zanetti R, Gallagher RP, Dwyer T, Rebbeck TR, Kanetsky PA, Wilcox H, Busam K, From L, Berwick M, for the GEM Study Group. *CDKN2A* germline mutations in individuals with cutaneous malignant melanoma. *Journal of Investigative Dermatology* 2007; in press.
21. Berwick M, Orlow I, Hummer AJ, Armstrong BK, Kricger A, Marrett LD, Millikan RC, Gruber SB, Anton-Culver H, Zanetti R, Gallagher RP, Dwyer T, Rebbeck TR, Kanetsky PA, Busam K, From L, Mujumdar U, Wilcox H, Begg CB, the GEM Study Group. The prevalence of *CDKN2A* germ-line mutations and relative risk for cutaneous malignant melanoma: an international population-based study. *Cancer Epidemiology Biomarkers and Prevention* 2006; **15**(8):1520–1525.
22. Goodnight JH. Computing MIVQUE0 estimates of variance components. *SAS Technical Report R-105*, SAS Institute Inc., Cary, NC, 1978.
23. Schabenberger O. Introducing the GLIMMIX procedure for generalized linear mixed models. *SUGI 30*, SAS Institute Inc., Cary, NC, 2005.
24. Greenland S. Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary testing, and empirical-Bayes regression. *Statistics in Medicine* 1993; **12**: 717–736.
25. SAS Institute Inc. *The GLIMMIX Procedure. Manual*. SAS Institute Inc.: Cary, NC, 2005.
26. Zeger SL, Karim RM. Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* 1991; **86**:79–86.
27. Natarajan R, Kass RE. Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association* 2000; **95**:227–337.
28. Gilks W. Derivative-free adaptive rejection sampling for Gibbs sampling. *Bayesian Statistics*, vol. 4. Oxford University Press: Oxford, U.K., 1992; 641–665.
29. Gamerman D. Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing* 1997; **7**:57–68.
30. Spiegelhalter DJ, Thomas A, Best NG, Lunn D. *WinBUGS User Manual, version 1.4*. MRC Biostatistics Unit, 2003. Available at: www.mrc-bsu.cam.ac.uk/bugs.
31. Kenward M, Roger J. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997; **53**:983–997.
32. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992; **7**:457–511.
33. Heidelberger P, Welch PD. Simulation run length control in the presence of an initial transient. *Operations Research* 1983; **31**:1109–1144.
34. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 1998; **7**:434–455.

35. Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 2006; **1**:515–533.
36. Parmigiani G, Berry DA, Aguilar O. Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. *American Journal of Human Genetics* 1998; **62**.
37. Goldgar DE, Easton DF, Deffenbaugh AM, Monteiro ANA, Tavtigian SV, Couch FJ, the Breast Cancer Information Core (BIC) Steering Committee. Integrated evaluation of DNA sequence of unknown clinical significance: application to BRCA1 and BRCA2. *American Journal of Human Genetics* 2004; **75**:535–544.
38. Rebbeck TR, Martinez ME, Sellers TA, Shields PG, Wild CP, Potter JD. Genetic variation and cancer: improving the environment for publication of association studies. *Cancer Epidemiology, Biomarkers and Prevention* 2004; **13**:1985–1986.
39. Whittemore AS. Genetic association studies: time for a new paradigm? *Cancer Epidemiology, Biomarkers and Prevention* 2005; **14**:1359.
40. Wacholder S. Publication environment and broad investigation of the genome. *Cancer Epidemiology, Biomarkers and Prevention* 2005; **14**:1361.
41. Pharoah PDP, Dunning AM, Ponder BAJ, Easton DF. The reliable identification of disease-gene associations. *Cancer Epidemiology, Biomarkers and Prevention* 2005; **14**:1362.
42. Byrnes G, Gurrin L, Dowty J, Hopper JL. Publication policy or publication bias. *Cancer Epidemiology, Biomarkers and Prevention* 2005; **14**:1363.
43. Begg CB. Reflections on publication criteria for genetic association studies. *Cancer Epidemiology, Biomarkers and Prevention* 2005; **14**:1364–1365.
44. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002; **64**:583–639.