Published in final edited form as:

Biometrics. 2011 June ; 67(2): 371–380. doi:10.1111/j.1541-0420.2010.01469.x.

Hierarchical Modeling for Estimating Relative Risks of Rare Genetic Variants: Properties of the Pseudo-Likelihood Method

Marinela Capanu* and Colin B. Begg

Memorial Sloan-Kettering Cancer Center, 307 E 63rd St, 3rd Floor, New York, NY 10021

Summary

Many major genes have been identified that strongly influence the risk of cancer. However, there are typically many different mutations that can occur in the gene, each of which may or may not confer increased risk. It is critical to identify which specific mutations are harmful, and which ones are harmless, so that individuals who learn from genetic testing that they have a mutation can be appropriately counseled. This is a challenging task, since new mutations are continually being identified, and there is typically relatively little evidence available about each individual mutation. In an earlier article we employed hierarchical modeling (Capanu et al. 2008) using the pseudo-likelihood and Gibbs sampling methods to estimate the relative risks of individual rare variants using data from a case-control study and showed that one can draw strength from the aggregating power of hierarchical models to distinguish the variants that contribute to cancer risk. However, further research is needed to validate the application of asymptotic methods to such sparse data. In this article we use simulations to study in detail the properties of the pseudo-likelihood method for this purpose. We also explore two alternative approaches: pseudo-likelihood with correction for the variance component estimate as proposed by Lin and Breslow (1996) and a hybrid pseudo-likelihood approach with Bayesian estimation of the variance component. We investigate the validity of these hierarchical modeling techniques by looking at the bias and coverage properties of the estimators as well as at the efficiency of the hierarchical modeling estimates relative to that of the maximum likelihood estimates. The results indicate that the estimates of the relative risks of very sparse variants have small bias, and that the estimated 95% confidence intervals are typically anti-conservative, though the actual coverage rates are generally above 90 per cent. The widths of the confidence intervals narrow as the residual variance in the second-stage model is reduced. The results also show that the hierarchical modeling estimates have shorter confidence intervals relative to estimates obtained from conventional logistic regression, and that these relative improvements increase as the variants become more rare.

Keywords

hierarchical models; pseudo-likelihood; Bayesian; genetic risk; rare variants

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

*capanum@mskcc.org.

Supplementary Materials

Web Appendix, Tables, and Figures referenced in Sections 2.3 and 3 are available under the Paper Information link at the Biometrics website <http://www.biometrics.tibs.org>.

disease etiology. The challenge lies in determining which genetic variants are most likely to be deleterious. This plays a critical role in identifying effective approaches to prevention and treatment, and allows genetic counselors to provide informed choices for interventions to reduce the risk of the disease. The focus of this article is the disease impact of variants that occur within a gene that is known to have a strong effect on disease risk, and our motivation is our involvement in major epidemiologic studies in melanoma [the Genes, Environment and Melanoma, GEM Study, Begg et al. (2006)] and breast cancer [Women's Environmental, Cancer, and Radiation Epidemiology, WECARE Study, Bernstein et al. (2004)]. In both studies, data have been collected on large numbers of cases and controls, and the germline DNA from these subjects has been sequenced for variants in major genes known to affect these diseases, notably *CDKN2A* for melanoma, and *BRCA1* and *BRCA2* for breast cancer.

Although many studies have provided important evidence of the cancer impact of these genes in aggregate (Bishop et al. 2006; Stratton and Rahman 2008), the problem is that in practice, when an individual in the population is screened and is discovered to be a carrier, the individual is a carrier of a specific variant. Thus, this individual's disease risk, and the risks of relatives who also turn out to be carriers, is dependent on the disease impact of the specific variant. Historically, evidence about individual variants has usually emerged from families, where a cluster of cancer occurrences have been observed, and the specific genetic variant that is represented in the family can be identified. Increasingly, evidence now accrues from large association (case-control) studies where rare variants are, by definition, observed very infrequently. On their own, these data are insufficient to provide meaningful risk predictions. However, by aggregating the results from individual rare variants on the basis of characteristics shared by groups of variants, using hierarchical modeling, it is possible in principle to leverage much more accurate predictions about the impact of the individual variants on disease risk.

Hierarchical modeling has been used previously in association studies (see, e.g., Hung et al. 2004; Aragaki et al. 1997; De Roos et al. 2003, among others), for fine-scale linkage-disequilibrium (LD) mapping of disease genes (Conti and Witte 2003), and to select single nucleotide polymorphic variants (SNPs) to test in subsequent stages of genomewide association studies (Lewinger et al. 2007). However, these previous applications have only dealt with SNPs with high minor allele frequency. Other related approaches have involved pedigree data, rather than data from a case-control study (Goldgar et al. 2004; Zhou et al. 2005).

In earlier work we demonstrated the feasibility of using hierarchical modeling based on pseudo-likelihood estimation to obtain the relative risks of the individual rare variants (Capanu et al. 2008). However, we acknowledged that the validity of this approach is uncertain due to the application of asymptotic methods to sparse data. In the present article we seek to address this issue. Can hierarchical modeling be used to obtain estimates of these relative risks with acceptable statistical properties? What are the practical impediments to utilizing this methodology? In our earlier article we also analyzed the data using a Bayesian approach with Gibbs sampling, as a comparator to the pseudo-likelihood method. Although this approach is conceptually appealing in that it does not rely on asymptotic approximations, it is exceptionally time-consuming in the context of the kinds of datasets under investigation. Consequently the focus of our work is the validity of the much faster pseudo-likelihood approach. We conduct simulations to examine the statistical properties of the relative risk estimators of individual rare variants, estimated from a hierarchical model using the pseudo-likelihood method and we also examine two modifications. In the first, we use an adaptation of Lin and Breslow's correction of the variance component of a generalized linear mixed model (Lin and Breslow 1996). In the second, we introduce a

computationally efficient hybrid approach involving use of a Markov chain method to estimate the variance component of the random effects representing the risk estimates of the individual variants and then conduct pseudo-likelihood estimation of the fixed and random effects using this estimated random effects variance.

The bias and coverage probabilities of the estimators are examined as a function of factors such as the overall sample size, the numbers of cases and controls observed to carry the variant, the total number of variants observed, the strength of the higher level predictors, the number of second-stage predictors, and the residual variance of the higher level model. We apply the methods to the melanoma data from the GEM Study. Section 2 introduces the hierarchical model under investigation and presents the three estimation methods. Section 3 describes the simulation study. Application to the melanoma data is provided in Section 4, and concluding remarks are given in Section 5.

2. Methods

The fundamental data configuration under investigation is a case-control study with N participants, with the case and control status of each participant denoted by the $N \times 1$ vector Y . In addition to a set of subject specific covariates, captured by the design matrix W and corresponding regression parameters γ , each participant is characterized by the presence or absence of each of a set of n genetic variants within the gene under investigation. The constellation of combinations of these variants in the participants is contained in the design matrix X , and β represents an $n \times 1$ vector of parameters representing the log relative risks conferred by each of the variants. Because the variants are typically rare, they are each observed in only a very few participants, in many cases only once. The ultimate objective of our analysis is to estimate the component log relative risks in the vector β , and the individual confidence intervals of these estimates. The goal of this article is to assess the validity of these estimates and confidence intervals.

The model we use throughout combines a first-stage logistic regression with a second-stage linear regression that characterizes the individual variants in terms of attributes that could distinguish their association with disease. Specifically, in the first-stage model, the outcome vector Y is related to β and γ through the logistic regression model:

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

In the second stage the individual log relative risks of the variants (β) are modeled as a function of biological features of the variant that may indicate the likely relevance of the variant to the functioning of the gene (for example, bioinformatic tools such as SIFT (Ng and Henikoff 2001), PolyPhen (Sunyaev et al. 2000), or Grantham scores (<http://agvgd.iarc.fr/>)). The second-stage model is defined by

$$\beta = Z\pi + \delta, \quad \text{with} \quad \delta \sim N(0, \tau^2 I_n), \quad (2)$$

where the $n \times p$ matrix Z captures the hierarchical information about the variants, π is a $p \times 1$ column vector of the second-stage parameters indicating the aggregated risk of variants categorized on the basis of the second-stage predictors, and δ is a vector of normally distributed 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. Here I_n denotes the $n \times n$ identity matrix.

The use of a linear second-stage model allows us to combine the first and second-stage models into a mixed effects logistic model. This is easily shown by substituting (2) into (1):

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

where γ and π are vectors of fixed effects, and δ is a random vector whose elements are normally distributed with zero mean and common variance τ^2 .

This model forms the basis of all our simulations. However, the properties of the model for evaluating rare variants is strongly dependent on our approach for parameter estimation. The estimation method also greatly affects computational speed, which in turn affects our ability to analyze data sets flexibly with respect to selection of covariates, subset analyses, and so forth. In the following, we describe three distinct estimation methods.

2.1 The Pseudo-Likelihood Method

The pseudo-likelihood method estimates the model parameters of a generalized linear mixed model (GLMM) using an iterative linearization technique which employs Taylor expansions to approximate the model with a model based on pseudo-data with fewer nonlinear components. The linearization method first constructs a pseudo-model and pseudo-data by using an expansion around the common estimate of the best linear unbiased predictors of the random effects. 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. More details of this algorithm and additional formulas are provided in Wolfinger and O'Connell (1993) and the Web Appendix, and a description of its application to fit hierarchical models can be found in Capanu et al. (2008). In applying this algorithm to the GLMM in (3), one obtains estimates for τ^2 as well as for π , δ , γ along with their corresponding covariance matrices. The coefficients for the individual variants can be estimated as $\beta = Z\pi + \delta$, which can then be easily used to derive the covariance matrix estimate and 95% confidence intervals for β (Witte et al. 2000).

The penalized quasi-likelihood method (PQL) of Breslow and Clayton (1993) is an alternative method for estimating generalized linear mixed models which involves approximations based on Laplace's method. Though arrived at through different expansions, the PQL and the pseudo-likelihood methods are in fact equivalent methods in the sense that the objective functions minimized by the two methods differ by a constant, thus leading to identical parameter estimates (Wolfinger and O'Connell 1993). It has been shown that the PQL estimate of the variance component (τ^2) is subject to serious bias especially for certain cases such as clustered binary data with clusters of small size. In view of this, the pseudo-likelihood estimate of τ^2 must be subject to the same bias, and our simulations of the pseudo-likelihood method verify this fact. However, accurate estimation of τ^2 is important since the estimation of the fixed and random effects, and their corresponding standard errors, are strongly dependent on τ^2 . We have performed simulations of the pseudo-likelihood method in which τ^2 was pre-specified at the true value, and we obtained very satisfactory results with low bias of the estimated relative risks of the individual rare variants and confidence intervals with nominal coverage (data not shown). These results emphasize the key role that estimation of τ^2 plays in the success of the model. In recognition of this issue we have examined two alternative estimation methods that are designed to improve estimation of τ^2 .

2.2 Lin and Breslow correction for the variance component

Breslow and Lin (1995) and Lin and Breslow (1996) proposed a correction of the PQL estimate of the variance component and showed that the correction can improve the performance of the PQL method for models with moderate values of τ^2 (τ^2 between 0.5 and 1 on the log relative risk scale). The corrected PQL estimate of the variance component is of the form $\hat{\tau}_{cp}^2 = C_p \hat{\tau}_p^2$, where $\hat{\tau}_p^2$ is the restricted maximum likelihood PQL estimate, while the correction term C_p is derived based on a series of linear expansions and depends on the fixed effects and random effects design matrices and on the variance functions associated with the generalized mixed model. In our simulations, we apply the Lin and Breslow correction to the pseudo-likelihood method.

2.3 Hybrid Bayesian estimation of the variance component

Besides bias in the estimation of the variance component, another problem with the pseudo-likelihood method is that it can produce negative estimates of the variance component (which are typically replaced with zero by some statistical software packages including SAS). This problem can occur in the estimation of any linear mixed model. Negative estimates are especially likely to occur for datasets with only a few observations per random effect category and these necessarily indicate underestimation of a true positive (or zero) variance component. If the pseudo-likelihood estimate of the variance component is zero, the method acts as if Y follows a fixed effects logistic model, and the random effects are estimated deterministically from the estimates from the 2nd level regression parameters. Thus, variants that share a set of categorical bioinformatic predictors would have the same estimated risk. In some configurations of our simulations, we observed a zero estimate almost forty percent of the time (see Web Table 1). In these cases, a Bayesian procedure may produce more sensible results (Hulting and Harville 1991; Harville and Carriquiry 1992). We investigated a hybrid approach with initial (non-zero) Bayesian estimation of the variance component followed by pseudo-likelihood estimation of the fixed and random effects using the Bayesian variance component estimator. As seen earlier in this section, the pseudo-likelihood method approximates the original generalized linear mixed model with a weighted linear mixed model. This weighted linear mixed model can then be estimated by restricted maximum likelihood estimation using conventional software for linear mixed models. Thus, we propose a three-step procedure: in step 1, we apply the pseudo-likelihood method; for the approximated linear mixed model obtained in step 1, we estimate τ^2 in step 2 using a Markov chain algorithm; finally, in step 3 we re-estimate the regression parameters using the pseudo-likelihood method with τ^2 pre-specified at its Bayesian estimate.

To achieve the Bayesian estimation in step 2, we generate a posterior sample from the marginal posterior density of τ^2 using a random walk Metropolis-Hastings algorithm with a non-informative Inverse-Gamma distribution with shape and scale parameters both set at 0.01, and use the posterior mean as the estimator. This step can be conducted using the PRIOR statement in PROC MIXED SAS, while step 3 can be implemented using the %glimmix macro with the option “gdata” which allows one to prespecify the variance-covariance matrix of the random effects (for details, see the SAS code provided in the Web Appendix). Note that for computational simplicity, we omitted from the Bayesian estimation subjects who did not possess one (or more) of the rare variants under investigation (see the SAS code provided in the Web Appendix for more details). As seen in (2), the second-stage model pertains only to genetic variants and the residual variance τ^2 measures the variation in effects between these variants; thus the wild type subjects (i.e. subjects who do not possess any of the rare variants) do not contribute to the variance parameter τ^2 .

3. Simulation Study

3.1 Description of GEM and WECARE Datasets

In examining the validity of hierarchical modeling in context, we opted to construct simulations that reflect closely two motivating examples: the GEM Study of melanoma and the WECARE Study of breast cancer. The GEM Study comprised 1189 individuals with multiple primary melanoma (who served as cases) and 2424 with single primary melanoma (who served as controls) screened for *CDKN2A* mutations. Sequencing identified 43 individual mutations, most of which were observed very infrequently, even in this study of over 3000 incident melanoma patients. Specifically, of the 43 variants, 20 were carried by a single subject, 13 were carried by 2 subjects, 4 occurred in 3 subjects, 6 occurred in at least 4 subjects, and 3 occurred in 7 or more subjects. We analyze this study in depth in Section 4.

The WECARE Study involved 705 cases, women with asynchronous contralateral breast cancer (CBC), and 1398 controls with unilateral breast cancer. Screening for *BRCA1* and *BRCA2* mutations identified 470 unique sequence variants, out of which 185 were missense variants (see Borg et al. 2010, for more details). Our hierarchical model is restricted to missense variants since in practice, the most accessible characteristics of variants are bioinformatic predictors typically available only for missense variants. Other variants such as truncating mutations or insertions are aggregated and classified as confounders in the first-stage model. Of the 182 rare missense variants (with $MAF \leq 2.5\%$), 115 variants (63%) were carried by only one subject, 29 were carried by 2 subjects, 11 were carried by 3 subjects, 6 were carried by 4 participants, 7 were carried by 5 to 9 subjects, and 14 were carried by 10 or more subjects.

3.2 Fundamental Simulation Strategy

For each proposed scenario we simulate data for 1000 repetitions of the case-control study. This results in a standard error of the simulated coverage rate for each 95% confidence interval of less than 0.7% when the true coverage rate is 95%. We generate datasets that mirror exactly the configurations of the GEM or WECARE data; specifically, the total number of participants, the total number of variants, and the relative frequencies with which a variant occurs are set to the corresponding values from the motivating studies. That is, N (the number of patients), n (the number of variants), the $N \times n$ design matrix X that indicates which variants are possessed by each patient, and the matrix W containing the patient factors, are chosen to be identical to those observed in the motivating studies and are kept fixed throughout the simulations. The parameters γ representing the effects of the patient covariates are set at the corresponding estimated values in the motivating datasets.

For each simulated dataset in the scenario under investigation we generate an $N \times 1$ vector $Y = \{y_i\}_i$ of binary indicators of case-control status given by

$$y_i = \begin{cases} 1 & \text{if } u \leq p_i \\ 0 & \text{otherwise} \end{cases}$$

where u is a uniformly distributed random number between 0 and 1, and $p_i = 1/(1 + \exp \{-(\alpha_0 + x_i\beta + w_i\gamma)\})$ represents the probability that the i^{th} subject with a given configuration of genetic variants (x_i , the i^{th} row of the matrix X) and a particular set of confounders (w_i , the i^{th} row of the matrix W) is a case. Note that $\alpha_0 = -(\log 2 + \bar{x}\beta + \bar{w}\gamma)$ is chosen so that the ratio of cases to controls matches the ratio of cases to controls in the motivating study. Here $\bar{x}$ and $\bar{w}$ denote the row vectors containing the column means of the matrices X and W ,

respectively. For each scenario the true log relative risks of the individual variants are randomly generated as $\beta \sim N(Z\pi, \tau^2 I_n)$.

Our strategy is based on the premise that the informativeness of the hierarchical model is driven by two critical factors, and in our simulations we systematically vary these two factors. The first is the predictive strength of the hierarchical covariates, represented by the hierarchical parameters π . The second is the residual variation in the second-stage model, i.e. the extent to which individual variants with the same hierarchical covariates differ in their disease relative risks. This is represented by the parameter τ^2 . Initially we used a single binary variable to represent the predictive strength, and so $\pi = (\pi_0, \pi_1)$. In this setting the Z matrix of hierarchical covariates contains 2 columns, first a column of ones corresponding to the second-stage model intercept, and second a column corresponding to a random Bernoulli variable with success probability of 0.75. The latter is chosen to mimic the GEM Study in which an important 2nd stage predictor was a binary variable indicating whether the variant was functional or not and approximately 75% of the variants were functional. For comparison purposes, for the WECARE data we use the same success probability as for the GEM data. Varying the success probability of the second-stage Bernoulli covariate did not result in notable changes (data not shown). The Bernoulli variable is generated separately for each simulation. The intercept of the 2nd stage model, π_0 , is kept fixed at the estimated value from the data set (0.002 for GEM and -0.066 for WECARE).

To evaluate the performance of the estimators as a function of π_1 and τ^2 we vary these parameters over a range of values that could commonly arise in genetic epidemiologic studies. We set $\pi_1 = 0.8, 1.4$, and 1.8 which correspond to relative risks of 2.2, 4, and 6. In our model these represent the extent to which risks of harmful variants on average exceed the risk in wild type subjects. The parameter τ was set to $\tau = 0.4, 0.8, 1.2$. As an example, if $\tau = 0.4$ then 95 percent of the true relative risks of the variants with the same bioinformatic characteristics will fall within almost fivefold range of one another (since $\exp(2 \cdot 1.96 \cdot \tau) = 4.8$). Similarly, $\tau = 0.8$ and 1.2 would correspond to the relative risks being within 23 and 110 fold of one another, respectively. Note that here and throughout the article, the variance τ^2 is expressed in standard deviation units τ .

To explore the effect of adding multiple second-stage covariates, we also generated second-stage models with 2 and 3 predictors, each binary indicators with success probability of 0.5 and coefficients of strength 1.4. To study the effects of larger numbers of variants, n , we performed simulations in which we increased the number of variants by a predetermined factor f and increased the overall sample size by the same factor while keeping the number of occurrences of each variant rare as in the motivating datasets. In our simulations we use $f = 2$ and $f = 3$. For example, in the case of the GEM data $f = 2$ would correspond to twice as many variants (86) variants and twice as many subjects (7226). We also evaluate the validity of the approach under two forms of misspecification of the 2nd stage model. In the first, the fitted model omits the 2nd stage predictor, and in the second the covariance structure possesses correlation between the relative risks of the variants.

3.3 Simulation Results

We evaluated the operating characteristics of the three hierarchical modeling estimation techniques described in Section 2: the pseudo-likelihood method (denoted “PL” in the tables), the pseudo-likelihood method with Breslow and Lin’s correction (denoted “BLPL”), and the hybrid Bayesian pseudo-likelihood approach (denoted “BPL”). Table 1 compares the three methods using the GEM data simulations with the coefficient of the 2nd stage predictor assumed to be $\pi_1 = 1.4$. For each of the three techniques, the table presents the mean bias, the mean coverage rates, and the mean lengths of 95 per cent confidence intervals of the log relative risks of the variants stratified by variant frequency (i.e., averaged

over all variants observed once in the dataset, then over variants observed twice, three times, and so forth). The table also reports the relative mean lengths of the confidence intervals, i.e. the mean length of the interval obtained from the hierarchical model divided by the mean length of the interval for the estimates from ordinary logistic regression, using only those variants for which an MLE estimate is defined, i.e. variants occurring in at least one case and one control. For example, using the pseudo-likelihood method and assuming $\tau = 0.8$, the mean bias of the log relative risks of variants carried by a single subject (top row) was -0.02 , the 95% confidence intervals had a mean coverage of 81% and a mean width of 2.60. For variants carried by 2 subjects, the confidence intervals from hierarchical modeling were on average 58% shorter than the corresponding maximum likelihood estimated confidence intervals (the ratio mean width was 0.42).

Overall, the table indicates that the methods are comparable in terms of bias of the log relative risks of variants, all exhibiting small biases, behavior which was also seen by Witte and Greenland (1996) in their hierarchical modeling simulations. The hybrid approach (BPL) has better coverage properties than both the other alternatives. For example, for $\tau = 0.8$, coverage of 95 per cent confidence intervals for variants carried by 6 subjects or less improves from 81–84% for the pseudo-likelihood method, to 83–84% for the pseudo-likelihood with Breslow and Lin's correction, and further to 89–91% for the Bayesian hybrid approach. These results are largely a consequence of reduced bias in the estimate of τ for the hybrid approach (see Web Table 1). However, bias in τ increases for the hybrid approach when τ is small, since the PL estimate of τ is often zero in this setting but the BPL always supplies a positive estimate. This reduction in the bias of τ is apparent for $\tau = 1.2$ but not for $\tau = 0.4$. A possible explanation is that for values such as $\tau = 0.4$ the pseudo-likelihood estimates of τ are zero very often (more than 36% of time), and as a result the hybrid approach has a tendency to overestimate τ in this setting, since it always substitutes positive estimates in these cases. Greenland (1997) observed similar behavior when one of the methods he investigated appeared to produce nearly unbiased estimates for the residual variance due to a high rate of zero estimates of τ .

For all three methods, regardless of the degree of informativeness of the second level model (i.e., regardless of the values of τ and π_1), the lengths of the confidence intervals decrease when the number of occurrences of the variants is increased, as one would expect. Moreover, for smaller values of τ these widths are further reduced, indicating smaller standard errors of the relative risks with decreased τ values. This makes sense, since a small τ provides stronger evidence for aggregation of the results from different variants with similar hierarchical characteristics. However the widths are not substantially affected by varying the strength of the higher level predictor, π_1 (see Web Tables 2–4 where simulations from Table 1 are expanded to explore the effect of π_1 on the results). Similarly, the mean lengths of hierarchical modeling confidence intervals relative to ML intervals (the column denoted "Ratio Width") decrease as τ decreases, and they are not noticeably affected by the degree of predictiveness of the higher level covariate, π_1 . As expected, all intervals obtained from the hierarchical model are shorter than the maximum likelihood estimates, and this improvement is most evident for the very sparse variants. All three methods exhibited similar root mean squared error which decreased with smaller values of τ (see Web Table 5). These results confirm that widths of confidence intervals are appropriate measures of performance for methods with small bias. Note also that the number of times the trials resulted in a zero estimate of τ increases with smaller values of τ . For example, for $\tau = 1.2$, the residual variance was estimated to be 0 around 4% of times; this number increases to about 16% for $\tau = 0.8$, and to around 36% for $\tau = 0.4$ (see Web Table 1).

In what follows, we focus only on the results from the hybrid pseudo-likelihood Bayesian approach, given its superior properties, and include results for the other two techniques as

Web Tables. Table 2 presents simulations that follow the configuration of the GEM data (left panel) and WECARE data (right panel), and include a range of values of π_1 and τ . For both the GEM and WECARE datasets, the estimators continue to exhibit only small biases which do not appear to be dependent on the rarity of the variants. The bias seems to increase for stronger second-stage predictors (larger values of π_1), but decrease with smaller values of τ . Web Figures 1 and 2 plot the estimated versus the true log relative risks for each of the 43 variants from the GEM data, in settings where the variants are randomized to be functional and non-functional, respectively. There is little evidence that the magnitude of the relative risk affects the bias. The coverage rates in Table 2 are improved (more so for the GEM data) as the second-stage model becomes better specified, (i.e. τ decreases), and become close to the nominal level of 95 per cent for $\tau = 0.4$. The anti-conservativeness of the coverage probabilities is improved with increasing frequency of occurrences of the variants. Thus, for the GEM data with $\tau = 1.2$ and $\pi_1 = 1.8$, the coverage improves from 86% for variants observed only once in the dataset to 92% in variants observed at least 7 times. For a fixed τ , varying the strength of the higher level covariate, π_1 , does not seem to noticeably change the coverage rates. Note however that when the 2nd stage model contains no other covariate besides the intercept (i.e. $\pi_1 = 0$) the coverages are maintained, but the bias of the relative risk estimates is considerably increased especially for the more rare variants and for larger values of τ (see Web Table 4 for the GEM data).

An important aspect of the simulations is to assess how the performances of the estimators are affected by the number of variants in the model. We have conducted simulations in which we have doubled or tripled the number of subjects in the GEM Study and correspondingly doubled or tripled the number of variants, and we observed similar behavior as that shown in Table 2 (data not shown). Witte and Greenland (1996) varied the number of second-stage covariates, rather than their predictive strength, and found that the coverage rates improved with larger numbers of second-stage covariates, though the confidence intervals became wider; this improvement was greatest especially for models with large numbers of random effects. We have examined the effect of a larger number of second-stage covariates by including separately 2 and 3 covariates in the higher level model (each a binary indicator with success probability of 0.5 and predictive strength coefficient of $\pi_1 = 1.4$). We too observed better coverage rates but wider confidence intervals with the addition of more covariates (Table 3). For a second-stage model with 3 predictors, the coverage rates were close to the nominal level for $\tau = 0.8$ and $\tau = 0.4$ and were at 90 per cent or above for $\tau = 1.2$.

We evaluated the validity of the BPL approach under two forms of misspecification. Under a model that omits an important second-stage covariate ($\pi_1 = 1.4$), the biases of the relative risk estimates, the widths of the confidence intervals as well as the ratio of the widths of the confidence intervals relative to the MLE are similar to those obtained from a true model that contains a second-stage covariate. However there is a noticeable decrease in coverage especially for smaller values of τ (see Web Table 6 in comparison to Web Table 4, $\pi_1 = 1.4$). Under the second type of misspecification in which the log relative risks of the variants were correlated with each other with common correlation coefficient $\rho = 0.25$ but are assumed to be independent, the estimates from the model that ignored this correlation did not suffer of an increase in bias or a loss of coverage (see Web Table 6).

4. Identifying Harmful Variants in the P16 Gene

The risk of melanoma is known to be substantially increased in families with mutations in the tumor suppressor gene *CDKN2A*. The GEM study of melanoma (described earlier in Section 3.1) is a case-control study of multiple primary melanoma designed to estimate the population relative risk overall, and to examine the impact of individual variants in this

gene. In an earlier article (Capanu et al. 2008) we analyzed the data using the PL method and compared the results with Gibbs sampling. For variants that occurred in at least one case and at least one control, maximum likelihood estimates from an ordinary logistic regression were also supplied. The primary results are reproduced in the 1st, 2nd, and 5th results columns of Table 4. In the other two result columns we provide estimates obtained with the Breslow and Lin method and the hybrid approach.

The table groups the variants in terms of their observed frequencies and includes all variants observed on at least three patients and a representative selection of the remaining variants observed only once or twice in the dataset. The hierarchical covariate is the “functional” status, indicating “yes” if the variant occurs in one of the coding exons and is either a missense single nucleotide substitution, or an insertion or a deletion. Besides the 43 indicators for the genetic variants, the first-stage model also contained potential confounder variables age, sex, and age by sex interaction, along with three common polymorphisms (see Capanu et al. 2008, for more details).

For most of the variants, the hybrid approach produces results closer to Gibbs sampling, which supports our simulation findings that the hybrid approach has better statistical properties than the other two competitors PL and BL. However, we note that the Gibbs sampling approach takes 4 minutes per iteration in WinBUGS (1.35 minutes per iteration in OpenBUGS on Linux) making it prohibitive to allow it to run for a large number of iterations to ensure convergence of the chain, while the hybrid approach converges and produces results in a few seconds. The results show that variants that are “functional” are typically assigned higher relative risks unless there is strong empirical evidence to contradict this, as there is, for example, for c.8_9ins, where all 6 patients with this variant are controls. The analysis appears to show that seemingly convincing conclusions about individual rare variants are possible, for example for c.334C>G and c.95T>C, both of which occur only twice. Note also that for the two most common variants (c.50+37G/C and (-)33G>C), all methods produce the same relative risks estimates and similar confidence intervals. However, for the rest of variants with a finite MLE, ordinary logistic regression produced more extreme estimates and confidence intervals than the other methods.

5. Discussion

We have conducted an extensive investigation to examine the small-sample properties of hierarchical modeling based on pseudo-likelihood estimation in the context of the analysis of very sparse random effects, an application that is motivated by epidemiologic case-control studies of major cancer genes. The factor that emerges as being the most influential in the performance of this approach is the estimate of the residual variance of the second-stage model. It is well known that the pseudo-likelihood method is subject to bias in the estimation of this variance component especially for clustered binary data with few observations per cluster. In the context of the motivating examples, the results from our simulations suggest that the pseudo-likelihood technique suffers from this limitation, with underestimated τ values resulting in anti-conservative confidence intervals for the hierarchical modeling estimates. Another shortcoming of this approach is that for certain configurations a zero estimate of τ occurs very frequently. Simulations in which τ was pre-specified at the true value indicate that the hierarchical modeling estimates enjoy good properties with low bias and coverage rates close to the nominal level. This emphasizes the important role that the estimation of the random effects variance has in the success of this hierarchical modeling approach. We have thus focused our attention on methods that provide a more accurate estimate of the residual variance. We explored the bias correction technique proposed by Lin and Breslow (1996), as well as a hybrid Bayesian three-step approach that is designed to

better handle settings in which τ is frequently estimated to be zero. The hybrid approach seems to perform better overall than the other two competitors.

From a broad perspective, our results are encouraging. Coverage rates with the hybrid approach were close to the nominal level when τ is small, though they tended to be anti-conservative for other configurations with larger residual variances. More specifically, the method seems to underestimate τ for larger values of τ and overestimate it for smaller values, with underestimated τ values resulting in anti-conservative confidence intervals for the hierarchical modeling estimates. Confidence intervals become narrower as the number of carriers of an individual variant increases, and they further narrow in settings where τ is small. Hierarchical modeling estimates are characterized by shorter confidence intervals than those obtained by conventional logistic regression and these improvements are most evident for the rarest variants. For the scenarios investigated, varying the second-stage predictor, represented by the parameter π_1 , did not result in substantial changes in the operating characteristics of the method. However, in interpreting the effect of π_1 on the results one should note that the strength of the second-stage predictor is inherently linked to the second-stage residual variation represented by τ^2 in that with a stronger second-stage predictor you would expect better specified models with smaller residual variance, whereas for models with weak predictors, the residual variance is expected to be higher.

We explored how the presence of multiple second-stage covariates impacts the performance of the estimators. Simulations with 2 and 3 second-stage predictors showed improved coverage rates. With 3 covariates, the coverage rates were close to the nominal level for the smaller residual variances and were above 90 per cent for the larger values of τ . These results offer encouragement, since with future improvement of bioinformatic tools to predict functional relevance, the number of available second-stage covariates may well increase and this in turn will increase our ability to predict the risk conferred by the individual rare variants. When the model was misspecified, the effect of unrecognized correlation between the risks of the variants did not affect the performance of the hybrid approach. However omission of an important 2nd stage covariate resulted in loss of coverage for smaller residual variances.

Given the daunting task of estimating relative risks of such sparse variants, the proposed hierarchical modeling hybrid Bayesian approach is promising, with low bias and with coverage rates for the individual estimates falling under 90 per cent only for the more sparse variants in models with larger residual variance. The method is straightforward to use and it provides a considerable efficiency boost over conventional logistic regression (which at the outset can only be used for variants that are carried by at least one case and one control in the study data). However, users of this methodology need to interpret confidence intervals for rare variants with some caution, as they may underestimate the uncertainty in the results.

Acknowledgments

We thank Duncan C. Thomas, Mithat Gönen, and the reviewers for their insightful suggestions. We also thank the investigators of the GEM Study (PI Marianne Berwick) and the WECARE Study (PI Jonine L. Bernstein) for use of their data. Supported by the National Cancer Institute, Awards CA131010, CA83180, and CA097397.

References

- 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 & Prevention*. 1997; 6:307–314.
- Begg CB, Hummer AJ, Mujumdar U, Armstrong BK, Kricker 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. [PubMed: 16556646]
- Bernstein JL, Langholz B, Haile RW, Bernstein L, Thomas DC, Stovall M, Malone KE, Lynch CF, Olsen JH, Anton-Culver H, Shore RE, Boice JD Jr, Berkowitz GS, Gatti RA, Teitelbaum SL, Smith SA, Rosenstein BS, Borresen-Dale AL, Concannon P, Thompson WD. Study design: evaluating gene-environment interactions in the etiology of breast cancer - the Wecare Study. *Breast Cancer Research*. 2004; 6:R199–R214. [PubMed: 15084244]
- Bishop JN, Harland M, Bishop DT. The genetics of melanoma. *British Journal of Hospital Medicine*. 2006; 67:299–304. [PubMed: 16821732]
- Borg A, Haile RW, Malone KE, Capanu M, Diep A, Törngren T, Teraoka S, Begg CB, Thomas DC, Concannon P, Møller M, Bernstein L, Tellhed L, Xue S, Olson ER, Liang X, Dolle J, Børresen-Dale AL, Bernstein JL. Characterization of *BRCA1* and *BRCA2* deleterious mutations and variants of unknown clinical significance in unilateral and bilateral breast cancer: the WECARE study. *Human Mutation*. 2010; 31:E1200–E1240. [PubMed: 20104584]
- Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*. 1993; 88:9–25.
- Breslow NE, Lin X. Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*. 1995; 82:81–91.
- Capanu M, Orlov I, Berwick M, Hummer AJ, Thomas DC, Begg CB. The use of hierarchical models for estimating relative risks of individual genetic variants: an application to a study of melanoma. *Statistics in Medicine*. 2008; 27:1973–1992. [PubMed: 18335566]
- Conti DV, Witte JS. Hierarchical modeling of linkage disequilibrium: genetic structure and spacial relations. *American Journal of Human Genetics*. 2003; 72:351–363. [PubMed: 12525994]
- 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 & Prevention*. 2003; 12:14–23.
- 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 variants of unknown clinical significance: Application to *BRCA1* and *BRCA2*. *American Journal of Human Genetics*. 2004; 75:535–544. [PubMed: 15290653]
- Greenland S. Second-stage least squares versus penalized quasi-likelihood for fitting hierarchical models for epidemiologic analyses. *Statistics in Medicine*. 1997; 16:515–526. [PubMed: 9089960]
- Harville DA, Carriquiry AL. Classical and Bayesian prediction as applied to an unbalanced mixed linear model. *Biometrics*. 1992; 48:987–1003.
- Hulting FL, Harville DA. Some Bayesian and non-Bayesian procedures for the analysis of comparative experiments and for small-area estimation: computational aspects, frequentist properties, and relationships. *Journal of the American Statistical Association*. 1991; 86:557–568.
- 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 & Prevention*. 2004; 13:1013–1021.
- Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genetic Epidemiology*. 2007; 31:1–12. [PubMed: 17096343]
- Lin X, Breslow NE. Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*. 1996; 91:1007–1016.
- Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Research*. 2001; 11:863–874. [PubMed: 11337480]
- Schabenberger, O. Introducing the GLIMMIX procedure for generalized linear mixed models. SAS Institute Inc; Cary, NC, SUGI 30: 2005.
- Stratton MR, Rahman N. The emerging landscape of breast cancer susceptibility. *Nature Genetics*. 2008; 40:17–22. [PubMed: 18163131]
- Sunyaev S, Ramensky V, Bork P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends in Genetics*. 2000; 16:198–200. [PubMed: 10782110]

- Witte JS, Greenland S. Simulation study of hierarchical regression. *Statistics in Medicine*. 1996; 15:1161–1170. [PubMed: 8804145]
- Witte JS, Greenland S, Kim L, Arab L. Multilevel modeling in epidemiology with GLIMMIX. *Epidemiology*. 2000; 11:684–688. [PubMed: 11055630]
- Wolfinger R, O’Connell M. Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*. 1993; 48:233–243.
- Zhou X, Iversen E Jr, Parmigiani G. Classification of missense mutations of disease genes. *Journal of the American Statistical Association*. 2005; 100:51–60. [PubMed: 18418466]

Table 1

The Pseudo-Likelihood method, Breslow and Lin correction, and a hybrid approach for the GEM data.

Method ^a	Number of Carriers	$\tau = 1.2$				$\tau = 0.8$				$\tau = 0.4$			
		Bias	Coverage	Width	Ratio Width	Bias	Coverage	Width	Ratio Width	Bias	Coverage	Width	Ratio Width
PL ^b	1	-0.03	0.82	3.39		-0.02	0.81	2.60		0.00	0.86	1.92	
	2	-0.04	0.83	3.13	0.53	-0.02	0.82	2.46	0.42	0.00	0.87	1.85	0.31
	3	-0.05	0.82	2.93	0.57	-0.01	0.82	2.34	0.46	0.01	0.86	1.78	0.35
	4,5,6	-0.03	0.85	2.57	0.64	-0.02	0.84	2.11	0.54	0.01	0.86	1.66	0.43
	4+	-0.02	0.87	2.29	0.70	-0.01	0.85	1.90	0.61	0.00	0.88	1.54	0.51
	7+	0.00	0.88	2.01	0.75	-0.01	0.86	1.70	0.68	0.00	0.90	1.41	0.59
BLPL ^c	1	-0.02	0.84	3.55		-0.01	0.83	2.77		-0.01	0.87	2.00	
	2	-0.02	0.84	3.26	0.55	-0.02	0.83	2.60	0.44	-0.01	0.87	1.93	0.33
	3	-0.04	0.85	3.04	0.59	-0.03	0.84	2.45	0.48	-0.01	0.87	1.85	0.36
	4,5,6	-0.02	0.86	2.65	0.66	-0.03	0.83	2.19	0.56	0.00	0.86	1.71	0.44
	4+	-0.02	0.87	2.35	0.71	-0.01	0.85	1.97	0.63	0.00	0.87	1.58	0.52
	7+	-0.03	0.88	2.06	0.77	0.01	0.87	1.75	0.69	0.00	0.88	1.44	0.60
BPL ^d	1	-0.05	0.86	3.75		-0.02	0.89	3.02		0.00	0.97	2.49	
	2	-0.04	0.86	3.43	0.58	-0.02	0.89	2.83	0.48	0.00	0.97	2.38	0.41
	3	-0.03	0.87	3.19	0.61	-0.02	0.90	2.67	0.53	0.01	0.97	2.28	0.45
	4,5,6	-0.04	0.88	2.77	0.68	-0.04	0.91	2.38	0.61	0.00	0.97	2.08	0.54
	4+	-0.03	0.89	2.44	0.74	-0.03	0.92	2.13	0.68	0.00	0.96	1.89	0.62
	7+	-0.02	0.90	2.12	0.79	-0.02	0.92	1.88	0.75	0.01	0.96	1.69	0.70

^aFor all configurations in the table the parameter representing predictive strength (π_1) is set to 1.4, corresponding to a relative risk of 4.

^bPL stands for the Pseudo-Likelihood method.

^cBLPL stands for Breslow and Lin correction to the Pseudo-Likelihood method.

^dBPL stands for the hybrid Bayesian Pseudo-Likelihood method.

Table 2

Operating characteristics of the hybrid approach for the GEM and WECARE datasets.

2 nd stage Predictor	Number of Carriers	GEM data						WECARE data					
		$\tau = 1.2$			$\tau = 0.8$			$\tau = 1.2$			$\tau = 0.8$		
		Bias	Coverage	Bias	Coverage	Bias	Coverage	Bias	Coverage	Bias	Coverage	Bias	Coverage
$\pi_1 = 0.8$	1	0.01	0.85	0.02	0.89	0.01	0.97	0.02	0.89	0.01	0.90	0.01	0.92
	2	0.02	0.86	0.02	0.89	0.00	0.97	0.01	0.89	0.00	0.90	0.01	0.92
	3	0.02	0.86	0.02	0.90	0.00	0.97	0.01	0.89	0.01	0.90	0.01	0.93
	4,5,6	0.03	0.87	0.02	0.91	-0.01	0.96	0.01	0.90	0.02	0.90	0.01	0.93
	4+	0.01	0.88	0.02	0.92	0.00	0.96	0.02	0.91	0.02	0.91	0.01	0.93
	7+	-0.01	0.90	0.02	0.92	0.01	0.96	0.02	0.92	0.02	0.91	0.01	0.93
$\pi_1 = 1.4$	1	-0.05	0.86	-0.02	0.89	0.00	0.97	-0.04	0.90	-0.01	0.90	0.01	0.93
	2	-0.04	0.86	-0.02	0.89	0.00	0.97	-0.03	0.90	-0.01	0.90	0.01	0.93
	3	-0.03	0.87	-0.02	0.90	0.01	0.97	-0.04	0.90	-0.02	0.90	0.02	0.93
	4,5,6	-0.04	0.88	-0.04	0.91	0.00	0.97	-0.03	0.90	-0.01	0.91	0.01	0.93
	4+	-0.03	0.89	-0.03	0.92	0.00	0.96	-0.01	0.91	0.00	0.91	0.01	0.93
	7+	-0.02	0.90	-0.02	0.92	0.01	0.96	0.00	0.91	0.01	0.92	0.01	0.93
$\pi_1 = 1.8$	1	-0.07	0.86	-0.05	0.89	0.01	0.98	-0.06	0.90	-0.02	0.90	0.02	0.93
	2	-0.06	0.87	-0.06	0.90	0.01	0.97	-0.07	0.91	-0.02	0.90	0.01	0.93
	3	-0.05	0.86	-0.05	0.90	0.01	0.98	-0.06	0.91	-0.03	0.91	0.02	0.93
	4,5,6	-0.07	0.87	-0.02	0.90	0.02	0.96	-0.03	0.91	-0.02	0.91	0.02	0.94
	4+	-0.05	0.90	-0.02	0.91	0.02	0.96	-0.02	0.92	-0.01	0.91	0.02	0.93
	7+	-0.03	0.92	-0.02	0.92	0.01	0.96	-0.01	0.92	-0.01	0.91	0.02	0.93

Table 3
Operating characteristics of the hybrid approach for the GEM data with 1, 2, or 3 second-stage predictors.

2 nd stage Predictor ^a	Number of Carriers	$\tau = 1.2$				$\tau = 0.8$				$\tau = 0.4$			
		Bias	Coverage	Width	Ratio Width	Bias	Coverage	Width	Ratio Width	Bias	Coverage	Width	Ratio Width
1 predictor	1	-0.05	0.86	3.75		-0.02	0.89	3.02		0.00	0.97	2.49	
	2	-0.04	0.86	3.43	0.58	-0.02	0.89	2.83	0.48	0.00	0.97	2.38	0.41
	3	-0.03	0.87	3.19	0.61	-0.02	0.90	2.67	0.53	0.01	0.97	2.28	0.45
	4,5,6	-0.04	0.88	2.77	0.68	-0.04	0.91	2.38	0.61	0.00	0.97	2.08	0.54
	4+	-0.03	0.89	2.44	0.74	-0.03	0.92	2.13	0.68	0.00	0.96	1.89	0.62
	7+	-0.02	0.90	2.12	0.79	-0.02	0.92	1.88	0.75	0.01	0.96	1.69	0.70
2 predictors	1	-0.06	0.88	4.05		-0.03	0.92	3.47		0.02	0.98	2.94	
	2	-0.06	0.88	3.71	0.61	-0.03	0.92	3.23	0.54	0.03	0.98	2.80	0.47
	3	-0.10	0.88	3.44	0.65	0.00	0.91	3.05	0.58	0.05	0.97	2.68	0.52
	4,5,6	-0.06	0.91	2.98	0.72	-0.02	0.92	2.69	0.66	0.03	0.97	2.44	0.61
	4+	-0.04	0.91	2.65	0.76	-0.02	0.93	2.41	0.72	0.03	0.97	2.20	0.68
	7+	-0.03	0.91	2.31	0.81	-0.01	0.93	2.13	0.78	0.04	0.97	1.96	0.75
3 predictors	1	-0.12	0.90	4.65		0.02	0.94	4.20		0.07	0.98	3.71	
	2	-0.08	0.90	4.27	0.67	0.04	0.94	3.93	0.62	0.06	0.98	3.51	0.56
	3	-0.09	0.90	3.95	0.70	0.03	0.94	3.68	0.66	0.08	0.97	3.32	0.61
	4,5,6	-0.07	0.91	3.51	0.76	0.03	0.95	3.28	0.74	0.07	0.97	3.02	0.69
	4+	-0.05	0.92	3.11	0.80	0.03	0.95	2.94	0.77	0.09	0.97	2.75	0.75
	7+	-0.03	0.93	2.71	0.84	0.04	0.94	2.60	0.81	0.10	0.97	2.47	0.80

^aFor each second-stage predictor the coefficient representing predictive strength was $\pi_1 = 1.4$.

Table 4

Relative risks and corresponding 95% confidence intervals for selected variants along with estimates of τ followed by their standard errors for the GEM dataset.

Variant	Controls	Cases	Functional	Odds Ratios (95% CI)				
				MLE ^a	PL ^b	BLPL ^c	BPL ^d	Gibbs sampling
c.50+37 G/C	15	6	no	0.8 (0.3–2.1)	0.8 (0.3–2.1)	0.8 (0.3–2.0)	0.8 (0.3–2.0)	0.8 (0.3–2.0)
(-) 33 G>C	9	6	no	1.2 (0.4–3.6)	1.2 (0.4–3.3)	1.2 (0.4–3.3)	1.2 (0.4–3.3)	1.2 (0.4–3.3)
(-) 34 G>T	2	7	no	13.5 (2.7–67.3)	6.4 (1.8–23.3)	6.9 (2.0–24.0)	7.6 (2.1–28.0)	8.9 (2–50.7)
c.301 G>T	1	5	yes	14.7 (1.6–132.3)	7.3 (1.5–35.4)	7.7 (1.7–34.9)	8.4 (1.7–41.8)	10.0 (1.8–92.9)
c.8_9ins	6	0	yes		0.7 (0.1–3.9)	0.6 (0.1–3.4)	0.5 (0.1–3.3)	0.3 (0.0–2.4)
c.123 G>A	4	1	no	0.4 (0.0–5.0)	0.6 (0.1–3.6)	0.6 (0.1–3.3)	0.5 (0.1–3.4)	0.5 (0.0–3.1)
(-) 14 C>T	2	1	no	1.2 (0.1–14.6)	0.8 (0.1–10.1)	1.1 (0.2–7.2)	1.1 (0.2–7.9)	1.1 (0.1–9.2)
(-) 25 C>T	2	0	no		0.7 (0.1–6.8)	0.7 (0.1–6.3)	0.6 (0.0–7.1)	0.3 (0.0–5.7)
c.373 G>C	3	0	yes		1.1 (0.2–7.6)	1.0 (0.1–6.6)	0.8 (0.1–6.8)	0.4 (0.0–4.9)
c.87_89delG	2	0	yes		1.0 (0.1–7.6)	0.9 (0.1–6.7)	0.8 (0.1–6.6)	0.4 (0.0–5.5)
c.318 G>A	2	1	yes	2.0 (0.2–23.1)	2.4 (0.4–15.0)	2.4 (0.4–13.9)	2.3 (0.4–15.2)	2.0 (0.2–15.0)
c.159 G>A	1	2	yes	2.3 (0.2–25.5)	2.6 (0.4–16.4)	2.5 (0.4–15.1)	2.5 (0.4–16.6)	2.6 (0.3–26.5)
c.159 G>C	1	1	yes	2.5 (0.1–41.4)	2.7 (0.4–19.3)	2.7 (0.4–18.0)	2.6 (0.3–20.4)	2.4 (0.2–28.0)
c.334 C>G	0	2	yes		8.0 (1.1–59.6)	8.7 (1.2–61.1)	10.5 (1.3–87.5)	16.9 (1.3–1072.8)
c.170 C>T	0	2	yes		4.7 (0.5–40.5)	4.9 (0.6–41.0)	5.5 (0.5–57.1)	8.4 (0.6–507.2)
c.95 T>C	0	2	yes		9.0 (1.2–65.3)	9.8 (1.4–67.0)	12.0 (1.5–96.2)	19.4 (1.5–1126.6)
(-) 70 G/A	0	1	no		2.0 (0.2–19.6)	2.1 (0.2–20.1)	2.4 (0.2–27.7)	4.1 (0.2–245.7)
c.146 T>G	0	1	yes		4.9 (0.5–44.9)	5.2 (0.6–46.4)	5.9 (0.5–66.0)	8.5 (0.5–641.6)
c.247 C>T	0	1	yes		5.9 (0.7–53.2)	6.3 (0.7–55.0)	7.4 (0.7–79.5)	10.6 (0.6–769.7)
c.67 G>A	0	1	yes		5.1 (0.6–46.1)	5.4 (0.6–47.5)	6.2 (0.6–67.6)	5.2 (0.3–370.2)
c.249 C>A	1	0	yes		1.7 (0.2–15.5)	1.6 (0.2–14.3)	1.4 (0.1–15.9)	0.8 (0.0–15.0)
c.370 C>T	1	0	yes		2.0 (0.2–19.1)	1.9 (0.2–18.4)	1.8 (0.1–21.6)	1.0 (0.0–17.7)
c.384 G>A	1	0	no		0.8 (0.1–8.3)	0.7 (0.1–8.1)	0.7 (0.0–9.4)	0.4 (0.0–9.4)
c.67 G>C	1	0	yes		2.0 (0.2–19.5)	1.9 (0.2–18.8)	1.8 (0.1–22.2)	1.0 (0.0–20.5)
2 nd stage			$\hat{\tau} (SE(\hat{\tau}))$	-	1.2 (0.38)	1.3 (0.40)	1.5 (0.36)	1.9 (0.88)

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

^b PL stands for the Pseudo-Likelihood method.

^c BLPL stands for Breslow and Lin correction to the Pseudo-Likelihood method.

^d BPL stands for the hybrid Bayesian Pseudo-Likelihood method.