Published in final edited form as:
Genet Epidemiol. 2010 January ; 34(1): 16–25. doi:10.1002/gepi.20429.

Bayesian Mixture Modeling of Gene-Environment and Gene-Gene Interactions

Jon Wakefield[1,2], **Frank De Vocht**[1,3], and **Rayjean J. Hung**[1,4]

[1] International Agency for Research on Cancer, Lyon, France

[2] Departments of Statistics and Biostatistics, University of Washington, Seattle, Washington, USA

[3] Occupational and Environmental Health Research Group, School of Translational Medicine, Faculty of Medical and Human Sciences, The University of Manchester

[4] Samuel Lunenfeld Research Institute, Toronto, Ontario, Canada

Abstract

With the advent of rapid and relatively cheap genotyping technologies there is now the opportunity to attempt to identify gene-environment and gene-gene interactions when the number of genes and environmental factors is potentially large. Unfortunately the dimensionality of the parameter space leads to a computational explosion in the number of possible interactions that may be investigated. The full model that includes all interactions and main effects can be unstable, with wide confidence intervals arising from the large number of estimated parameters. We describe a hierarchical mixture model that allows all interactions to be investigated simultaneously, but assumes the effects come from a mixture prior with two components, one that reflects small null effects and the second for epidemiologically significant effects. Effects from the former are effectively set to zero, hence increasing the power for the detection of real signals. The prior framework is very flexible, which allows substantive information to be incorporated into the analysis. We illustrate the methods first using simulation, and then on data from a case-control study of lung cancer in Central and Eastern Europe.

Keywords

Hierarchical models; Informative prior distributions; Markov chain Monte Carlo; Mean-variance trade-off

INTRODUCTION

Complex diseases such as cancers are the result of a multifactorial process, and the interplay between many risk factors including environmental exposures and genetic susceptibility (Adami et al., 2008; Weinberg, 2007). Specifically, the biological dosage of environmental exposures may be modified by an individuals biological response to the exposures, including their specific metabolism, DNA repair capacity and cell cycle control. On the other hand, genetic expression of specific cancer-related genes can be induced or altered by specific environmental exposures, and the extent of this alteration can be influenced by sequence variation. Thus it is important to identify interactions between various etiological factors for better understanding of the disease mechanism. A number of important gene-environment interactions have already been identified. One example concerns the interaction between alcohol dehydrogenase (ADH) genetic variants and alcohol consumption for upper aerodigestive cancer, in which increased alcohol consumption strengthens the association between ADH genetic variants and upper aerodigestive cancer risk (Hashibe et al., 2008). As a second example, Boccia et al. (2008) describe the interaction between

methylenetetrahydrofolate reductase (MTHFR) and folate intake with respect to gastric cancer risk, in which a more prominent genetic effect was observed when the folate intake level was low. Clayton and McKeigue (2001) provide a discussion of the differences between biological and statistical interpretations of "interaction". Here we concentrate on case-control data and refer to an interaction as a departure from main effects only in a logistic model.

In this paper we describe a method for modeling interactions when the number of single nucleotide polymorphisms (SNPs) is moderate (we have experimented with up to 100 SNPs and 5 exposures), for example in a candidate gene study, or following an initial screen in a genome-wide association study. An obvious strategy in such a situation is to fit separate models for each SNP/exposure combination. This requires a large number of models to be fitted, and a formal comparison between these models is not straightforward under a frequentist approach since the models are not nested. An additional, and more significant, problem is that the exposures will often be correlated and so fake associations will be induced due to confounding, since there is no control for additional exposures in a model containing a single exposure. These drawbacks suggest that a better approach is to include all of the SNPs, exposures and interactions in a single model. Unfortunately this leads to an increase in the standard errors of estimates, since there are a large number of parameters to simultaneously estimate, an example of the usual bias-variance trade-off in regression modeling. Numerical instability may also result, particularly with a large number of correlated exposures. In general, multiple comparisons issues must be considered, though common approaches such as the control of the family-wise error rate through a Bonferroni correction lead to a loss of power due to their overly conservative nature. A stepwise strategy in which main effects models are first fitted and then interactions are considered when the main effects are significant is fraught with problems in terms of interpretation, since the search through model space is data driven, which causes standard type I and type II error probability calculations to be incorrect (Miller, 1984).

Many methods have been suggested for the modeling of interactions (Motsinger et al., 2007). These range from conventional, parametric logistic regression driven approaches (Kraft et al., 2007; Marchini et al., 2005), to exploratory approaches such as multifactor dimensionality reduction (MDR) (Ritchie et al., 2001, 2003). MDR is a popular technique which we briefly describe in the context of detecting SNP-SNP interactions. For a fixed number of factors (SNPs), K, to include in the model MDR finds the optimal set by first splitting the data into multiple training and validation sets. Based on the training data, contingency tables corresponding to each possible combination of K factors are then constructed, with each cell in the table being labeled as high/low risk depending on whether the case/control ratio is >/< 1 within that cell. The "training error" is then calculated for each set of factors, and the set producing the lowest error is deemed the best for this value of K. Prediction errors and the consistency across different training sets are then computed for different values of K, to give the "best" K. MDR has recently been critiqued by Park and Hastie (2008). In particular they highlight that the power of MDR will suffer when the real effects are additive (since no structure is imposed on the form of the model). In addition, for high-dimensional tables many of the cells will contain zero cells, and the classification to high/low risk may be highly unstable. We would add the lack of flexibility in allowing prior information to be incorporated, and the ad hoc nature in labeling each cell as high/low risk.

Several tree-based methods have also been proposed to detect subgroups of interest, for example, classification and regression trees (CART) and random forests (Lunetta et al., 2004; Strobl et al., 2008; Wu et al., 2006). These type of methods have the advantage of dealing with large number of parameters with small sample sizes. However they too cannot

take into account available prior information (Ziegler et al., 2008). An additional critique is that subgroup effects detected in the tree-based methods are often difficult to interpret.

A number of methods have been proposed that build a semi-parametric model based on a logistic regression framework with a flexible regression model. Kooperberg and Ruczinski (2005) use logic regression, a technique for finding interesting combinations of binary regressors. Chen et al. (2007) proposed a method based on trees to detect gene-gene and gene-environment, while Chen et al. (2008) suggest another algorithmic method, support vector machines, to detect interactions. Again these approaches do not use prior information, and so cannot be tuned to specific applications.

Alternative suggestions begin by assuming independence of SNPs and exposures, in order to increase power. Methods for the case-only design that assumes and exploits the independence between SNPs and exposure, in order to increase power, have been proposed by a number of authors (Piegorsch et al., 1994: Umbach and Weinberg, 1997). Other methods assume independence and use case and control data (Chatterjee and Carroll, 1995; Chatterjee et al., 2006). The presence of gene-environmental non-independence can greatly bias the interaction estimates in this design, however (Albert et al., 2001: Kraft et al., 2007). To overcome this a hybrid estimator has recently been proposed (Mukherjee and Chatterjee, 2008), and appears promising in comparative simulation studies (Mukherjee et al., 2008). These methods have not been used in the multiple SNP/multiple exposure scenario.

Throughout we assume a rare disease and refer to odds ratios as relative risks. The approach we follow is Bayesian and is based on a mixture prior for the effect sizes. Specifically we assume that with probability $1 - \pi$ a single log relative risk arises from a normal prior centered at zero with a small variance (the null component of the prior), and with probability π the log relative risk arises from a zero mean normal with a larger variance reflecting anticipated effects (the non-null component). Such mixture priors have a long history in Bayesian linear model selection contexts (George and McCulloch, 1993, 1997; Mitchell and Beauchamp, 1988). The extension of such models in the context of modeling interactions in a linear model was suggested by Chipman (1996). Here we use a hierarchical model for the case-control context with a logistic regression model at the first stage of the hierarchy, with the mixture prior on the coefficients at the second stage. A similar model has been proposed in this context by Conti et al. (2003), though there are differences in the prior set-up, and in the reporting, as we will highlight.

We briefly describe two approaches to the detection of gene-environment interactions that we consider in detail later in the paper. The first is to fit single models containing a single SNP main effect, a single exposure main effect, and the interaction. Alternatively the full model containing all main effects and all interactions may be fitted. Unfortunately both of these strategies have serious drawbacks. The single models do not control for other variables, which can lead to serious problems of false positives, due to dependence between exposures. In the full model the standard errors can be very large due to the estimation of multiple parameters, which leads to a loss of power. The supplementary material contains a more formal critique of the full and single model strategies, and describes the difficulties of controlling meaningful criteria such as the false discovery rate (FDR) in a correlated multiple testing setting.

The structure of this paper is to first describe the Bayesian mixture model, before examining its behavior on simulated data. We then apply the model to data from a case-control study of lung cancer, before concluding with a discussion.

METHODS

Conventional Hierarchical Approach

Although we are primarily interested in modeling interactions, we motivate the mixture model by initially considering models for estimating the effects of multiple SNPs only. Let $Y_i = 1/0$ denote a case/control indicator with $i = 1, \ldots, n_1$ corresponding to cases and $i = n_1 + 1, \ldots, n_1 + n_0 = n$ to controls. Further let $x_{is} = 0/1/2$ be the minor allele count for SNP s, $s = 1, \ldots, S$, $i = 1, \ldots, n$.

A standard three-stage hierarchical model is given by:

Stage 1 *Data Model:* $Y_i|\alpha, \boldsymbol{\beta} \sim$ Binomial$(1, p_i)$ where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_S)$ and

$$\text{logit } p_i = \alpha + \sum_{s=1}^{S} x_{is}\beta_s$$

so that we have a linear predictor in the number of minor allele copies with $\exp(\beta_s)$ corresponding to the relative risk associated with a single copy of the minor allele for SNP s, $s = 1, \ldots, S$.

Stage 2 *Prior for Log Relative Risks:* Assume a common prior for the log relative risks:

$$\beta_s|\sigma_s^2 \sim N(0, \sigma_s^2), \quad s = 1, \ldots, S.$$

The s subscript is necessary here since later we will extend the notation to incorporate environmental exposures and interaction effects. We assign an improper flat prior to α.

Stage 3 *Hyperprior:* $\sigma_s^{-2} \sim$ Gamma(a_s, b_s) for suitable a_s, b_s

The latter stage may be neglected if σ_s^2 are fixed *a priori*, or if they are estimated from the data via empirical Bayes, or related methods (Greenland, 1992, 1993; Witte and Greenland, 1996). In the remainder of the paper we will fix the parameters of the prior model based on the context, since the majority of SNPs will be null, and so there will be limited information available to estimate the parameters of the prior.

We relate this model to the penalized logistic regression method of Park and Hastie (2008) which maximizes

$$L(\alpha, \beta) - \frac{\lambda}{2}\|\beta\|^2$$

for fixed λ where $L(\alpha, \boldsymbol{\beta}) = \log p(\mathbf{y}|\alpha, \boldsymbol{\beta})$ denote the log-likelihood. This method can be viewed from a Bayesian standpoint as finding the posterior mode with independent priors $\beta_s \sim_{iid} N(0, 1/\lambda)$, $s = 1, \ldots, S$, so that λ is the reciprocal of the variance. Sandwich estimation is used to obtain standard errors (which is more difficult to justify via a Bayesian argument), and cross-validation is used to choose λ.

A difficulty with the above three stage model is that it assumes that the prior is the same for all of the log relative risks β_s, when in the present context the vast majority of SNP effects will be zero, or very close to zero. We now describe a model that allows for each parameter to be effectively set to zero, thus aiding in interpretation, and allowing the remaining parameters to be estimated with greater power.

Mixture Hierarchical Approach

We describe a hierarchical mixture model which extends the above model by assuming that a proportion π_S of SNPs are non-null, with a common prior for their effects, while the remaining proportion $1 - \pi_S$ of null SNPs arise from a distribution with its mass very close to zero. Mixture models, with components corresponding to null and non-null signals, were suggested in an epidemiologlcal multiple testing situation by Thomas et al. (1985, 1992). A similar logistic model to that described below was proposed in the gene-environment context by Conti et al. (2003). Stage 1 is as in the conventional model with Stages 2 and 3 given by:

Stage 2 *Prior for Log Relative Risks*: We assume the mixture prior:

$$\beta_s|T_s, \sigma_s^2 \sim N\left(0, T_s\sigma_s^2+(1-T_s)\sigma_{\varepsilon_s}^2\right)$$

where T_s is an indicator for SNP s that takes the value 1 for a non-null SNP, and the value 0 for a null SNP. We assume that the non-null effects arise from $N(0, \sigma_s^2)$ with the null effects from $N(0, \sigma_{\varepsilon_s}^2)$. The latter is chosen so relative risks in the range $e^{\pm\varepsilon_S}$ for small ε_S can be treated as null. In what follows we choose $\sigma_{\varepsilon_s}^2 = 0.012^2$ which corresponds to the belief that the null relative risks lie in the range (0.98, 1.02) with probability 0.9. It is useful to allow values slightly either side of 1 in order to "soak up" small amounts of bias arising from, for example, small differential genotyping errors, or confounding due to population stratification.

Stage 3 *Prior for Non-Null/Null Indicators*: Without additional information we assume that

$$T_s|\pi_S \sim \text{Bernoulli}(\pi_S), \quad s=1,\ldots,S,$$

so that π_S is the proportion of non-null SNPs.

As we have noted, often the number of non-null SNPs will not be large, and so it will not be possible to estimate σ_s^{-2} and π_S from the data, so we fix them in advance. We set the standard deviation so that we expect $100(1 - q)$% of non-null SNPs to have relative risks that lie within $\exp(\pm R)$, to give $\sigma_S = \log R/\Phi^{-1}(1 - q/2)$. As an example, for 90% within $[e^{-2.5}, e^{2.5}]$ we obtain $\sigma_S = \log 2.5/1.645 = 0.557$. The anticipated proportion of non-null SNPs is given by $S \times \pi_S$, which aids in the fixing of π_S. For example, with $S = 50$ SNPs in total a best guess of 5 non-null SNPs gives $\pi_S = 0.10$. Note that under this prior set-up, the prior on the number of non-null SNPs is binomial with parameters S and π_S, which allows the prior probability of a particular number of non-null SNPs to be evaluated.

We note the following about this model:

1. Setting $T_s = 1$ for all s gives the conventional hierarchical model.

2. If we average out over the T_s (i.e. collapse stages 2 and 3 into a single stage) we obtain the mixture model

$$\beta_s|\pi_s, \sigma^2_{\varepsilon_s}, \sigma^2_s \sim (1-\pi_s)N(0, \sigma^2_{\varepsilon_s}) + \pi_s N(0, \sigma^2_s)$$

3. If we have prior information on particular SNPs being non-null we can relax the Stage 3 assumption that the indicators T_s are independent and identically distributed to independent with π_{Ss}, for $s = 1, \ldots, S$. Such information may reflect information such as whether the SNP is missense, or has been previously implicated with the disease. Setting $\pi_{Ss} = 1$ indicates that we believe SNP s is important.

Figure 1(a) shows the two components of the prior (for the choices discussed above, namely 95% points of 1.02 and 2.5 for the null and non-null components). The densities cross at a log relative risk value of 0.033 so that $|\beta|$'s less than this are more likely to be null, with values larger more likely to be non-null. For a given log relative risk β, one can evaluate the false discovery rate (FDR):

$$\Pr(T=0|\beta) = \frac{p(\beta|T=0)\pi_s}{p(\beta|T=0)\pi_s + p(\beta|T=1)(1-\pi_s)}$$

and this quantity is displayed in Figure 1(b) (with $\pi_S = 0.8$). The FDR associated with this prior shows desirable behavior in that it is monotonic decreasing from zero (Rice and Spiegelhalter, 2008). This is desirable since it means that as $|\beta|$ increases it becomes increasingly less likely that the coefficient corresponds to a false discovery.

Gene × Environment Interactions

Turning now to gene-environment interactions we let z_{ie} denote exposure e on individual i, $i = 1, \ldots, n$, $e = 1, \ldots, E$. We assume that continuous exposures are standardized to have mean 0 and standard deviation 1, so that associated log relative risks are on the same scale. We extend the model of the previous section:

Stage 1 *Data Model:* $Y_i|\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta} \sim$ Binomial(1, p_i) where

$$\text{logit } p_i = \alpha + \sum_{s=1}^{S} x_{is}\beta_s + \sum_{e=1}^{E} z_{ie}\gamma_e + \sum_{s=1}^{S}\sum_{e=1}^{E} x_{is}z_{ie}\delta_{se}$$

so that $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_E)$ are the main effects associated with exposure, and $\boldsymbol{\delta} = (\delta_{11}, \ldots, \delta_{SE})$ are the interactions. We have assumed that the interactions act in the same fashion on 1 or 2 copies of the minor allele, though the mixture formulation is easily extended to other interaction models.

Stage 2 *Prior for Non-Null/Null Indicators*:

$$\begin{aligned}
\beta_s|T_s, \sigma^2_s &\sim N\left(0, T_s\sigma^2_s + (1-T_s)\sigma^2_{\varepsilon_s}\right), \quad s=1,\ldots,S \\
\gamma_e|T_e, \sigma^2_e &\sim N\left(0, T_e\sigma^2_e + (1-T_e)\sigma^2_{\varepsilon_e}\right), \quad e=1,\ldots,E \\
\delta_{se}|T_{se}, \sigma^2_{se} &\sim N\left(0, T_{se}\sigma^2_{se} + (1-T_{se})\sigma^2_{\varepsilon_{se}}\right) \quad s=1,\ldots,S, e=1,\ldots,E
\end{aligned}$$

Stage 3 *Priors for Non-Null/Null Indicators*:

$$\begin{aligned} T_s|\pi_s &\sim_{iid} \text{Bernoulli}(\pi_s), \quad s=1,\ldots,S \\ T_e|\pi_e &\sim_{iid} \text{Bernoulli}(\pi_e), \quad e=1,\ldots,E \\ T_{se}|\pi_{se} &\sim_{iid} \text{Bernoulli}(\pi_{se}), \quad s=1,\ldots,S, \quad e=1,\ldots,E \end{aligned}$$

We also assume that the collections $\boldsymbol{T}_s$, $\boldsymbol{T}_e$ and $\boldsymbol{T}_{se}$ are independent. Again we fix $\sigma_s^2, \sigma_{se}^2, \sigma_{se}^2$, π_s, π_s and π_{se}.

We make four observations:

1. A special case of this model is to take $T_{se} = T_s T_e$, which equals 1 only if both main effects are in the model, and is zero otherwise, thus obeying the "hierarchy principle" in which an interaction term is only included if the main effect is present. A further extension is to take $T_{se} = T_s T_e \Delta$ where $\Delta < 1$ so that we have a reduced probability of an interaction, relative to the product of the SNP and exposure interactions (Chipman, 1996). In both the simulations and the real data example, the number of exposures is small, and we will assume that $\pi_e = 1$ so that we do not have a null prior for exposures, and we take $T_{se} = T_s T_e = T_s$. In this case we only need to specify π_s.
2. The above model can be applied in an obvious fashion to three- and higher-way interactions, and to the analysis of SNP-SNP interactions.
3. Interpretation of the results requires care in situations in which the majority of the probability mass of $\Pr(T|\boldsymbol{y})$ is not at $T = 0$ or $T = 1$. If the mass is distributed more equally between these points then interval estimates for the corresponding regression coefficient may be deceptive. One possibility is to report posterior intervals *given* that $T = 1$, along with $\Pr(T = 1|\boldsymbol{y})$.
4. The model of Conti et al. (2003) differs from the above in the following respects. We place a tight normal prior around the log odds ratio under the null, while Conti et al. set the parameter to zero. Under the alternative ($T = 1$) we fix the variance of the effect sizes, while Conti et al. estimate these variances and have a specific prior on variance inflation parameters. We fix these variances because we want to be as context-specific as possible. In the situations in which we are envisaging, the number of non-null effects will be small, and so the variance will be difficult to estimate. Similarly, Conti et al. assign Beta prior distributions to, and estimate, the mixture probabilities, whereas we fix these *a priori*. In terms of reporting, Conti et al. recommend, for a generic parameter θ, tabulation of $\mathrm{E}[\theta|T = 1, \boldsymbol{y}]$, $\mathrm{var}(\theta|T = 1, \boldsymbol{y})$ and $\Pr(\theta < 0|T = 1, \boldsymbol{y})$, whereas we will examine credible intervals for the unconditional parameter, for example, we report those values (l, u) such that $\Pr(l < \theta < u|\boldsymbol{y}) = 0.95$, along with $\Pr(\theta > 0|\boldsymbol{y})$. Conti et al also explicitly emphasize model averaging and calculate posterior probabilities of each model (where the class of models is obtained by considering all values of $T = 0/1$). We view the analysis as more exploratory in nature and hence concentrate on the posterior probabilities of coefficients being "significantly" different from zero.

Computation

The posterior distribution is given by

$$p(\alpha, \beta, \gamma, \delta, \boldsymbol{T}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\alpha, \beta, \gamma, \delta) p(\alpha) p(\beta|\boldsymbol{T}_{se}) p(\gamma|\boldsymbol{T}_e) p(\delta|\boldsymbol{T}_{se}) p(\boldsymbol{T}_{se})$$

The first stage logistic model means that marginalization short-cuts available for the linear model (which allow the random effects to be integrated out analytically) are not possible here, but Markov chain Monte Carlo (MCMC) techniques, in which a Markov chain is constructed whose stationary distribution is the posterior distribution, may be utilized.

As noted by George and McCulloch (1997) it is important that the ratios $\sigma^2_s/\sigma^2_{\varepsilon_s}, \sigma^2_e/\sigma^2_{\varepsilon_e}, \sigma^2_{se}/\sigma^2_{\varepsilon_{se}}$, should not be made too large since this will lead to poor convergence of the Markov chain, because the sampler can become "stuck" at values of β close to zero. Intuitively, if the two variances are not too different then the sampler will move easily between the $T = 0$ and $T = 1$ models, but if one is highly peaked it is difficult to move since the ratio of the peaked to the flatter densities is too large.

To assess convergence we recommend that multiple chains be run. In the examples that follow we run two chains, one with starting values $T = 0$ and each regression parameter set initially to zero, and the other setting $T = 1$ for those parameters that are significantly different from zero from the fit of the full model (with initial values for the significant regression parameters set to the estimates from this model). Figure 2 illustrates the mixing for a particular parameter, β, with associated indicator variable T, from two chains. These data were taken from one of the simulations of the next section. Panel (a) shows that $T = 0$ for the majority of the iterations, but there are visits to the $T = 1$ state. Panel (c) shows how β is more varied when $T = 1$. Panels (b) and (d) display histogram representations of the conditional posterior densities, $\beta|\mathbf{y}, T = 0$ and $\beta|\mathbf{y}, T = 1$, respectively. When $T = 0$ the posterior for β essentially corresponds to the narrow prior.

RESULTS

Simulation Study

We assume binary exposures and to simulate data, first fixed SNP and exposure prevalences, and then set a number of relative risks for SNP and exposure main effects, and interactions to be non-null. There are clearly many characteristics of the simulation that can be varied, and we explored only a subset. Null SNP minor allele frequencies (MAFs) were set to 0.3, with non-null SNPs having MAFs of 0.1 and relative risks of 1.2. The prevalences of exposed individuals was 0.1 in all cases and non-null exposure and interaction relative risks were set to 2. In the dependent exposure simulations the correlation was approximately 0.75. We examine the performance of three models: (1) $S \times E$ "single" models that contain four parameters only (the intercept, main effects for SNP and exposure, and the interaction); (2) the full model that contains all SNPs/exposures and interactions; (3) the Bayesian mixture model. We set $\pi_s = 0.2$ and $\pi_e = 1$ so that the non-null exposure prior is always used; we also have $T_{se} = T_s$ so that the hierarchy principle is obeyed.

For the mixture models a pair of Markov chains were run for a minimum of 50,000 iterations, discarding the first 10,000 as burn-in. A number of the models with a larger number of SNPs/exposures required longer runs.

Figure 3 displays 95% interval estimates for each parameter for simualtion 9 (since the same main effects are estimated multiple times under the single models, we report the interactions only for these models). This figure clearly shows the wide intervals that result from the full model, and the effective setting to zero of the majority of parameters under the mixture model. The supplementary material contains plots analogous to Figure 3, for all of the simulations.

Table 1 summarizes the results for each of the combinations of parameters we examine. We concentrate on inference for the interactions and report the number of false discoveries (type

I errors), and missed signals (type II errors), where these are (somewhat arbitrarily) evaluated based on 95% confidence/credible intervals. In each simulation, and for each of the interactions we also evaluate the posterior probabilities of exceedence of 0, and these are presented in Figure 4. For true null signals these should be close to 0.5, and for true non-null signals they will be close to zero (for protective effects) or one (for detrimental effects, which is the case in our simulations). We also consider the mean squared error of the interactions, which is given by the sum of the variance and the square of the bias:

$$\text{MSE}=\frac{1}{S\times E}\sum_{s=1}^{S}\sum_{e=1}^{E}(\widehat{\delta}_{se}-\delta_{se})^2=\frac{1}{S\times E}\sum_{s=1}^{S}\sum_{e=1}^{E}\left[\text{var}(\widehat{\delta}_{se})+(\widehat{\delta}_{se}-\delta_{se})^2\right]$$

where $S \times E$ is the number of interactions, and δ_{se} is the interaction parameter corresponding to SNP s and exposure e.

We make some general conclusions, and discuss each simulation in more detail in the supplementary material. In general:

- The single models perform better, in terms of MSE, than the full models, with the mixture models having orders of magnitude lower MSEs than either.
- The 95% interval estimates are widest for the full model, and shortest for the mixture model.
- There is slight bias towards zero in the Bayesian mixture model, due to the prior being centered at zero, but so far as MSE is concerned, this is more than offset by the reduced variance.
- The full model approach tends to have higher type I error rates than the single model analyses, while in only one simulation did the mixture model give a type I error.
- Using the 95% credible interval criteria the mixture model has a number of Type II errors. However, as Figure 4 shows, the mixture model is ranking the associations reasonably well. Figure 4 shows the posterior probability of exceedence of 0, for all of the interaction parameters. We see that although for a number of the simulations (numbers 8, 12, 16 and 17 in particular) signals are missed at the 95% level, they are detected at level 90% without any increase in the number of type I errors. In simulation 10, in which the exposures are correlated, the “wrong” SNP-exposure interaction is detected, but the correct signal is the second most significant. Figures 1 and 2 of the supplementary material give the p-values associated with each interaction and for each simulation for the single and full model examples, respectively. An advantage of examination of $\Pr(\delta > 0|\boldsymbol{y})$ is that for null signals these posterior probabilities tend to concentrate around 0.5. In contrast, under the null p-values are uniform on (0,1) and so it is more difficult to discern the true non-null signals from the null effects.
- Simulation 18 represents a situation similar to many in practice with 100 SNPs, two correlated exposures, and a single non-null interaction. For this simulation all three of the methods correctly identify the interaction, though the full model gives a very wide interval estimate. The full model also gives 20 false positives, while the single models give 7 and the mixture model 0.

Real Data

We analyze data previously described in Hung et al. (2004), from a study conducted in centers within six countries of Central and Eastern Europe including the Czech Republic, Hungary, Poland, Romania, Russia and Slovakia. Within each country an identical protocol was followed with a consecutive group of newly diagnosed cases of lung cancer and a comparable group of controls being recruited between 1998 and 2002. Controls in all centers except Warsaw were chosen among subjects admitted as in-patients or out-patients in the same hospital as the cases, and were frequency matched with the case group by sex, age (+/− 3 years), center and referral (or residence) area. The eligible diseases for controls were non-tobacco related diseases including minor surgical conditions, benign disorders, common infections, eye conditions (except cataract or diabetic retinopathy) common orthopaedic diseases (except osteoporosis), etc. In Warsaw, population controls were selected by random sampling from the Polish Electronic List of Residents. Both cases and controls underwent an identical face-to-face interview during which they completed a detailed questionnaire including sections on: (a) demographic variables; (b) medical history; (c) family history of cancer; (d) tobacco smoking and involuntary smoking; (e) alcohol drinking; (f) intake frequency of selected food items: (g) occupational history. Blood samples were collected at the time of interview.

For illustration of the model we here examine 20 SNPs, two of which, CHEK2I157T (rs17879961, Brennan et al. 2007) and loc123688 (rs8034191, Hung et al. 2008), are known to be associated with lung cancer. As exposure we take a binary smoking variable (0=never, 1=ever) which has a known and strong association with lung cancer. The model we fit has 48 parameters (plus an intercept): 20 main effects for SNPs, a main effect for smoking, 20 interactions between the SNPs and smoking, and seven additional variables coding countries 1–5 (so that country 6 is the baseline), gender and age. Since CHEK2I157T and loc123688 are known to be associated we set $T = 1$ for these SNPs, and for the remainder we take $\pi_S = 0.2$. We placed flat priors on the intercept and the log relative risks corresponding to age, gender and the country indicators. For simplicity we consider the 1752 individuals with full data.

Figure 5 shows 95% interval estimates for each of the 48 parameters from the full logistic model and single models (all fit via MLE), and the Bayes mixture model. There was numerical instability for the logistic models — for example, the standard errors for the main effect of CHEK2I157T, and the interaction between CHEK2I157T and smoking were both 395 (the *y*-axis in Figure 5 is truncated) for the full model. The single model containing this interaction was also highly unstable. Two interactions were flagged as “just” significant under the full model, and one by the single model. Under the Bayesian mixture model (for which the results are stable) no signal interaction was categorically flagged as “significant”. For 18 of the 20 interactions, the values of $\Pr(\delta|\boldsymbol{y})$ lay between 0.38 and 0.61 indicating no evidence of an interaction. Of the remaining SNPs, NAT2C282T gave a probability of 0.83. However, the interaction between loc123688 and smoking gave a posterior probability of 0.93 giving a hint of significance, and indicating that the harmful allele combined with ever smoking further increased the risk of lung cancer.

DISCUSSION

In this paper we have proposed the use of a Bayesian mixture model for modeling interactions, and have shown, via simulation, that this model has good performance when compared to fitting a full model, or multiple single models. All of the Bayesian mixture models were fitted using the freely-available `WinBUGS` software (Spiegelhalter et al., 1998). Example code is presented in the supplementary material, and is also available at: http://faculty.washington.edu/jonno/cv.html

In practice there are often missing genotype data, the MCMC approach is ideally suited to imputing these missing data see, for example, Lunn et al. (2006). We have assumed a binary phenotype (case-control data) but the hierarchical model could easily be altered for continuous phenotypes including survival data. Often in studies of environment and occupation the exposures are subject to measurement error. The incorporation of a parametric measurement error model is also straightforward within the Bayesian framework.

The choice of the π parameters will determine the operating characteristics (type I errors and power) and varying these parameters will give a more or less conservative procedure. For example, if we are in exploration mode we should set the π's to be relatively large, since we would rather have a longer list of possible interactions to investigate. Sensitivity of the results to the specific values of π chosen is an important step.

In the hierarchical model that we have described the number of relative risks is fixed and equal to the maximum number of main effects and interactions that we wish to consider. An alternative approach is to allow the model dimension to change as terms are "dropped" and "added" to the model. This is an attractive strategy but requires specialized reversible jump MCMC (Green, 1995) and great care in diagnosis of convergence/posterior summarization. Some progress has been made in the former with the `WinBUGS` software (Lunn et al., 2006), though logistic regression models are not currently available.

The implementation for the model we use is computationally expensive and so the use of the model in a genome-wide association study (GWAS) is not currently feasible. However, Lewinger et al. (2007) use a similar mixture model for prioritizing signals in a GWAS.

In summary, we have described a Bayesian mixture model framework to assess gene-environment and gene-gene interactions in higher dimensional case-control studies. We have shown that this framework performs much better than traditional methods, and dramatically reduces both the MSE and the number of false positives, especially when exposures are correlated. The discovery of interactions in the setting we have considered is an intrinsically difficult endeavor, however, and so we would recommend the use of the mixture model in tandem with complementary alternative methods such as recently-proposed tree-based approaches.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank the Cancer Research UK small research grant (grant reference C29425/A9638), while JW would like to thank the National Institutes of Health for support under grant 1 U01–HG004446–01. The constructive comments of a referee are also appreciated.

References

Adami, HO.; Hunter, D.; Trichopoulos, D. Textbook of Cancer Epidemiology. New York: Oxford University Press; 2008.

Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. Limitations of the case-only design for identifying gene-environment interactions. American Journal of Epidemiology. 2001; 154:687–693. [PubMed: 11590080]

Boccia S, Hung R, Ricciardi G, Gianfagna F, Ebert MP, Fang JY, Gao CM, Gotze T, Graziano F, Lacasana-Navarro MD, Lopez-Carrillo L, Qiao YL, Shen H, Stolzenberg-Solomon R, Takezaki T, Weng YR, Zhang FF, van Duijn CM, Boffetta PE, Taioli. Meta- and pooled analyses of the

methylenetetrahydrofolate reductase C677T and A1298C polymorphisms and gastric cancer risk: a huge-GSEC review. American Journal of Epidemiology. 2008; 167:505–516. [PubMed: 18162478]

Brennan P, McKay J, Moore L, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, Chow WH, Rothman N, Chabrier A, Gaborieau V, Odefrey F, Southey M, Hashibe M, Hall J, Boffetta P, Peto J, Peto R, Hung RJ. Uncommon CHEK2 mis-sense variant and reduced risk of tobacco-related cancers: case control study. Human Molecular Genetics. 2007; 16:1794–1801. [PubMed: 17517688]

Chatterjee N, Carroll RJ. Semiparametri maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika. 1995; 92:399–418.

Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. American Journal of Human Genetics. 2006; 79:1002–1016. [PubMed: 17186459]

Chen J, Yu K, Hsing A, Thernau TM. A partially linear tree-based regression model for assessing complex joint gene-gene and gene-environment effects. Genetic Epidemiology. 2007; 31:238–251. [PubMed: 17266115]

Chen SH, Sun J, Dimitriv L, Turner AR, Adams TS, Meyers DA, Chang BL, Zheng SL, Gronberg H, Xu J, Hsu FC. A support vector machine approach for detecting gene-gene interaction. Genetic Epidemiology. 2008; 32:152–167. [PubMed: 17968988]

Chipman H. Bayesian variable selection with related predictors. Canadian Journal of Statistics. 1996; 24:17–36.

Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. The Lancet. 2001; 358:1356–1360.

Conti DV, Cortessis V, Molitor J, Thomas DC. Bayesian modeling of complex metabolic pathways. Human Heredity. 2003; 56:83–93. [PubMed: 14614242]

George EI, McCulloch RE. Variable selection via Gibbs sampling. Journal of the American Statistical Association. 1993; 88:881–889.

George EI, McCulloch RE. Approaches for variable selection. Statistica Sinica. 1997; 7:339–374.

Green PJ. Reversible jump MCMC computation and Bayesian model determination. Biometrika. 1995; 82:711–732.

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. [PubMed: 1579760]

Greenland S. Methods for epidemiologic analyses of multiple exposures: a review and comparitive study of maximum-likelihood, preliminary-testing, and empirical Bayes regression. Statistics in Medicine. 1993; 12:717–736. [PubMed: 8516590]

Hashibe M, McKay JD, Curado MP, Oliveira JC, Koifman S, Koifman R, Zaridze D, Shangina O, Wunsch-Filho V, Eluf-Neto J, Levi JE, Matos E, Lagiou P, Lagiou A, Benhamou S, Bouchardy C, Szeszenia-Dabrowska N, Menezes A, Dall'Agnol MM, Merletti F, Richiardi L, Fernandez L, Lence J, Talamini R, Barzan L, Mates D, Mates IN, Kjaerheim K, Macfarlane GJ, Macfarlane TV, Simonato L, Canova C, Holcatova I, Agudo A, Castellsague X, Lowry R, Janout V, Kollarova H, Conway DI, McKinney PA, Znaor A, Fabianova E, Bencko V, Lissowska J, Chabrier A, Hung RJ, Gaborieau V, Boffetta P, Brennan P. Multiple ADH genes are associated with upper aerodigestive cancers. Nature Genetics. 2008; 40:707–709. [PubMed: 18500343]

Hung RJ, Brennan P, Malaveille C, Porru S, Donato F, Boffetta 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; 13:1013–1021.

Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, Chen C, Goodman G, Field JK, Liloglou T, Xinarianos G, Cassidy A, McLaughlin J, Liu G, Narod S, Krokan HE, Skorpen F, Elvestad MB, Hveem K, Vatten L, Linseisen J, Clavel-Chapelon F, Vineis P, de Mesquita HB, Bueno Lund E, Martinez C, Bingham S, Rasmuson T, Hainaut P, Riboli E, Ahrens W, Benhamou S, Lagiou P, Trichopoulos D, Holctov I, Merletti F, Kjaerheim K, Agudo A, Macfarlane G, Talamini R, Simonato L, Lowry R, Conway DI, Znaor A, Healy C, Zelenika D, Boland A, Delepine M, Foglio M, Lechner D, Matsuda F, Blanche H, Gut I, Heath S,

Lathrop M, Brennan P. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature. 2008; 452:633–637. [PubMed: 18385738]

Kooperberg C, Ruczinski I. Identifying interacting SNPs using Monte Carlo logic regression. Genetic Epidemiology. 2005; 28:157–170. [PubMed: 15532037]

Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. Human Heredity. 2007; 63:111–119. [PubMed: 17283440]

Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC. Hierarchical Bayes prioritization for marker associations from a genome-wide association scan for further investigation. Genetic Epidemiology. 2007; 31:871–882. [PubMed: 17654612]

Lunetta KL, Hayward LB, Segal J, Eerdewegh P Van. Screening large-scale association study data: exploiting interactions using random forests. BMC Genetics. 2004; 5:32. [PubMed: 15588316]

Lunn DJ, Whittaker JC, Best N. A Bayesian toolkit for genetic association studies. Genetic Epidemiology. 2006; 30:231–247. [PubMed: 16544290]

Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature Genetics. 2005; 37:413–417. [PubMed: 15793588]

Miller AJ. Selection of subsets of Regression Variables. Journal of the Royal Statistical Society, Series A. 1984; 147:389–425.

Mitchell TJ, Beauchamp JJ. Bayesian variable selection in linear regression (with discussion). Journal of the American Statistical Association. 1988; 83:1023–1036.

Motsinger AA, Ritchie MD, Reil DM. Novel methods for detecting epistasis in pharmacogenomics studies. Pharmacogenomics. 2007; 8:1229–1241. [PubMed: 17924838]

Mukherjee B, Ahn J, Fruber SB, Rennert G, Moreno V, Chatterjee N. Tests for gene-environment interaction from case-control data: a novel study of type I error, power and designs. Genetic Epidemiology. 2008; 32:1–14. [PubMed: 17630650]

Mukherjee B, Chatterjee N. Exploiting gene-environment independence for analysis of case-control studies: an empirical-Bayes type shrinkage estimator to trade off between bias and efficiency. Biometrics. 2008 On-line.

Park MY, Hastie T. Penalized logistic regression for detecting gene interactions. Biostatistics. 2008; 9:30–50. [PubMed: 17429103]

Piegorsch WW, Weinberg CR, Taylor J. Non hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Statistics in Medicine. 1994; 13:153–162. [PubMed: 8122051]

Rice K, Spiegelhalter D. Comment: Microarrays, empirical Bayes and the two groups model by B. Efron. Statistical Science. 2008; 23:41–44.

Ritchie M, Hahn L, Moore J. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. American Journal of Human Genetics. 2001; 69:138–147. [PubMed: 11404819]

Ritchie M, Hahn L, Moore J. Power of multifactor dimensionality reduction for detecing gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heteropgeneity. Genetic Epidemiology. 2003; 24:150–157. [PubMed: 12548676]

Spiegelhalter, DJ.; Thomas, A.; Best, NG. WinBUGS User Manual, version 1.1.1. Cambridge; UK: 1998.

Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics. 2008; 9:307. [PubMed: 18620558]

Thomas D, 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; 24:387–92. [PubMed: 1418924]

Thomas D, Siemiatycki J, Dewar R, Robins J, Goldberg M, Armstrong BG. The problem of multiple inference in studies designed to generate hypotheses. American Journal of Epidemiology. 1985; 122:1080–95. [PubMed: 4061442]

Umbach DM, Weinberg CR. Designing and analyzing case-control studies to exploit independence of genotype and exposure. Statistics in Medicine. 1997; 16:1731–1743. [PubMed: 9265696]

Weinberg, RA. The Biology of Cancer. Abingdon: Garland Science; 2007.

Witte JS, Greenland S. Simulation study of hierarchical regression. Statistics in Medicine. 1996; 15:1161–1170. [PubMed: 8804145]

Wu X, Gu J, Grossman HB, Amos CI, Etzel C, Huang M, Zhang Q, Millikan RE, Lerner S, Dinney CP, Spitz MR. Bladder cancer predisposition: a multigenic approach to DNA-repair and cell-cycle-control genes. American Journal of Human Genetics. 2006; 78:464–479. [PubMed: 16465622]

Ziegler A, Konig IR, Thompson JR. Biostatistical aspects of genome-wide association studies. Biomedical Journal. 2008; 50:8–28.

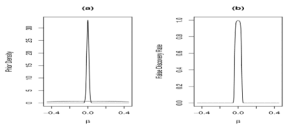

Figure 1.
(a) Two component mixture prior, and (b) false discovery rate Pr(null $|\beta$).

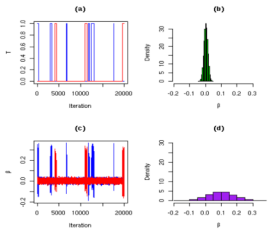

Figure 2.
(a) Realizations of indicator variable T associated with parameter β, for two Markov chains (displayed in blue and red). (b) Histogram representation of the posterior of the conditional $\beta|\boldsymbol{y}, T = 0$. (c) Realizations of the parameter β, for two Markov chains (displayed in blue and red). (d) Histogram representation of the posterior of $\beta|\boldsymbol{y}$, $T = 1$.

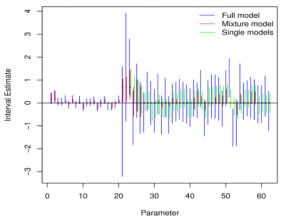

Figure 3.
Summaries of three modeling approaches for simulation 9. Parameters 1–20 represent SNP main effects, 21–22 are the exposure main effects, and 23–62 are the 40 interactions. The latter are ordered so that 23 and 24 correspond to the interaction between SNP 1 and exposures 1 and 2, respectively, etc. The five short horizontal lines represent the true effect sizes.

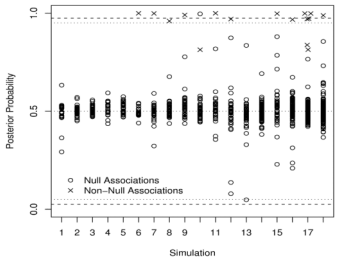

Figure 4.
Posterior probabilities that interactions exceed 0, $\Pr(\delta > 0|\boldsymbol{y})$. The null signals should cluster around 0.5. The top dashed line is at 0.975, which corresponds to the cut-off for a 95% credible interval. Horizontal lines are also drawn at probabilities of 0.95, 0.5, 0.05 and 0.025.

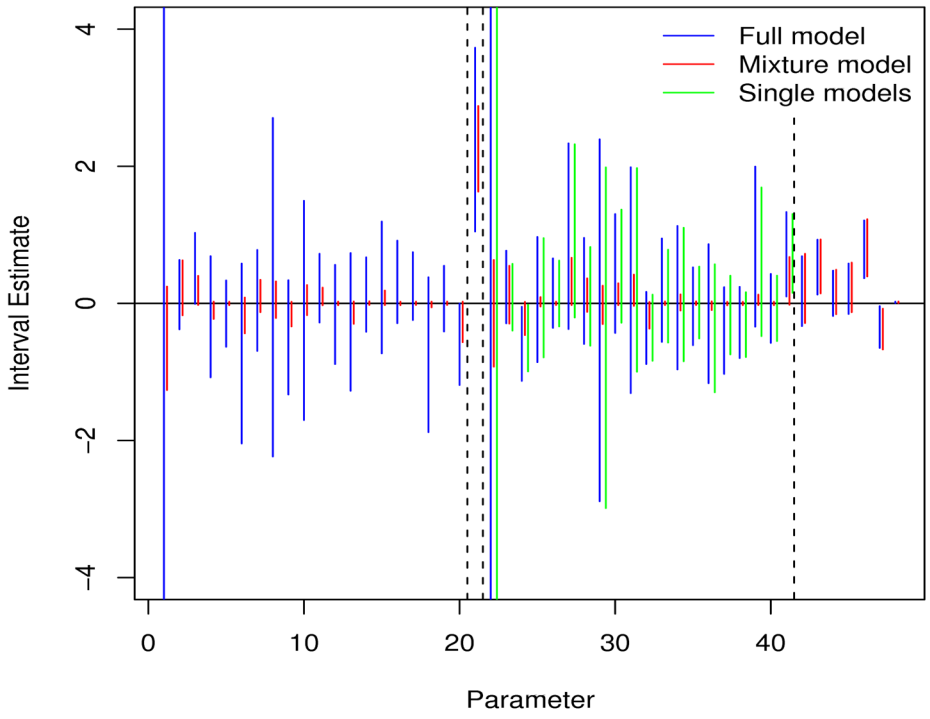

Figure 5.
95% interval estimates under the full and mixture models. The dashed vertical lines denote distinct sets of parameters: 20 main effects for SNPs, a main effect for smoking, 20 interactions between the SNPs and smoking, and seven additional variables coding centers 1–5, gender and age.

Table 1

Simulated data settings and summaries. For the non-null SNPs the relative risks are 1.2, while the relative risks for the non-null exposure main effects and interactions are 2. Exposure prevalences are 10%, as are MAFs for non-null SNPs. All other SNPs have MAFs of 0.3.

		Number/Non-null			Single			Full			Mixture		
	n	SNPs	Exposures	Interactions	MSE	Type I Errors	Type II Errors	MSE	Type I Errors	Type II Errors	MSE	Type I Errors	Type II Errors
1	1000	20/0	1/0	20/0	0.113	1 (0.05)	—	0.253	3 (0.15)	—	0.0110	0 (0)	—
2	1000	20/0	2/0★	40/0	0.0866	0 (0)	—	0.255	2 (0.05)	—	0.00291	0 (0)	—
3	1000	20/2	2/0★	40/0	0.100	0 (0)	—	0.580	2 (0.05)	—	0.00272	0 (0)	—
4	1000	20/1	1/1	20/0	0.0865	1 (0.05)	—	0.106	1 (0.05)	—	0.00302	0 (0)	—
5	1000	20/1	2/2★	40/0	0.0796	0 (0)	—	0.343	3 (0.08)	—	0.0150	0 (0)	—
6	1000	20/1	1/1	20/1	0.0664	0 (0)	0 (0)	0.121	0 (0)	0 (0)	0.00561	0 (0)	0 (0)
7	1000	20/2	2/1	40/1	0.105	1 (0.03)	0 (0)	0.183	2 (0.05)	0 (0)	0.00806	0 (0)	0 (0)
8	1000	20/2	2/1★	40/1	0.0934	2 (0.05)	0 (0)	0.154	0 (0)	1 (100)	0.0108	0 (0)	1 (100)
9	1000	20/2	2/2★	40/1	0.108	2 (0.05)	0 (0)	0.464	2 (0.05)	0 (0)	0.0185	0 (0)	0 (0)
10	2000	20/2	2/2★	40/1	0.0606	2 (0.05)	0 (0)	0.234	4 (0.10)	1 (100)	0.0226	1 (0.05)	1 (100)
11	2000	20/1	5/1	100/1	0.0576	7 (0.07)	0 (0)	0.0891	11 (0.11)	0 (0)	0.0149	0 (0)	0 (0)
12	2000	20/1	5/1★	100/1	0.0635	8 (0.08)	0 (0)	0.391	9 (0.09)	1 (100)	0.0163	0 (0)	1 (100)
13	2000	50/1	1/0	50/0	0.0581	3 (0.06)	—	0.0763	3 (0.06)	—	0.00556	0 (0)	—
14	2000	50/1	2/0★	100/0	0.0547	6 (0.06)	—	0.184	3 (0.03)	—	0.00673	0 (0)	—
15	2000	50/5	1/1	50/1	0.0477	6 (0.12)	0 (0)	0.0676	4 (0.08)	0 (0)	0.00704	0 (0)	0 (0)
16	2000	50/2	2/2★	100/1	0.0405	7 (0.09)	0 (0)	0.159	1 (0.01)	0 (0)	0.00481	0 (0)	1 (100)
17	2000	50/5	5/3★	250/6	0.101	9 (0.04)	1(17)	1.71	32 (0.13)	3 (50)	0.0427	0 (0)	4(67)
18	2000	100/2	2/2★	200/1	0.0418	7 (0.04)	0 (0)	0.718	20 (0.10)	0 (0)	0.00381	0 (0)	0 (0)

★denotes correlated exposures, with correlation approximately 0.75. The mean-squared errors (MSE) are averages over the interaction parameters. The number of type I and type II errors are given as an absolute number and as a percentage (in brackets), and are based on a 95% interval estimate.