STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:113–139
Published online 12 September 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3413

Semiparametric Bayesian modeling of random genetic effects in family-based association studies

Li Zhang[1,*,†], Bhramar Mukherjee[2], Bo Hu[1], Victor Moreno[3] and Kathleen A. Cooney[4]

[1]*Department of Quantitative Health Sciences, The Cleveland Clinic Foundation, Cleveland, OH 44195, U.S.A.*
[2]*Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
[3]*Unit of Biostatistics and Bioinformatics, Catalan Institute of Oncology, and Autonomous University of Barcelona, Barcelona, Spain*
[4]*Departments of Internal Medicine and Urology, University of Michigan Medical School, University of Michigan Comprehensive Cancer Center, Ann Arbor, MI 48109, U.S.A.*

SUMMARY

We consider the inference problem of estimating covariate and genetic effects in a family-based case-control study where families are ascertained on the basis of the number of cases within the family. However, our interest lies not only in estimating the fixed covariate effects but also in estimating the random effects parameters that account for varying correlations among family members. These random effects parameters, though weakly identifiable in a strict theoretical sense, are often hard to estimate due to the small number of observations per family. A hierarchical Bayesian paradigm is a very natural route in this context with multiple advantages compared with a classical mixed effects estimation strategy based on the integrated likelihood. We propose a fully flexible Bayesian approach allowing nonparametric modeling of the random effects distribution using a Dirichlet process prior and provide estimation of both fixed effect and random effects parameters using a Markov chain Monte Carlo numerical integration scheme. The nonparametric Bayesian approach not only provides inference that is less sensitive to parametric specification of the random effects distribution but also allows possible uncertainty around a specific genetic correlation structure. The Bayesian approach has certain computational advantages over its mixed-model counterparts. Data from the Prostate Cancer Genetics Project, a family-based study at the University of Michigan Comprehensive Cancer Center including families having one or more members with prostate cancer, are used to illustrate the proposed methods. A small-scale simulation study is carried out to compare the proposed nonparametric Bayes methodology with a parametric Bayesian alternative. Copyright © 2008 John Wiley & Sons, Ltd.

*Correspondence to: Li Zhang, The Department of Quantitative Health Sciences, Cleveland Clinic, Desk JJN3-01, 9500 Euclid Ave., Cleveland, OH 44195, U.S.A.
†E-mail: zhangl3@ccf.org

Contract/grant sponsor: NIH; contract/grant number: R03 CA130045-01
Contract/grant sponsor: NSF; contract/grant number: DMS 07-06935
Contract/grant sponsor: SPORE; contract/grant numbers: P50 CA69568, R01 CA79596

Received 14 November 2007
Accepted 21 July 2008

KEY WORDS: conditional logistic regression; Dirichlet process prior; integrated likelihood; matched case–control studies; random effects model

1. INTRODUCTION

Genetic epidemiology is a relatively new field that applies conventional epidemiologic designs and methods to explore the role genetic factors play in determining the etiology of a disease. Both theoretical and empirical studies have shown that traditional linkage studies may be inferior in power compared with studies directly utilizing allele status. On the other hand, population-based case–control association studies are subject to bias due to population stratification. As a compromise between linkage studies and population-based case–control studies, family-based association designs have received great attention recently due to their potentially higher power to identify complex disease genes and their robustness in the presence of population substructure [1–3].

A common phenomenon in genetic epidemiologic research is that sampled families are not representative of the targeted population as they are ascertained through probands with known phenotypic values. It is well known in the literature that statistical inference without proper ascertainment corrections would lead to biased estimations of the key parameters of interest. One simple remedy is to condition on the observed phenotypic values of the probands. In family-based case–control studies, a natural approach to account for the family effect will be to conduct a matched case–control analysis with controls selected from the same family and to use conditional logistic regression (CLR) conditional on the number of cases in each family (there could be more than one person affected by the disease in a family). However, one must use caution if controls are selected from outside of the case-ascertainment region [4, 5].

Despite the need to estimate family-specific random effects parameters, a mixed-model approach can lead to substantial gain in efficiency, relative to the conditional likelihood [6, 7]. Whittemore, Zhao *et al.* and Neuhaus *et al.* [8–10] proposed marginal or population-averaged models to analyze family-based data. However, often we are interested in family-specific effects that relate the probability of the response to changes in covariates within a family. Neuhaus *et al.* [11] compared and contrasted the estimates from the family-specific model and the marginal model. A standard route to model a correlated binary response as illustrated in [12] is to introduce a random effect as a linear predictor in a generalized linear model. Akin to this approach, Pfeiffer *et al.* [13] proposed a two-level mixed effects model to estimate environmental effects while accounting for varying genetic correlations among family members and adjusting for ascertainment by conditioning on the number of cases in the family. Pfeiffer *et al.* [13] based their analysis on the marginal conditional likelihood after integrating with respect to the joint random effects distribution of individual and family-level genetic effects. This approach took into account unmeasured familial and genetic effects that induce correlated responses and yielded consistent estimators of covariate effects under certain conditions, even with a misspecified random effects distribution.

The approach presented in [13] has certain flexibilities but comes with the drawback of computational complexity as one has to approximate the integrated likelihood by Monte Carlo samples. Although the Monte Carlo approach worked well in the examples presented therein, larger Monte Carlo samples or other methods may be needed for larger pedigrees. The optimization algorithm proposed in [13] requires fixing certain parameters and then searching for an optimum in the space of other parameters to overcome numerical instability; this conditional grid search may not always be quite efficient. We propose a full Bayesian approach to construct a hierarchical

DOI: 10.1002/sim

pedigree structure assuming priors on the genetic random effects, which offers an appealing alternative.

Pfeiffer *et al.* [13] modeled the covariance matrix of the genetic random effects as a function of the degree of kinship between members in each family by assuming no dominance component of the genetic variance [14]. Thus, their inference regarding the parameters related to the individual-level random effects was reduced to inference on only one scalar common variance parameter σ_g^2 as they assume a fixed correlation structure. Instead of assuming a fixed correlation structure, we introduce generation-specific variances, and interclass (between two generations, e.g. parent–offspring) and intraclass (within the same generation, e.g. offspring–offspring) correlations. Neuhaus *et al.* [15] presented family-specific models in a similar general structure and provided an elegant semiparametric likelihood estimation strategy. Pfeiffer *et al.*'s [13] two-level mixed model can be viewed as a special case of this more general class of models. Our proposed Bayesian methodology is not restricted to any specific model description or genetic correlation structure. However, we consider for illustration purposes the two specific situations: (i) parent–offspring familial data and (ii) Pfeiffer *et al.*'s [13] mixed effects model with a general pedigree structure.

Though the hierarchical Bayesian approach with parametric priors on the random effects is also novel to this specific problem, we take our approach an additional step further by nonparametric modeling of the random effects distribution using a Dirichlet process (DP) prior [16]. Pfeiffer *et al.* [13] pointed out that in many cases one does not know the precise nature of the genetic influences and hence the distribution of familial or individual-level genetic effects. The estimated random effects for each individual and family will be modified by changing the distribution of the random effects. This point is quite critical because there are many applications in which estimates of the random effects parameters themselves are desired. Given the nature of the current problem in mind, we allow this additional layer of model uncertainty via a flexible nonparametric Bayesian approach to attain robust inference. There has been a significant volume of recent literature on parametric Bayesian approaches to random effects logistic models [17, 18]. Bayesian nonparametric modeling of random effects distribution has been considered by several authors [19–21], but the application to family-based studies becomes especially interesting due to the sparsity of the information in each family, the familial correlation structure and the ascertainment correction in the likelihood.

In this paper, we provide a fairly general framework for nonparametric Bayesian modeling of the random effects distribution for family-based association studies. The primary advantages of our hierarchical Bayesian approach are: (i) it allows the possibility of incorporating prior information on the correlation and variance components parameters, which are hard to estimate due to limited observations per family, (ii) the DP prior works as an automated data-adaptive dimension reduction technique to handle the family-specific parameters as well as provide a model-robust alternative for the random effects distribution and (iii) it provides a comprehensive computing algorithm based on the exact posterior distribution of model parameters as enumerated via a Markov chain Monte Carlo (MCMC) numerical integration scheme avoiding complex optimization and approximation issues involved with the classical integrated likelihood approach.

The rest of the paper is organized as follows. In Section 2, we present the likelihood, the integrated likelihoods and the conditional likelihoods under different model structures. We introduce the proposed Bayesian approaches with description of priors, and details of parametric and nonparametric modeling of the random effects distribution in Section 3. In Section 4, we apply our proposed method to data from the University of Michigan Prostate Cancer Genetics Project (PCGP), a family-based study of inherited prostate cancer susceptibility, and then end the section with a

DOI: 10.1002/sim

small-scale simulation study to illustrate the advantage of our Bayesian nonparametric method. Section 5 contains concluding discussion, while some proofs, model extension and computational details are relegated to the Appendix.

2. MODELS AND LIKELIHOODS

Let the family data consist of a binary disease status variable Y_{ij}, together with a collection of covariate vectors X_{ij} for the jth member of the ith family, $i=1,\ldots,I$ and $j=1,\ldots,n_i$. Let $\mathbf{Y}_i=(Y_{i1},\ldots,Y_{in_i})^{\mathrm{T}}$ and $\mathbf{X}_i=(X_{i1},\ldots,X_{in_i})^{\mathrm{T}}$ denote the data corresponding to the ith family. We specify a vector of random parameters $\mathbf{b}_i$ specific to the ith family, capturing familial random effects. In its most general specification, we essentially consider the following mixed effects logistic model [22, 23]:

$$\log\left\{\frac{p_{ij}}{1-p_{ij}}\right\}=\mu+\beta X_{ij}+\mathbf{Z}_{ij}\mathbf{b}_i \tag{1}$$

where $p_{ij}=P(Y_{ij}=1|\mathbf{b}_i,X_{ij},\mathbf{Z}_{ij})$ and β stands for the effects of the covariates on the disease status. With an appropriate choice of $\mathbf{Z}_{ij}$, we can model different pedigree structures and correlations [15].

2.1. Mixed model for parent–offspring familial data

Case-siblings and case-parents designs are fairly common family-based designs that may be viewed as special cases of the parent–offspring data that we consider. We propose the following mixed effects model to account for generation effects (e.g. parent and offspring variance components) and different correlations among family members. We introduce a model with interclass (parent–offspring) correlation within family, induced through the random effects and an intraclass (offspring–offspring) correlation through disease responses (Y-values) in the bivariate familial data.

For simplicity in illustration, in the description below, we consider $n_i=4$ members in each family, among whom two are parents and the other two are offspring. Let b_{i1} and b_{i2} denote the random genetic effects for the parent and the offspring in each family i, respectively. $Z_{ij1}=1, Z_{ij2}=0$, if the jth member in family i is the parent; and $Z_{ij1}=0, Z_{ij2}=1$, if the jth member in family i is the offspring. $\mathbf{Z}_i$ is a 4×2 matrix of indicator variables with (Z_{ij1},Z_{ij2}) as the jth row, $j=1,\ldots,4$. Thus (1) becomes

$$\log\left\{\frac{p_{ij}}{1-p_{ij}}\right\}=\mu+\beta X_{ij}+Z_{ij1}b_{i1}+Z_{ij2}b_{i2} \tag{2}$$

$\mathbf{b}_i$ follows a bivariate distribution with the expectation of $\mathbf{b}_i=(b_{i1},b_{i2})^{\mathrm{T}}$ being $\boldsymbol{\mu}_b=(\mu_{b1},\mu_{b2})^{\mathrm{T}}$ and the variance–covariance matrix

$$\Sigma_b=\begin{pmatrix}\sigma_p^2 & \rho_{pc}\sigma_p\sigma_c\\ \rho_{pc}\sigma_p\sigma_c & \sigma_c^2\end{pmatrix} \tag{3}$$

DOI: 10.1002/sim

Table I. The joint probability $P(Y_{i3}=d_{i3}, Y_{i4}=d_{i4}|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i)$ under the mixed model for parent–offspring familial data.

Y_{i3}	Y_{i4}	$P(Y_{i3}=d_{i3}, Y_{i4}=d_{i4}\|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i)$
1	1	$p_{i3}p_{i4}+\rho^*_{cc}\sqrt{p_{i3}(1-p_{i3})p_{i4}(1-p_{i4})}$
1	0	$p_{i3}(1-p_{i4})-\rho^*_{cc}\sqrt{p_{i3}(1-p_{i3})p_{i4}(1-p_{i4})}$
0	1	$(1-p_{i3})p_{i4}-\rho^*_{cc}\sqrt{p_{i3}(1-p_{i3})p_{i4}(1-p_{i4})}$
0	0	$(1-p_{i3})(1-p_{i4})+\rho^*_{cc}\sqrt{p_{i3}(1-p_{i3})p_{i4}(1-p_{i4})}$

Note that $\boldsymbol{\mu}_b$ is not part of fixed effect of β. Σ_b allows separate variance components for parents (σ^2_p) and offspring (σ^2_c) and the correlation between random effects b_{i1} and b_{i2} corresponding to parents and offspring, denoted by ρ_{pc}.

In each family we assume that conditional on the random effects, the parents are unrelated, but there could be potential correlation among offspring, which in this model we define by ρ^*_{cc} through the direct correlation between Y-values rather than being imposed on the logistic scale via the random effects. Therefore, conditional on $\mathbf{b}_i$, the responses Y_{ij} of two parents within a given family i are independent; so are the responses of the subjects from two different generations in family i, but the responses Y_{ij} for two offspring are correlated. Hence without loss of generality, let the first two members in each family be the parents, $P(Y_{i1}, Y_{i2}|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i) = P(Y_{i1}|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i)P(Y_{i2}|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i)$, and $P(Y_{ij}, Y_{ij'}|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i) = P(Y_{ij}|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i)P(Y_{ij'}|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i)$ for $j=1,2$ and $j'=3,4$, but $P(Y_{i3}, Y_{i4}|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i) \neq P(Y_{i3}|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i)P(Y_{i4}|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i)$. The calculation of the joint probability of Y_{i3} and Y_{i4} given $\mathbf{b}_i$ is just treating Y_{i3} and Y_{i4} as two correlated binary variables with correlation ρ^*_{cc}. The results are shown in Table I (for detailed calculation, see Appendix A.1).

Thus, the joint probability for each family i can be written as the following:

$$\begin{aligned} &P(\mathbf{Y}_i=\mathbf{d}_i|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i) \\ &\quad = P(Y_{i1}=d_{i1}|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i)P(Y_{i2}=d_{i2}|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i)P(Y_{i3}=d_{i3}, Y_{i4}=d_{i4}|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i) \\ &\quad = p_{i1}p_{i2} \times P(Y_{i3}=d_{i3}, Y_{i4}=d_{i4}|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i) \end{aligned} \tag{4}$$

In many situations, we may have more than two offspring in each family. In the Appendix, we show how to calculate the joint probability $P(Y_{i3}, Y_{i4}, Y_{i5}|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i)$, i.e. when we have three offspring in each family, with some discussion of possibilities to extend the structure to the case of a general number of offspring. However, for illustration purposes, we restrict our attention to this special case. Please note that the essential idea of introducing varying correlations across degree of relation could be extended to any larger pedigree structures.

The model we propose for parent–offspring data departs from the standard class of mixed effect models where the Y-values are typically independent, conditional on the family-specific random effects. In our model conditional independence holds except for the offspring–offspring correlations, which remain even after conditioning on the random effects. In that sense, this model may be viewed as a hybrid blend of a population-averaged and family-specific approaches. Note also that the parent–offspring interclass correlation ρ_{pc} may be viewed as a measure of genetic

DOI: 10.1002/sim

correlation, while the offspring–offspring intraclass correlation ρ_{cc}^* measures the correlation of the disease responses between the offspring and does not have an interpretation in terms of pure genetic correlations only.

2.2. Mixed model of Pfeiffer et al. [13] with modified covariance structure for familial random effects

We now describe the two-level mixed model proposed by Pfeiffer *et al.* [13]. Let a_i denote the random familial effect for each family i and g_{ij} stand for an individual random genetic effect for the jth individual in the ith family. The g_{ij}'s are correlated within the ith family, but independent across families. Model (1) can be rewritten as follows:

$$\log\left\{\frac{P(Y_{ij}=1|a_i, g_{ij}, X_{ij})}{P(Y_{ij}=0|a_i, g_{ij}, X_{ij})}\right\}=\mu+\beta X_{ij}+a_i+g_{ij} \tag{5}$$

where a_i and $\mathbf{g}_i=(g_{i1},\ldots,g_{in_i})$ have expectation zero, and $(a_i, \mathbf{g}_i)$ are assumed to be independent.

Following [14] and assuming no dominance component of the variance, Pfeiffer *et al.* [13] modeled the covariance matrix of the $\mathbf{g}_i$'s $(i=1,\ldots,I), \Sigma_i$, as a function of degree of kinship between members in the family:

$$\text{cov}(g_{ij}, g_{il})=\sigma_g^2(\mathbf{R}_i)_{j,l}=\frac{\sigma_g^2}{2^{k(j,l)}} \tag{6}$$

where $k(j,l)$ denotes the degree of kinship between members j and l in the ith family. For example, $k(j,j)=0$ and $k(j,l)=1$ if j and l are first-degree relatives. For unrelated members, such as spouses, $k(j,l)=\infty$.

Instead of fixing the genetic correlations, we rather introduce the covariance matrix of the $\mathbf{g}_i$'s $(i=1,\ldots,I)$ with the variance parameters for each generation and the correlation parameters, which can flexibly formulate most pedigree structures. For example, suppose in each family i there are two parents and n_i-2 offspring and without loss of generality, let the first two members in each family be parents, thus $\Sigma_g^{n_i}$ has the structure of

$$\Sigma_g^{n_i}=\begin{pmatrix} \sigma_p^2 & 0 & \rho_{pc}\sigma_p\sigma_c & \cdots & \rho_{pc}\sigma_p\sigma_c \\ 0 & \sigma_p^2 & \rho_{pc}\sigma_p\sigma_c & \cdots & \rho_{pc}\sigma_p\sigma_c \\ \rho_{pc}\sigma_p\sigma_c & \rho_{pc}\sigma_p\sigma_c & \sigma_c^2 & \cdots & \rho_{cc}\sigma_c^2 \\ \vdots & \vdots & \cdots & \ddots & \vdots \\ \rho_{pc}\sigma_p\sigma_c & \rho_{pc}\sigma_p\sigma_c & \rho_{cc}\sigma_c^2 & \cdots & \sigma_c^2 \end{pmatrix} \tag{7}$$

where similar to the parent–offspring familial model, σ_p^2 and σ_c^2 are the variances for parents and offspring, respectively, and ρ_{pc} is the correlation between the random effects corresponding to the parent and offspring. But, different from ρ_{cc}^*, ρ_{cc} is the correlation between the random

DOI: 10.1002/sim

effects corresponding to the two different offspring. Thus, in contrast to the parent–offspring model discussed in Section 2.1, conditional on the random effects a_i and $\mathbf{g}_i$, the disease outcomes are independent. Note here that both ρ_{pc} and ρ_{cc} are interpretable as genetic correlations.

The joint probability corresponding to each family i is simply the product of disease probabilities of each family member:

$$P(\mathbf{Y}_i=\mathbf{d}_i|\mathbf{X}_i,a_i,\mathbf{g}_i)=\prod_{j=1}^{n_i}P(Y_{ij}=d_{ij}|X_{ij},a_i,g_{ij})$$
$$=\prod_{j=1}^{n_i}\frac{\exp\{d_{ij}(\mu+\beta X_{ij}+a_i+g_{ij})\}}{1+\exp\{\mu+\beta X_{ij}+a_i+g_{ij}\}} \tag{8}$$

The entire likelihood is the product of these family-specific contributions, namely, $\prod_{i=1}^{I}P(\mathbf{Y}_i=\mathbf{d}_i|\mathbf{X}_i,a_i,\mathbf{g}_i)$.

In the real data analysis, another set of family pedigree structures will be presented, which we would discuss in Section 4. The crux of the modeling approach is to data-adaptively estimate the genetic correlation structure and use available prior information on these parameters.

2.3. Conditional likelihood

In family-based case–control studies, a natural approach to account for the ascertainment effect will be to conduct a matched case–control analysis with controls selected from the same family and to use CLR conditional on the number of cases in the family. Statistical techniques for analyzing matched case–control data were first developed in [24]. The generated conditional likelihood is free of the nuisance parameters and yields the optimum estimating function [25] for estimating β.

With finer association structures across family members such as in [13], the conditioning is applied to the marginal likelihood of the data, and this marginal likelihood is obtained by integrating with respect to the joint random effects distribution. For example, for the mixed model in (5), the marginal probability of the disease in the ith family can be written as

$$\Pr(\mathbf{Y}_i=\mathbf{d}_i|\mathbf{X}_i)=\int P(\mathbf{Y}_i=\mathbf{d}_i|\mathbf{X}_i,a_i,\mathbf{g}_i)\,\mathrm{d}F(a_i,\mathbf{g}_i) \tag{9}$$

where $F(a_i,\mathbf{g}_i)$ is the joint distribution of random effects $a_i,\mathbf{g}_i$. If the number of affected family members is $\sum_{j=1}^{n_i}Y_{ij}=m_i$, then the conditional likelihood for family i is given by

$$\Pr\left(\mathbf{Y}_i=\mathbf{d}_i|\mathbf{X}_i,\sum_{j=1}^{n_i}Y_{ij}=m_i\right)=\frac{\Pr(\mathbf{Y}_i=\mathbf{d}_i,\sum_{j=1}^{n_i}Y_{ij}=m_i|\mathbf{X}_i)}{\Pr(\sum_{j=1}^{n_i}Y_{ij}=m_i|\mathbf{X}_i)} \tag{10}$$

The full conditional likelihood is the product of such I likelihoods as given in (10). For illustration purposes and to avoid cumbersome conditioning notations, in the following, we assume that in each family there are exactly two cases, i.e. $m_i=2$.

One can continue estimation of the parameters based on maximizing the above integrated likelihood in a classical frequentist framework. But the main challenge is integration over the joint random effects. Pfeiffer *et al.* [13] used Monte Carlo integration. For the mixed model proposed by Pfeiffer *et al.* [13], the above marginal conditional likelihood can be explicitly written as

DOI: 10.1002/sim

follows:

$$\prod_{i=1}^{I} \frac{\exp\{\beta(\sum_{j=1}^{n_i} d_{ij}X_{ij})\}\int \dfrac{\exp(2a_i+\sum_{j=1}^{n_i} d_{ij}g_{ij})}{\prod_{j=1}^{n_i}\{1+\exp(\mu+a_i+\beta X_{ij}+g_{ij})\}}\,\mathrm{d}F(a_i,\mathbf{g}_i)}{\sum_{k,l\in\mathscr{R}_i}\exp\{\beta(X_{ik}+X_{il})\}\int \dfrac{\exp(2a_i+g_{ik}+g_{il})}{\prod_{j=1}^{n_i}\{1+\exp(\mu+a_i+\beta X_{ij}+g_{ij})\}}\,\mathrm{d}F(a_i,\mathbf{g}_i)}$$

The summation in the denominator is overall $n_i(n_i-1)/2$ pairs in the set $\mathscr{R}_i$ that consists of selections of two possible 'cases' from any of the n_i family members. For each family i, they drew independent, identically distributed samples $a_i^{(k)}$ and $\mathbf{g}_i^{(k)}$ from the random effects distributions for each $k=1,\ldots,N$, and used the approximation

$$\int \frac{\exp(2a_i+\sum_{j=1}^{n_i} d_{ij}g_{ij})}{\prod_{j=1}^{n_i}\{1+\exp(\mu+a_i+\beta X_{ij}+g_{ij})\}}\,\mathrm{d}F(a_i,\mathbf{g}_i)\approx\frac{1}{N}\sum_{k=1}^{N}\frac{\exp(2a_i^{(k)}+\sum_{j=1}^{n_i} d_{ij}g_{ij}^{(k)})}{\prod_{j=1}^{n_i}\{1+\exp(\mu+a_i^{(k)}+\beta X_{ij}+g_{ij}^{(k)})\}}$$

They chose $N=100$ and used the same Monte Carlo sample for the numerator and denominator of the conditional likelihood of each family to ensure that the conditional likelihood was smooth in β. Different families were evaluated using independent Monte Carlo samples.

However, there is often very little information on the genetic random effects in the ascertainment corrected likelihood (which is the conditional likelihood conditioning on the ascertainment event), i.e. leading to numerical instabilities and computational challenges. This is because less information in the likelihood makes it harder to maximize the likelihood surface and obtain MLEs for the parameters but in the Bayesian context leads to prior-sensitive inference. Hence, instead, we consider a full Bayesian alternative by assuming a hierarchical prior structure on the random effects, but we continue to implement inference based on the conditional likelihood as we propose in the following, which is not a marginal conditional likelihood as proposed in [13]. Treating the conditional likelihood as a valid likelihood and proceeding with Bayesian inference may raise some concerns; however, [26] provides an interpretation of the conditional likelihood as a marginal likelihood and characterizes the nuisance distribution on a_i, which renders this equivalence and provides a justification for using this likelihood as an initial point in conducting Bayesian inference. A referee has pointed out the possibility of using a direct retrospective likelihood by posing a model for the exposure distribution conditional on disease status. However, such an approach may lead to robustness issues when the exposure vector is high dimensional and a mixture of categorical and continuous variables, and thus we choose the stratified prospective model as the basis of our inference.

Model 1

The conditional likelihood for the model in (2) corresponding to parent–offspring familial data is

$$L(\beta,\mathbf{b}_1,\ldots,\mathbf{b}_n,\rho_{cc}^*|\cdot)=\prod_{i=1}^{I} P\left(\mathbf{Y}_i=\mathbf{d}_i|\mathbf{X}_i,\mathbf{Z}_i,\mathbf{b}_i,\sum_{j=1}^{4}Y_{ij}=2\right) \tag{11}$$

where

$$P\left(\mathbf{Y}_i=\mathbf{d}_i|\mathbf{X}_i,\mathbf{Z}_i,\mathbf{b}_i,\sum_{j=1}^{4}Y_{ij}=2\right)=\frac{P(\mathbf{d}_i|\mathbf{X}_i,\mathbf{Z}_i,\mathbf{b}_i)}{\sum_{k=1}^{6}P(\mathbf{D}_k|\mathbf{X}_i,\mathbf{Z}_i,\mathbf{b}_i)} \tag{12}$$

DOI: 10.1002/sim

The summation in the denominator is over all $4(4-1)/2=6$ pairs in the set that consists of selections of two possible 'cases' from any of the four family members, i.e. $\mathbf{D}_k$ is the kth row of the matrix

$$\mathbf{D}=\begin{pmatrix}1&1&0&0\\1&0&1&0\\1&0&0&1\\0&1&1&0\\0&1&0&1\\0&0&1&1\end{pmatrix} \tag{13}$$

Hence,

$$\begin{aligned}L(\beta,\mathbf{b}_1,\ldots,\mathbf{b}_n,\rho^*_{cc}|\cdot)=\prod_{i=1}^{I}&\exp\{d_{i1}(\beta X_{i1}+b_{i1})+d_{i2}(\beta X_{i2}+b_{i1})\}\\&\times\left[\exp\{d_{i3}(\beta X_{i3}+b_{i2})+d_{i4}(\beta X_{i4}+b_{i2})\}\right.\\&\left.+(-\rho^*_{cc})^{d_{i3}+d_{i4}}\exp\left\{\frac{1}{2}\beta(X_{i3}+X_{i4})+b_{i2}\right)\right\}\Bigg]\\&\times\left\{\sum_{k=1}^{6}\exp\{D_{k1}(\beta X_{i1}+b_{i1})+D_{k2}(\beta X_{i2}+b_{i1})\}\right.\\&\times\left[\exp\{D_{k3}(\beta X_{i3}+b_{i2})+D_{k4}(\beta X_{i4}+b_{i2})\}\right.\\&\left.\left.+(-\rho^*_{cc})^{D_{k3}+D_{k4}}\exp\left\{\frac{1}{2}\beta(X_{i3}+X_{i4})+b_{i2}\right)\right\}\right]\Bigg\}^{-1}\end{aligned} \tag{14}$$

The above conditional likelihood can be easily extended to any choice of $m_i=1,\ldots,4$ and $\sum_{j=1}^{4}Y_{ij}\geqslant m_i$. For example, in the case when we have more than one diseased individual in each family, the summation in the denominator of (12) is over all $\sum_{m=2}^{4}C_m^4=11$ combinations in the set that consists of selections of two, three and four possible 'cases' from any of the four family members. In that case, $\mathbf{D}$ would be an 11×4 matrix.

Model 2

For the mixed model of [13], the conditional likelihood is

$$L(\beta,\mathbf{g}_1,\ldots,\mathbf{g}_n|\cdot)=\prod_{i=1}^{I}P\left(Y_{i1},\ldots,Y_{in_i}|a_i,g_{i1},\ldots,g_{in_i},X_{i1},\ldots,X_{in_i},\sum_{j=1}^{n_i}Y_i=2\right) \tag{15}$$

DOI: 10.1002/sim

where

$$
\begin{aligned}
&P\left(Y_{i1},\ldots,Y_{in_i}|a_i,g_{i1},\ldots,g_{in_i},X_{i1},\ldots,X_{in_i},\sum_{j=1}^{n_i}Y_i=2\right)\\
&=\frac{P(Y_{i1},\ldots,Y_{in_i},\sum_{j=1}^{n_i}Y_i=2|a_i,g_{i1},\ldots,g_{in_i},X_{i1},\ldots,X_{in_i})}{P(\sum_{j=1}^{n_i}Y_i=2|a_i,g_{i1},\ldots,g_{in_i},X_{i1},\ldots,X_{in_i})}\\
&=\frac{\sum_{j=1}^{n_i}\exp\{d_{ij}(g_{ij}+\beta X_{ij})\}}{\sum_{l,k\in\mathscr{R}_i}\exp\{g_{il}+g_{ik}+\beta(X_{il}+X_{ik})\}}
\end{aligned}
\tag{16}
$$

Note that this conditional likelihood only involves the random effects parameters $\mathbf{g}_i=(g_{i1},\ldots,g_{in_i})^{\mathrm{T}}$, $(i=1,\ldots,I)$ and β, while μ and a_i $(i=1,\ldots,I)$ are canceled out. Similarly, the conditional likelihood can be extended to any choices of $m_i=1,\ldots,n_i$ and $\sum_{j=1}^{n_i}Y_{ij}\geqslant m_i$.

3. BAYESIAN ESTIMATION METHOD

In a Bayesian paradigm, inferential interests lie in the posterior distribution of the fixed effects as well as the random effects. Posterior inference corresponding to the regression coefficient β is generally straightforward by choosing a normal prior and starting with the conditional likelihood we described in the previous section. The major question arising in Bayesian analysis concerns the sensitivity of the results to the chosen priors on the random effects; hence, modeling the random effects distribution is substantially more challenging in this context. In the following we discuss several choices.

3.1. Bayesian parametric modeling

The intuitive and traditional prior on the random effects is the normal distribution centered at their mean with a specific covariance matrix. More specifically, for the model in (2) corresponding to the parent–offspring familial data, we consider a bivariate normal (BVN) prior on the random effects, i.e. $\mathbf{b}_i\sim\mathrm{N}_2(\boldsymbol{\mu}_b,\Sigma_b)$. To avoid certain numerical computational problems, we reparameterize the random effects as, $\mathbf{b}_i=\boldsymbol{\mu}_b+\Sigma_b^{1/2}\mathbf{c}_i$ with the prior on $\mathbf{c}_i$ being $\mathbf{c}_i\sim\mathrm{N}_2(\mathbf{0},\mathbf{I}_2)$, $\mathbf{I}_n$ denoting an $n\times n$ identity matrix. We consider normal priors on μ_{b1} and μ_{b2}, $\log\{\rho^*_{cc}/(1-\rho^*_{cc})\}\sim\mathrm{N}(m_{cc},s^2_{cc})$, and the following set of priors on the hyperparameters of the covariance matrix Σ_b:

$$
\begin{aligned}
\log(\sigma_p^2)&\sim\mathrm{N}(m_p,s_p^2)\\
\log(\sigma_c^2)&\sim\mathrm{N}(m_c,s_c^2)\\
\log\left\{\frac{\rho_{pc}}{1-\rho_{pc}}\right\}&\sim\mathrm{N}(m_{pc},s_{pc}^2)
\end{aligned}
\tag{17}
$$

Generally, the correlation parameters could assume any values between -1 and 1; however, within a family, the correlations are positive, which is reflected through our prior structure. Furthermore, we will note in our real data analysis that we can center the priors corresponding to the correlation

DOI: 10.1002/sim

parameters on the genetic covariance structure [14] proposed, but allow possible uncertainty around that plausible structure.

Similarly, for the mixed model of [13], we consider $\mathbf{c}_i = (\Sigma_g^{n_i})^{-1/2}\mathbf{g}_i \sim \mathrm{N}_{n_i}(\mathbf{0}, \mathbf{I}_{n_i})$, with the similar set of priors (17) and $\log\{\rho_{cc}/(1-\rho_{cc})\} \sim \mathrm{N}(m_{cc}, s_{cc}^2)$ on the hyperparameters of the covariance matrix $\Sigma_g^{n_i}$ but with the restriction on ρ_{pc} and ρ_{cc} to keep $\Sigma_g^{n_i}$ positive definite.

3.2. Bayesian nonparametric modeling

The nonparametric Bayesian approach for modeling the distribution of random effects $\mathbf{b}_i$ starts by specifying a prior distribution on the space of all possible distribution functions for the random effects. This can be accomplished by assuming a DP prior on the space of random effects distributions. The construction and properties of DP priors are discussed in [16, 27]. The practical application of such priors to the random effect context has often focused on longitudinal data [17, 18, 20]. In the following, we describe two modeling approaches that we have implemented.

The DP prior on the random effects distribution: We first introduce a DP prior directly on $\mathbf{b}_i$:

$$\mathbf{b}_i | G \overset{\text{iid}}{\sim} G, \quad i = 1, \ldots, I$$

$$G | \alpha, G_0 \sim \mathrm{DP}(\alpha, G_0)$$

where G serving as a prior on $\mathbf{b}_i, i = 1, \ldots, I$, is itself a random probability measure. We assume that G is realization of a DP with a scalar precision parameter $\alpha \geqslant 0$ and a base measure (or base prior) $G_0 = E[G]$. In practice, the base measure G_0 specifies one's 'best guess' of an underlying model of the variation in $\mathbf{b}_i$'s, and α specifies the extent to which G_0 holds. Loosely speaking, DP may be thought of as a prior on a function space, the space of all prior distribution functions with common support. In this sense, DP specifies prior uncertainty in G, which we consider as a normal distribution. The precision parameter α corresponding to DP prior plays an especially important role in the distribution of $\mathbf{b}_i$'s: higher values of α lead to a higher probability of more unique values of $\mathbf{b}_i$'s. Following [28], we assume a Gamma(a_α, b_α) on α and follow the resampling scheme proposed therein.

Dirichlet process mixture (DPM) model for the random effects distribution: The DPM structure on $\mathbf{b}_i$ can be expressed by the following hierarchical description:

$$\mathbf{b}_i | \boldsymbol{\theta}_i \sim F(\boldsymbol{\theta}_i)$$

$$\boldsymbol{\theta}_i | G \overset{\text{iid}}{\sim} G, \quad i = 1, \ldots, I$$

$$G | \alpha, G_0 \sim \mathrm{DP}(\alpha, G_0)$$

Now, the base measure G_0 is defined as, under G, $\boldsymbol{\theta}_i$ follows some distribution. We again assume a Gamma(a_α, b_α) on α.

A property of the DP prior is that the random probability measure G is almost surely discrete, leading to the following properties that reinterpret the DPM model structure (see [29]): (i) Any realization of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_I$ generated from G lies in a set of $K \leqslant I$ distinct values; (ii) These K distinct values are random samples from the base prior G_0; (iii) $K \leqslant I$ is drawn from an implicitly determined prior distribution depending on the precision parameter α and I; and (iv) Given $K \leqslant I$, the I values are selected from the set according to a uniform multinomial distribution. There are two extreme cases that lead DPM to the fully parametric case: (1) As $\alpha \to \infty$, $K \to I$ and $G \to G_0$,

DOI: 10.1002/sim

so that the base measure is the prior distribution for $\boldsymbol{\theta}_i$; (2) As $\alpha \to 0$, $K \to 1$ and all $\boldsymbol{\theta}_i$ $(i = 1, \ldots, I)$ equal.

Note that by assuming DP priors on random effects distribution, we assume the random effects for different families could share the same values. But if the random effect vector for different families have different sizes, the DP priors as specified above cannot be assumed. However by DPM, each random effect follows the same family of continuous distributions but has its own hyperparameters, e.g. σ^2, and we consider DP priors on those hyperparameters. Thus, under DPM we avoid the inherent discreteness of the random effects distribution as is true for the direct DP prior.

Now we specifically discuss how to apply DP and DPM to the proposed models in Section 2.

A. Nonparametric modeling for parent–offspring familial data:

(1) *DP Model*: To avoid some numerical computational problems, we first reparameterize $\mathbf{b}_i = \boldsymbol{\mu}_b + \Sigma_b^{1/2}\mathbf{c}_i$ and introduce a DP prior on $\mathbf{c}_i$ in the following hierarchical manner:

$$\mathbf{c}_i | G \overset{\text{iid}}{\sim} G, \quad i = 1, \ldots, I$$

$$G | \alpha, G_0 \sim \text{DP}(\alpha, G_0)$$

$$G_0 \sim \text{N}_2(\mathbf{0}, \mathbf{I}_2)$$

The priors on $\boldsymbol{\mu}_b$ and Σ_b are the same as in Section 3.1.

(2) *DPM Model*: First, we choose a prior for each $\mathbf{b}_i$ as $\mathbf{b}_i \sim \text{N}_2(\boldsymbol{\mu}_{bi}, \Sigma_{bi})$, where

$$\boldsymbol{\mu}_{bi} = \begin{pmatrix} \mu_{b1i} \\ \mu_{b2i} \end{pmatrix} \quad \text{and} \quad \Sigma_{bi} = \begin{pmatrix} \sigma_{pi}^2 & \rho_{pc}\sigma_{pi}\sigma_{ci} \\ \rho_{pc}\sigma_{pi}\sigma_{ci} & \sigma_{ci}^2 \end{pmatrix}$$

We consider DP prior on $\boldsymbol{\theta}_i = (\mu_{b1i}, \mu_{b2i}, \sigma_{pi}^2, \sigma_{ci}^2)$:

$$\mu_{b1i}, \mu_{b2i}, \log(\sigma_{pi}^2), \log(\sigma_{ci}^2) | G \overset{\text{iid}}{\sim} G, \quad i = 1, \ldots, I$$

$$G | \alpha, G_0 \sim \text{DP}(\alpha, G_0)$$

Now, the base measure G_0 is defined as, under G, μ_{b1i}, μ_{b2i}, $\log(\sigma_{pi}^2)$ and $\log(\sigma_{ci}^2)$ are normal distributions. Under both models, the prior to ρ_{pc} is unchanged. Note that when $\boldsymbol{\mu}_b \neq 0$, β is not purely interpreted as the fixed effect.

B. Nonparametric modeling for the mixed model in [*13*]

For illustration purposes, here we consider the pedigree structure as in (7).

(1) *DP Prior*: To apply DP prior to this model, we need to restrict all families to have the same size n, i.e. n_i's equal to n. We also reparameterize $\mathbf{g}_i = (\Sigma_g^n)^{1/2}\mathbf{c}_i$ and introduce DP on $\mathbf{c}_i$:

$$\mathbf{c}_i | G \overset{\text{iid}}{\sim} G, \quad i = 1, \ldots, I$$

$$G | \alpha, G_0 \sim \text{DP}(\alpha, G_0)$$

$$G_0 \sim N_n(\mathbf{0}, \mathbf{I}_n)$$

with the priors on hyperparameters of Σ_g^n as in (17).

DOI: 10.1002/sim

(2) *DPM Prior*: First, we choose a prior for each $\mathbf{g}_i$ as $\mathbf{g}_i \sim \mathrm{N}_{n_i}(\mathbf{0}, \Sigma_{g_i})$. If each family has two parents and $n_i - 2$ offspring, then Σ_{g_i} has the same structure as in (7) except that each family has its own σ^2_{pi} and σ^2_{ci}, assuming that the first two members are parents and the rest are offspring:

$$\Sigma_{g_i} = \begin{pmatrix} \sigma^2_{pi} & 0 & \rho_{pc}\sigma_{pi}\sigma_{ci} & \cdots & \rho_{pc}\sigma_{pi}\sigma_{ci} \\ 0 & \sigma^2_{pi} & \rho_{pc}\sigma_{pi}\sigma_{ci} & \cdots & \rho_{pc}\sigma_{pi}\sigma_{ci} \\ \rho_{pc}\sigma_{pi}\sigma_{ci} & \rho_{pc}\sigma_{pi}\sigma_{ci} & \sigma^2_{ci} & \cdots & \rho_{cc}\sigma^2_{ci} \\ \vdots & \vdots & \cdots & \ddots & \vdots \\ \rho_{pc}\sigma_{pi}\sigma_{ci} & \rho_{pc}\sigma_{pi}\sigma_{ci} & \rho_{cc}\sigma^2_{ci} & \cdots & \sigma^2_{ci} \end{pmatrix} \tag{18}$$

Then, we consider a DP prior on $\boldsymbol{\theta}_i = (\sigma^2_{pi}, \sigma^2_{ci})$:

$$\log(\sigma^2_{pi}), \log(\sigma^2_{ci}) | G \overset{\text{iid}}{\sim} G, \quad i = 1, \ldots, I$$

$$G | \alpha, G_0 \sim \mathrm{DP}(\alpha, G_0)$$

Now, the base measure G_0 is defined as, under G, $\log(\sigma^2_{pi})$ and $\log(\sigma^2_{ci})$ are normal distributions. Under both models, the priors to ρ_{pc} and ρ_{cc} are unchanged as in parametric modeling.

None of the full conditional distributions follows a standard distributional form; hence, posterior inference is made by using the MCMC numerical integration technique. To update the parameters in DP or DPM, we use Algorithm 5 prescribed by Neal [30]. We describe the computational details of our algorithm in the Appendix.

4. EXAMPLE: ANALYSIS OF PROSTATE CANCER DATA

The data set is selected from families participating in the PCGP. The PCGP was initiated in 1995 to define the molecular basis of hereditary prostate cancer, including families with one or more identified cases of prostate cancer. From this database we selected 46 families, (i) with pedigree sizes ranging from 4 to 6, (ii) which had at least one affected family member and (iii) which only presented relationships of brother–brother and/or father–son. There are a total of 205 observations, among which 191 are white and 14 are black/African American. All family members have been tested for prostate cancer with serum prostate-specific antigen (PSA) measurement. For affected members we considered the last available PSA measurement before diagnosis of prostate cancer, whereas for unaffected family members we considered the most recent PSA measurement as the covariate of interest. In the original data set, about 20 per cent subjects were missing PSA measurement. We imputed the missing PSA values based on disease status, relationship to proband and age. We also noticed that the distribution of PSA values is right skewed with several extreme large values; thus, we analyze the data with PSA values transformed to log scale.

We analyze the data by applying the Bayesian approach to the conditional likelihood, which corresponds to the mixed model (5) [13]. This conditional likelihood adjusts for the ascertainment, at least one affected member in each family, i.e. $\sum_{j=1}^{n_i} Y_{ij} \geqslant 1$. Without loss of generality and

DOI: 10.1002/sim

assuming the first observation in the family is the father, the covariance matrix of $\mathbf{g}_i$ has the following structure:

$$\Sigma_g^{1,n_i} = \begin{pmatrix} \sigma_p^2 & \rho_{pc}\sigma_p\sigma_c & \cdots & \rho_{pc}\sigma_p\sigma_c \\ \rho_{pc}\sigma_p\sigma_c & \sigma_c^2 & \cdots & \rho_{cc}\sigma_c^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{pc}\sigma_p\sigma_c & \rho_{cc}\sigma_c^2 & \cdots & \sigma_c^2 \end{pmatrix} \tag{19}$$

for families which have information of the father and $n_i - 1$ sons with $n_i = 4, 5, 6$; and

$$\Sigma_g^{2,n_i} = \sigma_c^2 \begin{pmatrix} 1 & \rho_{cc} & \cdots & \rho_{cc} \\ \rho_{cc} & 1 & \cdots & \rho_{cc} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{cc} & \rho_{cc} & \cdots & 1 \end{pmatrix} \tag{20}$$

for families which only have brother–brother relationship with $n_i = 4, 5, 6$.

We are interested in both the effect of PSA measurement on prostate cancer and the genetic correlations among the family members. We consider $\mathrm{N}(0, 4)$ as the prior of β, while for the random effect $\mathbf{g}_i$, we consider both parametric and nonparametric modeling. For parametric modeling, we consider a multivariate normal (MVN) distribution with mean zero, and (19) or (20) as a covariance matrix according to the pedigree structures, denoting this model as mMVN. For the priors on the hyperparameters, we consider the priors on σ_p^2 and σ_c^2 in such a way that they could range from 0.4 to 7.4, a reasonable range for a variance but with a relative large variability. However, for the correlation parameters, following [14] and assuming no dominance component variance, we center the priors on 0.5, which corresponds to the correlation between the first-degree relatives, but allow some uncertainty through the stochastic hierarchy. Thus, $\log(\sigma_p^2)$ (or $\log(\sigma_c^2)$) $\sim \mathrm{N}(0.5, 0.25)$. The priors on the correlation parameters are truncated normal distributions: $\mathrm{logit}(\rho_{pc}) \sim \mathrm{N}(0, 0.25)$ and $\mathrm{logit}(\rho_{cc}) \sim \mathrm{N}(0, 0.25)$ with the restriction $\rho_{pc} - \rho_{cc} < (1 - \rho_{cc})/4$ to keep $\Sigma_g^{n_i}$ positive definite. For nonparametric modeling (denoted as mDPM), we consider DPM on $\mathbf{g}_i$; hence, the only changes are the priors on σ_{pi}^2 and σ_{ci}^2. Because not all the families have the father information, we separately put DP priors on σ_{pi}^2 and σ_{ci}^2:

$$\begin{aligned} \sigma_{pi}^2 | G_p &\overset{\text{iid}}{\sim} G_p, \quad i = 1, \ldots, 29 \\ G_p | \alpha_p, G_{p0} &\sim \mathrm{DP}(\alpha_p, G_{p0}) \\ \sigma_{ci}^2 | G_c &\overset{\text{iid}}{\sim} G_c, \quad i = 1, \ldots, 46 \\ G_c | \alpha_c, G_{c0} &\sim \mathrm{DP}(\alpha_c, G_{c0}) \end{aligned}$$

with both the base measures G_{p0} and G_{c0} being $\mathrm{lognormal}(0.5, 0.25)$ and α_p (and α_c) $\sim \mathrm{Gamma}(1, 2)$.

We also directly apply [13] proposed variance–covariance matrix in (6), $\sigma_g^2 \mathbf{R}_i$, with unknown σ_g^2 and known correlation matrices $\mathbf{R}_i$. Note that the data only involves first-degree relatives; thus,

DOI: 10.1002/sim

the correlation matrices $\mathbf{R}_i$'s are fixed with diagonal elements equal to 1 and others being 0.5, whereas the dimensions of these matrices range from 4 to 6. Similar to the above, we consider both parametric (pMVN) and nonparametric modeling (pDPM) for the random effects $\mathbf{g}_i$, which in fact reduce to the parametric choice of priors $\log(\sigma_g^2) \sim N(0.5, 0.25)$ and nonparametric choice of priors

$$\sigma_{gi}^2 | G_g \overset{\text{iid}}{\sim} G_g, \quad i = 1, \ldots, 46$$

$$G_g | \alpha_g, G_{g0} \sim \text{DP}(\alpha_g, G_{g0})$$

with the base measures G_{g0} being lognormal$(0.5, 0.25)$ and $\alpha_g \sim$Gamma$(1, 2)$.

To compare the results with the MLEs obtained by the traditional CLR method (denoted by CMLE), we additionally analyze the data applying a logistic regression model with PSA being the predictor, which is stratified by families and obtain the estimate of the covariate (PSA measurement) effect by maximizing the conditional likelihood. This is done by implementing the *clogit* function in the package *survival* in R (http://www.r-project.org/) by maximizing the conditional likelihood, treating the families as strata.

The corresponding results are presented in Table II. The traditional conditional maximum likelihood method is designed to obtain the estimate of the covariate effect β, while all the Bayesian methods can additionally obtain the inference of the variance–covariance matrix of the random effects. Note that the CMLE of β is much larger with a larger standard error, indicating that there may be appreciable genetic variability captured by the Bayesian methods.

Among the results from the four Bayesian methods, we observe that, with our nonparametric model (mDPM and pDPM), the estimates of β are relatively smaller and have smaller posterior deviance. But comparing the results from mMVN and mDPM (with our proposed covariance matrix) with pMVN and pDPM (with Pfeiffer *et al.* [13] proposed covariance matrix) correspondingly, there is little difference in the estimate of β, which is probably because the variances of each generation do not have much difference (estimates of σ_p^2 and σ_c^2 are 1.57 and 1.62, respectively, while the estimate of σ_g^2 is 1.43) and the genetic correlations are 0.54 and 0.58, which are close to predefined correlation 0.5 in [13]. The nonparametric and parametric models perform comparably to capture the random effect distribution; however, instead of providing the estimates of variances (σ_{gi}^2, or σ_{pi}^2 and σ_{ci}^2), DPM (both mDPM and pDPM) selects the number of distinct values of variances (K's) in a data-adaptive way depending on the extent of family-specific effects on the random effects. For instance, in mDPM, the number of distinct σ_{ci}^2's, K_c, is equal to 36 (<46, the number of families), which means that not all the families have the distinct effects on σ_{ci}^2's, while the large α_c value, 86.53, tells us that there is large variation in the family-specific effects on σ_{ci}^2's.

In summarizing the results, we observe that there is a large increase in the risk of prostate cancer for the people with a higher PSA measurement. The estimated odds ratio obtained by mDPM is $\exp(1.38) = 3.97$ for one unit increase in log (PSA measurement). All the five models present the statistical significance in the effect of PSA measurement on prostate cancer, though the estimated odds ratio obtained by CLR method is almost unbelievably astronomic, $\exp(3.26) = 26.05$. The estimated genetic correlation between father and son ρ_{pc} is 0.54, which is slightly smaller than the correlation between siblings $\rho_{cc} = 0.58$, and both estimated correlations are larger than those predefined, 0.5, as in [13].

DOI: 10.1002/sim

Table II. The analysis results of the partial (46 families) prostate cancer data of the University of Michigan Prostate Cancer Genetics Project.

Model	Parameter estimates			
CMLE*	β=3.26 (0.61) [2.06,4.46]			
pMVN†	β=1.49 (0.22) [1.14,1.97]	σ_g^2=1.43 (0.75) [0.44,3.27]		
pDPM‡	β=1.42 (0.20) [1.06,1.82]			
	K_g=35.73 (3.30) [29,42]	α_g=87.95 (16.81) [56.89,123.76]		
mMVN§	β=1.51 (0.23) [1.11,2.11]	σ_p^2=1.57 (0.72) [0.53,3.14]	σ_c^2=1.62 (0.74) [0.65,3.37]	
	ρ_{pc}=0.49 (0.12) [0.25,0.69]	ρ_{cc}=0.56 (0.11) [0.32,0.76]		
mDPM¶	β=1.38 (0.20) [0.96,1.74]	ρ_{pc}=0.54 (0.11) [0.37,0.67]	ρ_{cc}=0.58 (0.10) [0.42,0.66]	
	K_p=22.74 (2.77) [17,27]	α_p=58.96 (14.23) [34.70,91.07]	K_c=35.49 (3.40) [28,41]	α_c=86.53 (16.22) [57.44,121.83]

The results, except those obtained by CMLE, are attained by applying the conditional likelihood based on the mixed model of [13]. The results presented are the estimates with standard error or standard posterior deviance in parentheses '()', and 95 per cent confidence interval or highest posterior density (HPD) interval in bracket '[]'.

*CMLE: The method of maximizing the conditional likelihood based on the traditional logistic regression model stratified by families. It is done by implementing *clogit* in R.

†pMVN: The proposed parametric Bayesian method with the multivariate normal prior on random effects, which have [13] proposed covariance matrix.

‡pDPM: The proposed nonparametric Bayesian method with the Dirichlet process mixture prior on random effects, which have [13] proposed covariance matrix.

§mMVN: The proposed parametric Bayesian method with the multivariate normal prior on random effects, which have the modified covariance matrix.

¶mDPM: The proposed nonparametric Bayesian method with the Dirichlet process mixture prior on random effects, which have the modified covariance matrix.

Figure 1 shows the posterior distributions of the parameters based on mDPM model. Figure 2 presents histograms for the predictive density of $g_{n+1,1}$ and $g_{n+1,2}$ given the data based on both pDPM and mDPM models. (Since in the data set, we either have the first subject is father and rest are sons, or have all subjects in the same generation; hence, we just pick two elements $g_{n+1,1}$ and $g_{n+1,2}$ from the random effect vector.) Note that under DPM, for example, pDPM, $\sigma_{g_i}^2|G_g \sim G_g, i=1,\ldots,I$ with $G_g \sim \mathrm{DP}(\alpha_g G_{g0})$, $E(G_{g0}|\text{data})$ is in fact the posterior predictive distribution $p(\sigma_{g_{I+1}}^2|\text{data})$. Furthermore, in the hierarchical models, the posterior predictive distribution of the random effect $\mathbf{g}$ can be obtained based on the future draws of σ_g^2, i.e. $\sigma_{g_{I+1}}^2$. Hence for each of the last 1000 MCMC runs, we generate $\sigma_{g_{I+1}}^2$ from the corresponding predictive distribution, then draw $\mathbf{g}_{I+1}$ from an MVN with mean equal to 0 and an according covariance matrix by plugging $\sigma_{g_{I+1}}^2$. The histogram is based on these 1000 generations of $g_{I+1,1}$ ($g_{I+1,2}$) values. We also plot the curves based on the density of the normal distributions with the σ^2 values from the corresponding

DOI: 10.1002/sim

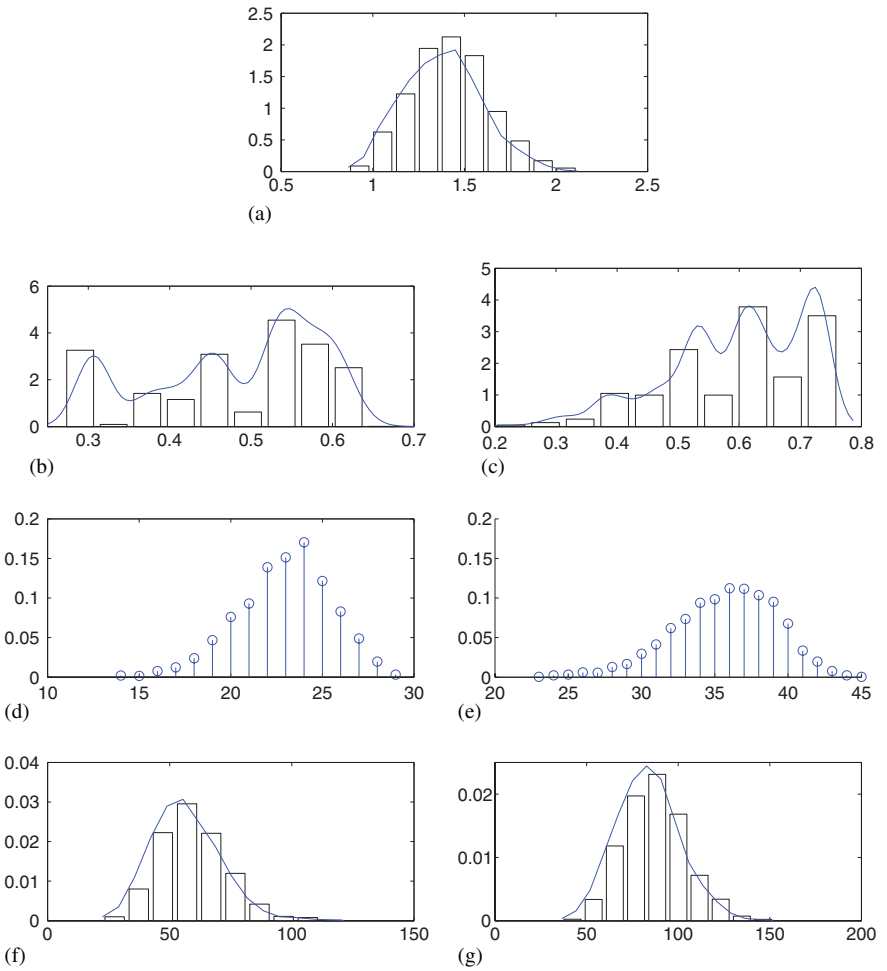

Figure 1. Posterior distribution of the parameters for prostate cancer data analyzed by the mixed model of Pfeiffer *et al.* [13] with the proposed Bayesian nonparametric model DPM on the random effect with the proposed covariance matrix (mDPM). Histogram of last 1000 MCMC values for each parameter overlaid with smoothed kernel density estimate: (a) posterior distribution of β; (b) posterior distribution of ρ_{pc}; (c) posterior distribution of ρ_{cc}; (d) posterior distribution of K_p; (e) posterior distribution of K_c; (f) posterior distribution of α_p; and (g) posterior distribution of α_c.

parametric Bayesian models, i.e. for pDPM, we plot N(0, 1.43) (a and b); for mDPM, N(0, 1.57) (c) and N(0, 1.62) (d).

4.1. Simulation study

SIMULATION Setting 1: Parent–offspring familial data. To illustrate our proposed methods, now we present numerical evidence in the form of simulation studies. We first consider to simulate parent–offspring familial data as stated in Section 2.1, where we assume in each family there

DOI: 10.1002/sim

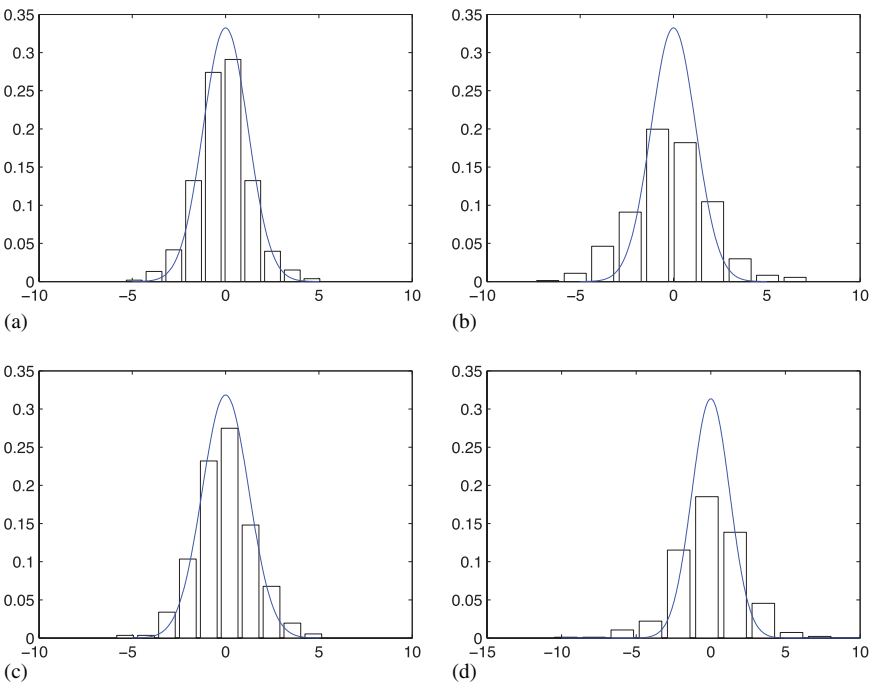

Figure 2. Posterior predictive densities of $g_{n+1,1}$ and $g_{n+1,2}$ for the Prostate cancer data analyzed by the mixed model of Pfeiffer *et al.* [13] with the proposed Bayesian nonparametric models (pDPM and mDPM). The histogram is based on $g_{n+1,1}$ and $g_{n+1,2}$ values, which are generated from the corresponding predictive distribution from each of the last 1000 MCMC runs. The curves are the densities of the $N(0, 1.43)$ (a and b), $N(0, 1.57)$ (c) and $N(0, 1.62)$ (d): (a) posterior predictive density of $g_{n+1,1}$ under pDPM[1]; (b) posterior predictive density of $g_{n+1,2}$ under pDPM[1]; (c) Posterior predictive density of $g_{n+1,1}$ under mDPM[2]; and (d) posterior predictive Density of $g_{n+1,2}$ under mDPM[2].

are two parents and two offspring, each having a binary covariate X_{ij} $(j=1,2,3,4)$ with the prevalence of 0.5. For generating the random effects $\mathbf{b}_i=(b_{i1},b_{i2})^{\mathrm{T}}$, we use the following bivariate distributions:

(1) *Mixture of three BVNs*:

$$\frac{3}{7}\mathrm{N}_2\left(\begin{pmatrix}-2\\-2\end{pmatrix},\begin{pmatrix}0.25 & 0.075\\0.075 & 0.09\end{pmatrix}\right)+\frac{1}{7}\mathrm{N}_2\left(\begin{pmatrix}-1\\-1\end{pmatrix},\begin{pmatrix}0.09 & 0.003\\0.003 & 0.04\end{pmatrix}\right)$$
$$+\frac{3}{7}\mathrm{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}0.25 & 0.075\\0.075 & 0.09\end{pmatrix}\right) \tag{21}$$

with common $\rho_{pc}=0.5$ in the variance–covariance matrix for each family.

First for each family, we generate the covariate $\mathbf{X}_i=(X_{i1},\ldots,X_{i4})^{\mathrm{T}}$ and random effects $\mathbf{b}_i$. With these information in hand, following the logistic regression model for disease risk (2) and the joint probability as shown in Table I, we generate the disease outcome for each individual. We

DOI: 10.1002/sim

set a common offspring–offspring genetic correlation $\rho^*_{cc}=0.25$ for each family. Then, we select families with exactly two diseased members from the simulated population. We consider varied scales of the fixed effect, $\beta=0, 0.5$ and 1. We simulate 200 data sets for each scenario with each data set having 100 families.

We apply the two proposed Bayesian semiparametric models, i.e. nonparametric modeling on $\mathbf{b}_i$ (DP and DPM), to the simulated data, with priors set as the following: $\beta\sim\mathrm{N}(0,9)$, $\mathrm{logit}(\rho_{pc})\sim\mathrm{N}(0,0.25)$, $\mathrm{logit}(\rho^*_{cc})\sim\mathrm{N}(-1,0.25)$, μ_b (or μ_{bi})$\sim\mathrm{N}(0,4)$, $\log(\sigma^2_p)$ (or $\log(\sigma^2_{pi})$), $\log(\sigma^2_c)$ (or $\log(\sigma^2_{ci})$)$\sim\mathrm{N}(0,0.25)$, $\alpha\sim\mathrm{Gamma}(1,2)$. To compare with Bayesian parametric model, we also analyze with the traditional prior, a BVN, i.e. $\mathbf{b}_i\sim\mathrm{N}_2(\boldsymbol{\mu}_b,\Sigma_b)$ on the random effects. We consider the following reparameterization: $\mathbf{b}_i=\boldsymbol{\mu}_b+\Sigma_b^{1/2}\mathbf{c}_i$ with the prior on $\mathbf{c}_i$ being $\mathrm{N}_2(\mathbf{0},\mathbf{I_2})$.

We set the priors on σ^2_p, σ^2_c, ρ_{pc} and ρ^*_{cc} as described above. The results in Table III are fairly clear. All three Bayesian models provide comparable results and the varied fixed effects do not have an influence on the performance of the models, even when $\beta=0$. The CLR method also provides similar results except no estimation of ρ_{pc} and ρ^*_{cc}, this is because the variances of the random effects were set relative small, though we did encounter some convergence problems.

To see the influence of the sample size (the number of families selected) as well as the effect of the variances of the random effects, we also performed simulations of 100 families and 20 families, where we fix $\beta=1$, and used the following random distribution:

(2) *Mixture of three BVNs*:

$$\frac{10}{12}\mathrm{N}_2\left(\begin{pmatrix}-4\\-4\end{pmatrix},\begin{pmatrix}2 & \sqrt{2}/2\\ \sqrt{2}/2 & 1\end{pmatrix}\right)+\frac{1}{12}\mathrm{N}_2\left(\begin{pmatrix}-1\\-1\end{pmatrix},\begin{pmatrix}1 & \sqrt{0.5}/2\\ \sqrt{0.5}/2 & 0.5\end{pmatrix}\right)$$
$$+\frac{1}{12}\mathrm{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & \sqrt{1.5}/2\\ \sqrt{1.5}/2 & 1.5\end{pmatrix}\right) \tag{22}$$

Table IV shows the corresponding results. We observe that the estimates of β obtained by the CLR method are biased toward null with larger MSEs even when the sample size is large. This phenomenon shows that larger variances in random effects could bring more biases in the estimates of β but not in the estimation of the correlation parameters. We also see that with the number of families selected increased, the point estimation of β has less biases. We notice that the Bayesian semiparametric modeling (DP and DPM) always provides better estimation of β than the Bayesian parametric modeling, and much better than the CLR method. In addition, as we expected, the larger the sample size, the smaller the posterior deviances and MSEs. Hence, when we suspect the data with large variances of the random effects, but with a relative small sample size, we would recommend to consider the Bayesian semiparametric modeling.

SIMULATION Setting 2: Mixed effects model motivated by the real data analysis. Following the mixed model of Pfeiffer *et al.* [13], we also performed a simulation study based on the real data results and the exact same genetic structure, i.e. either the families having father and rest being sons, or the families only having brother–brother relationship, with $n_i=4,5,6$. First, we draw random effects g_i following MVN distribution with mean equal to 0 and covariance as (19) or (20) according to the family structure, where we set $\rho_{pc}=0.5$, $\rho_{cc}=0.6$, σ^2_p to 2, and 1 with probabilities $\frac{1}{3}$ and $\frac{2}{3}$, respectively; and σ^2_c to 2, 1.5 and 0.5 with equal probabilities. Then

DOI: 10.1002/sim

Table III. Simulation scenarios as stated in Section 4.1 under simulation setting 1.

	Model		β	ρ_{pc}	ρ^*_{cc}
$\beta=1$	BVN*	Post mean	1.1073	0.5000	0.2470
		Post std. dev.	0.2273	0.0825	0.0500
		MSE	0.0652	6.3e−5	1.5e−4
	DP†	Post mean	1.0685	0.5006	0.2552
		Post dev.	0.2279	0.0831	0.0515
		MSE	0.0618	5.3e−5	1.2e−4
	DPM‡	Post mean	1.1105	0.4879	0.2511
		Post std. dev.	0.2263	0.0736	0.0474
		MSE	0.0640	4.4e−5	1.1e−4
	CMLE§	Mean	0.9384		
		Std. err	0.2620		
		MSE	0.0806		
$\beta=0.5$	BVN*	Post mean	0.6149	0.5139	0.2468
		Post std. dev.	0.2190	0.0829	0.0440
		MSE	0.0576	6.0e−5	1.3e−4
	DP†	Post mean	0.5645	0.5111	0.2496
		Post std. dev.	0.2138	0.0836	0.0452
		MSE	0.0526	5.3e−5	1.0e−4
	DPM‡	Post mean	0.6065	0.5073	0.2417
		Post std. dev.	0.2161	0.0745	0.0420
		MSE	0.0613	5.8e−5	1.5e−4
	CMLE§	Mean	0.5410		
		Std. err	0.2845		
		MSE	0.07822		
$\beta=0$	BVN*	Post mean	0.0407	0.5054	0.2446
		Post std. dev.	0.2146	0.0497	0.0335
		MSE	0.0459	5.7e−5	1.1e−4
	DP†	Post mean	−0.0120	0.5051	0.2488
		Post std. dev.	0.2102	0.0499	0.0341
		MSE	0.0610	6.3e−5	9.1e−5
	DPM‡	Post mean	−0.0297	0.5046	0.2444
		Post std. dev.	0.2100	0.0486	0.0336
		MSE	0.0933	5.2e−5	1.2e−4
	CMLE§	Mean	−0.0172		
		Std. err	0.2819		
		MSE	0.0758		

$\mathbf{b}_i$ generated from mixture of three bivariate normals (21), $\rho_{pc}=0.5$, $\rho^*_{cc}=0.25$ and $I=100$. The results, except those obtained by CMLE, are attained by applying the conditional likelihood based on the mixed model for parental–offspring familial data. *post mean*, *post std. dev.* and *MSE* denote the average of posterior mean, standard deviance estimate and the estimated mean-squared error based on 200 replications.

*BVN: The proposed parametric Bayesian method with the bivariate normal prior on random effects.

†DP: The proposed nonparametric Bayesian method with the Dirichlet process prior on random effects.

‡DPM: The proposed nonparametric Bayesian method with the Dirichlet process mixture prior on random effects.

§CMLE: The method of maximizing the conditional likelihood based on the traditional logistic regression model stratified by families. It is done by implementing *clogit* in R.

DOI: 10.1002/sim

Table IV. Simulation scenarios as stated in Section 4.1 under simulation setting 1.

	Model		$\beta=1$	$\rho_{pc}=0.5$	$\rho^*_{cc}=0.25$
$I=20$	BVN*	Post mean	1.2037	0.4986	0.2623
		Post std. dev.	0.5429	0.1161	0.0831
		MSE	0.4971	3.0e−4	0.0013
	DP†	Post mean	1.1148	0.4984	0.2692
		Post std. dev.	0.5254	0.1179	0.0868
		MSE	0.4436	1.9e−4	0.0014
	DPM‡	Post mean	1.1407	0.5048	0.2654
		Post std. dev.	0.5437	0.1161	0.0833
		MSE	0.4566	2.7e−4	0.0018
	CMLE§	Mean	0.8307		
		Std. err	0.6519		
		MSE	0.6009		
$I=100$	BVN*	Post mean	1.0798	0.4920	0.2484
		Post std. dev.	0.2394	0.1207	0.0640
		MSE	0.0662	8.7e−4	0.0020
	DP†	Post mean	0.9504	0.5028	0.2769
		Post std. dev.	0.2123	0.1204	0.0794
		MSE	0.0516	1.5e−4	0.0021
	DPM‡	Post mean	1.0383	0.4903	0.2690
		Post std. dev.	0.2298	0.1113	0.0667
		MSE	0.0589	0.0011	0.0025
	CMLE§	Mean	0.8673		
		Std. err	0.2403		
		MSE	0.0739		

$\mathbf{b}_i$ generated from mixture of three bivariate normals in (22), $\beta=1$, $\rho_{pc}=0.5$ and $\rho^*_{cc}=0.25$, but different sample sizes $I=20$ and 100. The results, except those obtained by CMLE§, are attained by applying the conditional likelihood based on the mixed model for parental–offspring familial data. *post mean*, *post std. dev.* and *MSE* denote the average of posterior mean, standard deviance estimate and the estimated mean-squared error based on 200 replications, respectively.

*BVN: The proposed parametric Bayesian method with the bivariate normal prior on random effects.

†DP: The proposed nonparametric Bayesian method with the Dirichlet process prior on random effects.

‡DPM: The proposed nonparametric Bayesian method with the Dirichlet process mixture prior on random effects.

§CMLE: The method of maximizing the conditional likelihood based on the traditional logistic regression model stratified by families. It is done by implementing *clogit* in R.

based on $P(Y_{ij}=1|g_{ij}, X_{ij})=\{1+\exp(\mu+\beta X_{ij}+g_{ij})\}^{-1}$, we simulate the disease status Y_{ij} by setting $\beta=1.5$, $\mu=-2$ and $X=\log(\text{PSA})$ (obtained from the real data). Note that we only keep the families that have at least one diseased subject, and we have 46 families and 205 subjects as in the real data set. We analyze the simulated data by implementing the five models mentioned above, and perform 200 such simulations. The results are presented in Table V. Our proposed Bayesian methods provide more accurate estimates of the fixed effect β with smaller posterior standard deviances and MSEs. By using the modified covariance structure, we can additionally obtain inference on the correlation parameters. However, due to the limited number of simulated data sets, the results of the simulation study should be evaluated with caution.

DOI: 10.1002/sim

Table V. Simulation scenarios as stated in Section 4.1 under simulation setting 2.

Model		$\beta=1.5$	$\rho_{pc}=0.5$	$\rho_{cc}=0.6$
CMLE*	Mean	1.4223		
	Std. err	0.2515		
	MSE	0.0830		
pMVN†	Post mean	1.4542		
	Post std. dev.	0.2289		
	MSE	0.0367		
mMVN‡	Post mean	1.4635	0.5001	0.5648
	Post std. dev.	0.2214	0.0490	0.0487
	MSE	0.0393	3.4e−5	0.0057
pDPM§	Post mean	1.4794		
	Post std. dev.	0.2090		
	MSE	0.0435		
mDPM¶	Post mean	1.5564	0.4981	0.5897
	Post std. dev.	0.2359	0.0292	0.0301
	MSE	0.0517	8.2e−5	0.0046

The results, except those obtained by CMLE, are attained by applying the conditional likelihood based on the mixed model of [13]. *post mean*, *post std. dev.* and *MSE* denote the average of posterior mean, standard deviance estimate and the estimated mean-squared error based on 200 replications.

*CMLE: The method of maximizing the conditional likelihood based on the traditional logistic regression model stratified by families. It is done by implementing *clogit* in R.

†pMVN: The proposed parametric Bayesian method with the multivariate normal prior on random effects, which have [13] proposed covariance matrix.

‡pDPM: The proposed nonparametric Bayesian method with the Dirichlet process mixture prior on random effects, which have [13] proposed covariance matrix.

§mMVN: The proposed parametric Bayesian method with the multivariate normal prior on random effects, which have the modified covariance matrix.

¶mDPM: The proposed nonparametric Bayesian method with the Dirichlet process mixture prior on random effects, which have the modified covariance matrix.

5. DISCUSSION

In this paper, we have applied both Bayesian parametric and nonparametric techniques to address an important class of models, the random effects model for family-based association studies. The contribution of this paper revolves around building different flexible Bayesian models for the distribution random effects and capturing various genetic correlation structures.

The parent–offspring familial model is a generation-specific model, explaining the genetic information but not isolating the familial effects. We obtain inferences regarding the variances related to each generation as well as the different genetic correlations. Srivastava [31] has proposed maximum likelihood estimation for the interclass correlation in familial data, and Srivastava *et al.* [32] derived asymptotical normal estimators of the interclass and intraclass correlations. In this paper, besides presenting an alternative approach in Bayesian domain to obtain the estimates of both intraclass and interclass correlations, we also introduce the intraclass correlations through outcomes (Y-values) directly. The nice thing about this model is that the dimension of the random effects is small and always fixed, i.e. 2, which would reduce computational complexity. Based on the two-level mixed model by Pfeiffer *et al.* [13], introducing genetic correction parameters

DOI: 10.1002/sim

to a more general family-specific mixed model is also appealing. This general model can handle different structures with more flexible degree of kinship and larger pedigrees. Instead of assuming fixed correlation values, we proposed a less stringent genetic correlation structure and are able to estimate the degree to which family members are correlated.

Moreover, our Bayesian technique involved specifying a nonparametric prior for the distribution of the random effects by using a DP prior. Although assuming a DP prior directly on the random effects distribution is attractive and could obtain the estimates of parameters of the covariance matrix, such as generation-specific variances and correlations, the proposed DPM is also competitive. In our applications with the DPM and unequal family sizes, we apply the DP prior on the normal variances, which brings more flexibility to modeling the distributions of the random effects. In fact, we can also consider the DP prior on the correlation parameters if we suspect the uncertainty in the correlations, which is not illustrated in this paper.

To conclude, due to lack of information on the genetic effects parameters in an ascertainment corrected likelihood, a Bayesian approach that can possibly incorporate information on the genetic parameters is a useful tool. The advantages are greater flexibility and more precise estimation in the presence of credible prior information. The potential disadvantages include sensitivity to the prior and computational burden. The lack of information on the random effect-related parameters often leads to sensitivity of these parameter estimates subject to prior choices; however, the inference on the risk parameters generally remains robust.

APPENDIX A

A.1. Calculation of Table I

Suppose X and Y are binary variables. Let $\Pr(Y=1)=P_X$ and $\Pr(Y=1)=P_Y$; thus, $E(X)=P_X$, $E(Y)=P_Y$, $\mathrm{var}(X)=P_X(1-P_X)$ and $\mathrm{var}(Y)=P_Y(1-P_Y)$. Since

$$\rho_{XY}=\frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}=\frac{E(XY)-E(X)E(Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}$$

we have

$$\begin{aligned}
P(X=1,Y=1)&=E(XY)=P_XP_Y+\rho_{XY}\sqrt{P_X(1-P_X)P_Y(1-P_Y)}\\
P(X=1,Y=0)&=P(X=1)-P(X=1,Y=1)=P_X(1-P_Y)-\rho_{XY}\sqrt{P_X(1-P_X)P_Y(1-P_Y)}\\
P(X=0,Y=1)&=P(Y=1)-P(X=1,Y=1)=(1-P_X)P_Y-\rho_{XY}\sqrt{P_X(1-P_X)P_Y(1-P_Y)}\\
P(X=0,Y=0)&=1-P(X=1,Y=0)-P(X=0,Y=1)-P(X=1,Y=1)\\
&=(1-P_X)(1-P_Y)+\rho_{XY}\sqrt{P_X(1-P_X)P_Y(1-P_Y)}
\end{aligned}$$

A.2. Extension to the case of three offspring for the mixed model for parent–offspring familial data

To extend to the case of three offspring, the key part is to obtain the joint probability of three binary random variables. Suppose there are three binary variables X, Y and Z, following the same definitions as in A.1 and letting $\Pr(Z=1)=P_Z$. Note that the results in A.1 work for either of the two binary variables.

DOI: 10.1002/sim

We define $A=1$ if $X=1$ and $Y=1$, and $A=0$ otherwise; hence, A is also a binary variable with $P_A=\Pr(A=1)=P(X=1,Y=1)$, which can be obtained as shown in A.1. We denote the correlation between A and Z as ρ_{AZ}; thus, we have

$$P(X=1,Y=1,Z=1)=P(A=1,Z=1)=P_AP_Z+\rho_{AZ}\sqrt{P_A(1-P_A)P_Z(1-P_Z)}$$

$$P(X=1,Y=1,Z=0)=P(A=1,Z=0)=P_A(1-P_Z)-\rho_{AZ}\sqrt{P_A(1-P_A)P_Z(1-P_Z)}$$

Similarly, we define

$$B=1 \quad \text{if } X=1, \quad Y=0 \text{ and } B=0 \quad \text{otherwise}$$

$$C=1 \quad \text{if } X=0, \quad Y=1 \text{ and } C=0 \quad \text{otherwise}$$

$$D=1 \quad \text{if } X=0, \quad Y=0 \text{ and } D=0 \quad \text{otherwise}$$

with the correlations with Z being ρ_{BZ}, ρ_{CZ} and ρ_{DZ}, respectively. Then, we can obtain the other six joint probabilities.

Note that $P(A=1,Z=1)=P(X=1,Z=1)-P(B=1,Z=1)$; we have

$$\rho_{BZ}=\frac{(P_A+P_B-P_X)P_Z+\rho_{AZ}\sqrt{P_A(1-P_A)P_Z(1-P_Z)}-\rho\sqrt{P_X(1-P_X)P_Z(1-P_Z)}}{\sqrt{P_B(1-P_B)P_Z(1-P_Z)}} \tag{A1}$$

and $P(A=1,Z=1)=P(Y=1,Z=1)-P(C=1,Z=1)$, we have

$$\rho_{CZ}=\frac{(P_A+P_C-P_Y)P_Z+\rho_{AZ}\sqrt{P_A(1-P_A)P_Z(1-P_Z)}-\rho\sqrt{P_Y(1-P_Y)P_Z(1-P_Z)}}{\sqrt{P_C(1-P_C)P_Z(1-P_Z)}} \tag{A2}$$

We also have $P_A+P_B+P_C+P_D=1$, $0<P_Z<1$, and

$$\begin{aligned}P_Z&=\Pr(Z=1)\\&=P(A=1,Z=1)+P(B=1,Z=1)+P(C=1,Z=1)+P(D=1,Z=1)\\&=(P_A+P_B+P_C+P_D)P_Z+\sqrt{P_Z(1-P_Z)}(\rho_{AZ}\sqrt{P_A(1-P_A)}+\rho_{BZ}\sqrt{P_B(1-P_B)}\\&\quad+\rho_{CZ}\sqrt{P_C(1-P_C)}+\rho_{DZ}\sqrt{P_D(1-P_D)})\end{aligned}$$

Therefore,

$$\rho_{AZ}\sqrt{P_A(1-P_A)}+\rho_{BZ}\sqrt{P_B(1-P_B)}+\rho_{CZ}\sqrt{P_C(1-P_C)}+\rho_{DZ}\sqrt{P_D(1-P_D)}=0 \tag{A3}$$

Note that now we have one more parameter ρ_{AZ} in the likelihood, which though is not a parameter of interest. Theoretically, we can implement the same calculation to the case of four offspring, or even more, though we would encounter very complicated formulations with more parameters.

A.3. Computational details of the proposed algorithm

Drawing observations from the posterior of DP, following Algorithm 5 in [*30*]: The basic model applies to data $y_1,\ldots,y_I$, where y_i $(i=1,\ldots,I)$ may be multivariate. We model the distribution

DOI: 10.1002/sim

from which y_i is drawn as a mixture of distributions of the form $F(y|\boldsymbol{\phi})$, with the mixing distribution over $\boldsymbol{\phi}$ being G. Hence, we give the following model:

$$\begin{aligned} y_i|\boldsymbol{\phi}_i &\sim F(y_i|\boldsymbol{\phi}_i) \\ \boldsymbol{\phi}_i|G &\sim G \\ G &\sim \mathrm{DP}(\alpha G_0) \end{aligned} \tag{A4}$$

Let $\boldsymbol{\omega}=(\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_K)$ denote the set of distinct $\boldsymbol{\phi}_i$'s, where $K\leqslant I$ is the number of distinct elements in $\boldsymbol{\phi}=(\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_I)$. Let $\mathbf{s}=(s_1,\ldots,s_I)$ denote the vector of configuration indicators defined by $s_i=k$ if and only if $\boldsymbol{\phi}_i=\boldsymbol{\omega}_k$, $i=1,\ldots,I$. In this connection we use the term 'cluster' where kth cluster is defined as $I_k=\{i:s_i=k\}$ and define n_k as the size of the kth cluster. We use $-i$ to denote the situation when the observation i is removed. For example, $n_{-i,k}$ is the size of the kth cluster after removing $\boldsymbol{\phi}_i$.

Let the state of the Markov chain consist of $\mathbf{s}=(s_1,\ldots,s_I)$ and $\boldsymbol{\phi}=(\boldsymbol{\phi}_s:s\in\{s_1,\ldots,s_I\})$. Repeatedly the sample is as follows:

- For $i=1,\ldots,I$, repeat the following update of s_i R times: Draw a candidate s_i^* from the conditional prior for s_i given by

$$\begin{aligned} P(s_i=s|s_{-i}) &= \frac{n_{-i,s}}{I-1+\alpha} \quad \text{if } s=s_j \quad \text{for some } j \\ P(s_i\neq s_j|s_{-i}) &= \frac{\alpha}{I-1+\alpha} \quad \text{for all } j \end{aligned} \tag{A5}$$

  If this s_i^* is not in $(s_1,\ldots,s_I)$, choose a value for $\boldsymbol{\phi}_{s_i^*}$ from the base measure G_0. Compute the following acceptance probability:

$$a(s_i^*,s_i)=\min\left\{1,\frac{F(y_i|\boldsymbol{\phi}_{s_i^*})}{F(y_i|\boldsymbol{\phi}_{s_i})}\right\} \tag{A6}$$

  and set the new value of s_i to s_i^* with this probability; otherwise leave s_i unchanged.
- Once the configuration indicators and the associated clusters are determined, we move on to update $\boldsymbol{\omega}$'s. The full conditional distribution of $\boldsymbol{\omega}_k$

$$P(\boldsymbol{\omega}_k|\cdot)\propto dG_0(\boldsymbol{\omega}_k)\prod_{\{i:s_i=k\}}F(y_i|\boldsymbol{\phi}_i=\boldsymbol{\omega}_k) \tag{A7}$$

  which is not in a standard form; therefore we use Metropolis–Hastings (M–H) algorithm to update $\boldsymbol{\omega}_k$'s.

The steps of a cycle of Gibbs sampler under three different models for parent–offspring familial data are illustrated as follows:

1. BVN: Under this model, draw all the parameters following the usual M–H algorithm.
2. DP:
   - Step 2.1. Draw β and ρ_{cc} following the usual M–H algorithm;
   - Step 2.2. Draw $\mathbf{c}_i$'s $(i=1,\ldots,I)$ following Algorithm 5 in [30] as illustrated above. Note, here $\boldsymbol{\phi}_i=\mathbf{c}_i$, G_0 is a standard BVN and the distribution F is as in (11);

DOI: 10.1002/sim

- Step 2.3. Update α:

  (1) Sample η from $p(\eta|\alpha, K) \propto \eta^{\alpha}(1-\eta)^{I-1}$;
  (2) Sample α from $\pi_{\eta}\text{Gamma}(a_{\alpha}+K, b_{\alpha}-\log(\eta))+(1-\pi_{\eta})\text{Gamma}(a_{\alpha}+K-1, b_{\alpha}-\log(\eta))$, where $\pi_{\eta}/(1-\pi_{\eta})=(a_{\alpha}+K-1)/\{I(b_{\alpha}-\log(\eta))\}$;

- Step 2.4. Draw hyperparameters $\boldsymbol{\mu}_b$, σ_c^2, σ_p^2 and ρ_{pc} following the usual M–H algorithm.

3. DPM: Steps 3.1 and 3.3 are the same as steps 2.1 and 2.3, respectively, whereas Step 3.2 is split up in the following three steps:

   - Step 3.2.1. Draw $\mathbf{b}_i$'s $(i=1,\ldots,I)$ following the usual M–H algorithm;
   - Step 3.2.2. Drawing $\boldsymbol{\theta}_i=(\mu_{b1i}, \mu_{b2i}, \sigma_{ci}^2, \sigma_{pi}^2)$ $(i=1,\ldots,I)$ following Algorithm 5 in [30] as illustrated above. Note, here $\boldsymbol{\phi}_i=\boldsymbol{\theta}_i$ and the distribution F is the distribution of $\mathbf{b}_i$, which is a BVN distribution with the parameters $\boldsymbol{\theta}_i$;
   - Step 3.2.3. Drawing hyperparameter ρ_{pc} following the usual M–H algorithm.

ACKNOWLEDGEMENTS

The research of Bhramar Mukherjee was supported by NIH R03 CA130045-01 and NSF DMS 07-06935. The research of Kathleen A. Cooney was supported by SPORE P50 CA69568 and R01 CA79596. The first two authors made equal contributions to the development of this manuscript.

REFERENCES

1. Witte JS, Gauderman J, Thomas DC. Asymptotic bias and efficiency in case–control studies of candidates genes and gene-environment interactions: basic family design. *American Journal of Epidemiology* 1999; **149**:693–705.
2. Umbach DM, Weinberg CR. The use of case-parent triads to study joint effects of genotype and exposure. *American Journal of Human Genetics* 2000; **66**:251–261.
3. Zhao H. Family-based association studies. *Statistical Methods in Medical Research* 2000; **9**:563–587.
4. Siegmund KD, Langholz B. Ascertainment bias in family-based case–control studies. *American Journal of Epidemiology* 2002; **155**(9):875–880.
5. Siegmund KD, Langholz B, Kraft P, Thomas DC. Testing linkage disequilibrium in sibships. *American Journal of Human Genetics* 2000; **67**:244–248.
6. Kraft P, Thomas DC. Bias and efficiency in family-based gene-characterization studies: conditional, prospective, and retrospective likelihoods. *American Journal of Human Genetics* 2000; **66**:1119–1131.
7. Hancock DB, Martin ER, Li Y, Scott WK. Methods for interaction analyses using family-based case–control data: conditional logistic regression versus generalized estimating equations. *Genetic Epidemiology* 2007; **31**:883–893.
8. Whittemore AS. Logistic regression of family data from case–control studies. *Biometrika* 1995; **82**:57–67 (correction 1997; **84**:989–990).
9. Zhao LP, Hsu L, Holte S, Chen Y, Quiaoit F, Prentice RL. Combined association and aggregation analysis of data from case–control family studies. *Biometrika* 1998; **85**:299–315.
10. Neuhaus J, Scott AJ, Wild CJ. The analysis of retrospective family studies. *Biometrika* 2002; **89**:23–37.
11. Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review* 1991; **59**:25–35.
12. Diggle P, Heagert P, Liang K, Zeger SL. *Analysis of Longitudinal Data*. Oxford University Press: New York, 1994.
13. Pfeiffer R, Gail MH, Pee D. Inference for covariates that accounts for ascertainment and random genetic effects in family studies. *Biometrika* 2001; **88**:933–948.
14. Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* 1918; **52**:399–433.
15. Neuhaus J, Scott AJ, Wild CJ. Family-specific approaches to the analysis of case–control family data. *Biometrics* 2006; **62**:488–494.

DOI: 10.1002/sim

16. Ferguson TS. A Bayesian analysis of some nonparametric problems. *Annals of Statistics* 1973; **1**:209–230.
17. O'Brien SM, Dunson DB. Bayesian multivariate logistic regression. *Biometrics* 2004; **60**:739–746.
18. Kinney SK, Dunson DB. Fixed and random effects selection in linear and logistic models. *Biometrics* 2007; **63**:690–698.
19. Mukhopadhyay S, Gelfand AE. Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association* 1997; **92**:633–639.
20. Kleinman KP, Ibrahim JG. A semiparametric Bayesian approach to the random effects model. *Biometrics* 1998; **54**:921–938.
21. Dorazio RM, Mukherjee B, Zhang L, Ghosh M, Jelks HL, Jordan F. Modeling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. *Biometrics* 2007. DOI: 10.1111/j.1541-0420.2007.00873.x (published online: 03 August 2007).
22. Stiratelli R, Laird NM, Ware JH. Random-effects models for serial observations with binary response. *Biometrics* 1984; **40**:961–971.
23. Anderson DA, Aitkin M. Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society*, *Series B*, *Methodological* 1985; **47**:203–210.
24. Breslow NE, Day NE, Halvorsen KT, Prentice RL, Sabai C. Estimation of multiple relative risk functions in matched case–control studies. *American Journal of Epidemiology* 1978; **108**:299–307.
25. Godambe VP. Conditional likelihood and unconditional optimum estimating equations. *Biometrika* 1976; **63**: 277–284.
26. Rice KM. Equivalence between conditional and mixture approaches to the Rasch model and matched case–control studies, with applications. *Journal of the American Statistical Association* 2004; **99**:510–522.
27. Sethuraman J. A constructive definition of Dirichlet priors. *Statistica Sinica* 1994; **4**:639–650.
28. Escobar MD, West M. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 1995; **90**:577–588.
29. Antoniak CE. Mixtures of Dirichlet processes with applications to non-parametric problems. *The Annals of Statistics* 1974; **2**:1152–1174.
30. Neal RM. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* 2000; **9**:249–265.
31. Srivastava M. Estimation of interclass correlations in familial data. *Biometrika* 1984; **71**:177–185.
32. Srivastava M, Keen KJ, Katapa RS. Estimation of interclass and intraclass correlations in multivariate familial data. *Biometrics* 1988; **44**:141–150.

DOI: 10.1002/sim