Estimating Fetal and Maternal Genetic Contributions to Premature Birth From Multiparous Pregnancy Histories of Twins Using MCMC and Maximum-Likelihood Approaches

Timothy P. York,[1,2] Jerome F. Strauss III,[3] Michael C. Neale,[1,2,4] and Lindon J. Eaves[1,2,4]

[1] Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University School of Medicine, Richmond, Virginia, United States of America

[2] Department of Human and Molecular Genetics, Virginia Commonwealth University School of Medicine, Richmond, Virginia, United States of America

[3] Department of Obstetrics and Gynecology, Virginia Commonwealth University School of Medicine, Richmond, Virginia, United States of America

[4] Department of Psychiatry, Virginia Commonwealth University School of Medicine, Richmond, Virginia, United States of America

The analysis of genetic and environmental contributions to preterm birth is not straightforward in family studies, as etiology could involve both maternal and fetal genes. Markov Chain Monte Carlo (MCMC) methods are presented as a flexible approach for defining user-specified covariance structures to handle multiple random effects and hierarchical dependencies inherent in children of twin (COT) studies of pregnancy outcomes. The proposed method is easily modified to allow for the study of gestational age as a continuous trait and as a binary outcome reflecting the presence or absence of preterm birth. Estimation of fetal and maternal genetic factors and the effect of the environment are demonstrated using MCMC methods implemented in WinBUGS and maximum likelihood methods in a Virginia COT sample comprising 7,061 births. In summary, although the contribution of maternal and fetal genetic factors was supported using both outcomes, additional births and/or extended relationships are required to precisely estimate both genetic effects simultaneously. We anticipate the flexibility of MCMC methods to handle increasingly complex models to be of particular relevance for the study of birth outcomes.

Keywords: preterm birth, fetal, maternal, genetic, environment, MCMC, ML

The prevalence of spontaneous preterm birth has risen progressively in the United States over the past decade (Behrman & Butler, 2007). Preterm birth is a multifactorial disease and genetic factors and/or gene–environment interactions are believed to play a significant role in its etiology (Anum, Springel, et al., 2009; Himes & Simhan, 2008; Menon, 2008). Several genes that have been proposed to contribute to prematurity and disparities among populations need to be validated as causally linked to preterm birth, while others remain to be identified (Anum, Springel, et al., 2009). A better understanding of the genetic epidemiology of this significant pregnancy complication will guide gene-finding efforts and enable more precise risk predictions.

The application of existing methodology to estimate the genetic contribution of this complex disease is not straightforward since etiology could involve contributions of both maternal and fetal genes in addition to environmental factors and their interactions. A number of family-based approaches take advantage of individuals that share differing degrees of maternal and paternal derived genes to distinguish these genetic sources (L. J. Eaves, et al., 2005; Maes et al., 1997; Nance & Corey, 1976; Pawitan et al., 2004; Silberg & Eaves, 2004). In the children of twins (COT) design, the offspring of monozygotic twins, in common with other biological half-siblings, share one-fourth of the additive genetic component of variance, while the offspring of dizygotic twins, like other first cousins, share only one-eighth the additive genetic component. If maternal genetic factors are in operation, they will inflate the correlations of maternal half-siblings and cousins related through sisters, relative to those of paternal half-siblings and cousins related through

Timothy P. York, Jerome F. Strauss III, Michael C. Neale, and Lindon J. Eaves

brothers and unlike-sex siblings. The latter groups related through male ancestors share only the effects of the fetal genotype (Haley et al., 1981).

There have been few attempts to estimate either maternal or fetal genetic sources to explain differences in gestational age at delivery from COT studies (Clausson et al., 2000; Kistka et al., 2008; Treloar et al., 2000). Heritability estimates for maternal genetic influences range from 17% to 36%, while no COT study to date has demonstrated a contribution of fetal genes. In these studies, for reasons of data availability or convenience of statistical analysis, analysis is based only on the first birth or births are pooled into a single measure. It has been demonstrated in some cases that including additional family members increases the power to detect the additive effect of genes and shared environmental factors (Posthuma & Boomsma, 2000). Although data on repeated births of the same mother present additional analytic challenges, the inclusion of these observations will increase the information content of each cousinship. For example, the covariance of siblings is composed of both maternal and fetal genetic sources, which will allow for a more precise estimate of these variance components. The nonindependent nature of repeated measures, in this case multiple pregnancies within the same mother and parents (twins) within the same family, are routinely treated using mixed effect models (Galway, 2006). This approach accommodates hierarchical structure, such as family data that is by nature nested, and allows for parameters to vary at more than one level by the modeling of random effects. In contrast to fixed effects, where an unknown constant is estimated from the data, the parameters that constitute the distribution of the random effect are estimated. In the context of quantitative genetic analysis these random effect parameters are indicators of genetic and environmental variance components. Although mixed effects models have been widely used in other applications, their application to the analysis of twin data is relatively recent with the increasing availability of computing power and software to implement these methods (Feng et al., 2009; Rabe-Hesketh et al., 2008; Visscher et al., 2004).

There is an extensive literature on fitting structural models to continuous (multivariate normal) data from kinships (L. J. Eaves et al., 1978; Fulker, 1973; Martin & Eaves, 1977). Such models are easily fitted to multivariate normal data from kinships using any of a number of programs for structural modeling such as Mx (Neale et al., 1999; Neale & Cardon, 1992), M-Plus (Muthen & Muthen, 1998), Mendel (Lange et al., 1988) and LISREL (Joreskog & Sorbom, 1996). However, estimation by numerical optimization of the likelihood is far less tractable for non-normal outcomes, such as categorical data, symptom counts and survival data because computation of the likelihood itself requires integration over all values for the (multivariate) distribution of liabil-ity for each family. Although this task is feasible for hierarchical models when the number of dimensions ('individuals') is relatively low, it becomes very labor-intensive for larger kinships, where the number of random effects to estimate increases with pedigree complexity, and number of outcomes to the point where computation of confidence intervals is virtually precluded by the need for repeated optimization of the same likelihood function for small fixed changes in each of the model parameters.

A series of applications have demonstrated the Markov Chain Monte Carlo (MCMC) approach implemented in the freely available WinBUGS program (D. Spiegelhalter et al., 2003) allows for the efficient calculation of parameter estimates and model comparison statistics for a wide range of genetic problems that have proved virtually intractable on far more powerful main-frame computers by conventional maximum-likelihood (ML; L. Eaves & Erkanli, 2003; L. Eaves, et al., 2005; L. Eaves, et al., 2003). See the reference Gilks et al. (Gilks et al., 1996) for a detailed account of MCMC methods in practice and references Eaves & Erkanli (2003) and Eaves et al. (2005) for a summary of the key features of this method as applied to genetic problems. MCMC has been widely used in other genetic applications including linkage analysis (Brock et al., 2007; George et al., 2005; Sobel & Lange, 1996) and model selection in following up genome wide association studies (Lunn et al., 2006). The key to MCMC is the construction of a series of successive samples (i.e., the 'Markov Chain') whose distribution over repeated samples converges to the target, posterior distribution of parameters conditional upon the data (i.e., the 'stationary distribution'). Thus, not only do the means of the distribution yield the estimates of the parameters, but also the repeated samples may be used to estimate sampling errors, confidence intervals and other parameters of the sampling distribution of the estimates that lie beyond the practical limitations of ML algorithms. A variety of algorithms have evolved to implement MCMC, including the Gibbs sampler (Gilks et al., 1996).

In this study, we demonstrate the application of MCMC methods to estimate the genetic and environmental parameters of interest in pregnancy history data and address the inherent modeling difficulties that are a consequence of the nested twin family structure and occurrences of multiple births of the same mother. Using birth records from offspring of a sample of Virginia twins we illustrate how to estimate fetal and maternal genetic sources and the contribution of the familial environment to differences in the timing of birth. The outcome measure is treated as both a continuous measure (gestational age) and a binary trait (preterm birth liability) to examine differences in parameter estimates that may occur when thresholds are applied to normally distributed data. Identical variance component models are fit using ML methods for comparison.

MCMC Method to Estimate Fetal and Maternal Genetic Effects

Genetic Model

Inspection of the algebraic expectations in Table 1 for the six essential correlations shows that they contain the information necessary to estimate the contribution of fetal (V_F) genetic, maternal (V_M) genetic and shared familial (V_C) environmental effects in COT studies. For example, the difference between the correlation in pregnancy outcome for female MZ twins and that for the spouses of male MZ twins provides an estimate of the contribution of maternal genotype. Similarly, under the simple model there are several contrasts that provide information about the effect of the fetal genotype, for example between the correlation of spouses of male MZ twins and the correlation between pregnancies of the spouses of male DZ twins. The shared environmental component is estimated from the sibling correlation and reflects the idiosyncratic environmental influences of the nuclear family shared by all offspring.

Six assumptions are made initially that can be relaxed in extensions of the proposed model: (1) with the exception of effects of the maternal genotype and the residual shared environmental, all other environmental effects are pregnancy-specific and uncorrelated between successive pregnancies of the same mother; (2) separate genes contribute to the maternal and fetal genetic components of pregnancy outcome; (3) the influence of fetal and maternal genetic differences are the same for male and female fetuses (i.e., genetic effects are autosomal and not X-linked or sex-limited); (4) genetic effects are additive; (5) mating is random; (6) apart from measured covariates (e.g., SES), other random aspects of parental phenotype do not affect pregnancy outcome (i.e., no 'vertical cultural inheritance'). In the absence of multigenerational pregnancy histories (e.g., the pregnancy histories of grandmothers), it is still possible to test for joint failure of assumptions 1, 4, 5 and 6 because each will lead to an apparent excess of the nonfetal contribution to differences ('shared sibling environment') and inflate the correlation in the outcomes of successive pregnancies from the same mother relative to those of twin mothers. Sex-dependent gene expression will lead to lower correlation between the outcomes of unlike-sex than like-sex fetuses. In theory, sex-limited effects of fetal and maternal genotype may also be resolved by analyzing the profile of resemblance between siblings and cousins as a function of whether the pregnancies involve like- or unlike-sex fetuses.

The latent genetic and environmental variables to be estimated contribute to the differences in pregnancy outcomes as defined by the structural model. The model follows a hierarchical form since multiple births serve as repeated measures within the same mother and the estimation of fetal and maternal contributions are constrained to follow expectations by twin type as presented in Table 1. Appendix A shows the implementation of this model as a BUGS script and the logic of the code reflects the process by which genetic and environmental effects are distributed in families (c.f., Eaves & Erkanli, 2003). The notes enumerate the principal steps in the process and the code implements the algebra of the underlying model and supplies details of the matrix algebra and array handling. The model for the kth pregnancy of the jth twin of the ith pair can be written as a variance components model,

$$y_{ijk} = \mu + F_{ijk} + M_{ijk} + C_{ijk} + E_{ijk} \qquad (1)$$

where the overall mean is μ, $F_{ijk} \sim N(0,\sigma_F^2)$ is the fetal genetic component, $M_{ijk} \sim N(0,\sigma_M^2)$ is the maternal genetic component, $C_{ijk} \sim N(0,\sigma_C^2)$ is the shared environmental component and $E_{ijk} \sim N(0,\sigma_E^2)$ is the unique environmental component. Generally, for each nuclear family the number of pregnancies, k, is not restricted. Since the variance components are assumed to be independent their sum equals the total variance, var (y_{ijk}). Covariances between twin offspring within a twin or within a cousinship are specified as a function of twin type (Table 1). Different covariance matrices of random effects can be easily specified in WinBUGS with the use of separate looping structures for each twin type implemented during the sampling stage of the algorithm (see Appendix A). Because of the widespread use of the multifactorial threshold model in kinship studies, we also implemented this model in our application by expressing the probability of a binary outcome as a probit function of the normal liability of each pregnancy given the random effects of maternal and fetal genotype and shared sibling environment.

Parameter Estimation and Model Comparison

Application of MCMC to genetic problems has been facilitated enormously by the development of freely distributed software package by staff of the MRC BUGS project in Cambridge, England (http://www.mrc-bsu.cam.ac.uk/bugs/). WinBUGS is the PC implementation (Spielgelhalter et al., 2003) that employs a flexible and transparent 'R-like' code (see example code in Appendix A). All MCMC models in WinBUGS are first subjected to a 'burn-in' period to mitigate the influence of initial values. Numerical estimates of para-

Table 1

Expected Covariance Between Pregnancy Outcomes as a Function of Relationship Between Offspring

Parental Relationship	Fetal relationship	Expected covariance
MZ female twins	Cousin ('Half-sibling')	¼ V_F + V_M
DZ female twins	Cousin	⅛ V_F + ½ V_M
MZ male twins	Cousin ('Half-sibling')	¼ V_F
DZ male twins	Cousin	⅛ V_F
DZ male-female twins	Cousin	⅛ V_F
Sibship	Sibling	½ V_F + V_M + V_C

Note: V_F = fetal genetic, V_M = maternal genetic, V_C = shared familial environment

Table 2

MCMC and ML Model Comparison Summary Statistics

Model	Trait distribution	k	MCMC DIC	MCMC Δ DIC[a]	ML $-2LL$	ML p value[b]	ML AIC
V_F, V_M, V_C	Continuous	5	24027.9	—	24436.120	—	10324.120
V_F, V_M	Continuous	4	24013.1	−14.8	24436.437	0.574	10322.437
V_M	Continuous	3	24006.4	−21.5	24437.300	0.554	10321.300
V_F	Continuous	3	24023.3	−4.6	24450.932	0.001	10334.932
V_F, V_M, V_C	Binary	5	2178.6	—	2766.777	—	−11353.223
V_F, V_M	Binary	4	2178.9	0.3	2766.769	1.00	−11355.231
V_M	Binary	3	2594.8	416.2	2768.255	0.478	−11355.745
V_F	Binary	3	1595.8	−582.8	2767.463	0.710	−11356.537

Note: [a]Δ DIC = change in deviance information criteria or [b]chi-square test of −2 log-likelihood difference from submodel versus full model containing V_F, V_M and V_C.

meters are obtained by subsequent sampling from the (presumed) stationary distribution. Successive samples of parameter values from the joint posterior distribution over all iterations are monitored from which summary statistics, such as means, standard errors and confidence intervals, are calculated. In contrast to the amount of computation required to generate second derivatives and plot contours of likelihood surfaces, the ease with which the sampling properties of the estimates are available to MCMC is a significant by-product. A related benefit is the ease with which the sampling distributions of functions of the model parameters (e.g., heritability estimates, proportions of variance) can be derived.

The familiar likelihood-ratio tests of ML are not readily available in MCMC because the approach also estimates a very large number of random effects as well as the fixed effects and other model parameters. The deviance information criterion (DIC), which uses an empirical estimate of the number of model parameters (pD), is used to penalize improvements in likelihood for changes in model complexity. Spiegelhalter et al. propose using the DIC for model comparison in a Bayesian framework (Spiegelhalter et al., 2002). The DIC is analogous to the widely used Akaike Information Criteria (Akaike, 1987) that seeks to optimize the balance between goodness of fit and model parsimony. The model with the smallest DIC value is estimated to best predict a replicate dataset while being mindful of parsimony. A 'rule of thumb' suggests that differences of 10 in the DIC of successive models indicate improvements in predictive value that need to be taken seriously and changes of 5 are at least suggestive. It should be cautioned the use of DIC is not recommended when the sampled posterior distributions are skewed, bimodal or truncated (Ntzoufras, 2009; Spiegelhalter et al., 2003), which from our experience is in itself an indicator of poor model performance resulting from over-fitting or lack of sufficient data to estimate parameters with adequate precision.

Application of MCMC Parameter Estimation to Pregnancy Histories of Virginia Twins

Sample

Pregnancy histories of twin parents were obtained by merging birth records from the Virginia Department of Health Office of Vital Records (VDH) with registered participants in the Mid-Atlantic Twin Registry (MATR) at Virginia Commonwealth University School of Medicine. The details of the MATR and sample characteristics are described elsewhere (Anderson et al., 2002). Matches were performed at the VDH by merging the parent SSN on the birth record for offspring born after 1989 with the SSN of twins provided by the MATR. Virginia Commonwealth University IRB approved sample collection and study design (VCU IRB# HM11443). Offspring exclusion criteria, based on data available from birth certificates, include multiple birth, any congenital anomalies, hydramnios/oligohydramnios, pregnancies complicated by pregnancy induced hypertension and eclampsia, Rh sensitization, abruptio placenta and placenta previa, or any medically necessitated preterm delivery. To avoid the influence of extreme values that may be associated with different etiologic factors we omitted gestational ages that deviated beyond 2.5 standard deviations from the mean. A binary preterm outcome measure was indicated as gestational age less than 37 completed weeks. The size of sibships, k, was limited to the first four births (99.5% of all births) in order to balance gains with computational load.

Results

The data-merge between the VDH and MATR resulted in 7,061 total births from 810 (264) MZ female, 529 (131) DZ female, 680 (166) MZ male, 465 (90) DZ male and 1201 (208) unlike-sex pairs. Numbers in parenthesis indicate the number of pairs where at least one birth was reported for both twins. The prevalence of preterm birth was 5.7%, average gestational age was 39.0 (SD = 1.4), mothers' age at birth of first child 27.6 (SD = 5.5), fathers' age at birth of first child 30.6 (SD = 6.0) and self-reported

Table 3

Parameter Estimates and 95% Confidence Intervals of Best Fitting Models for MCMC and ML Methods

Trait distribution	Method	Fetal genetic	Maternal genetic	Unique environment
Continuous	MCMC	—	0.526 (0.452, 0.600)	1.412 (1.344, 1.484)
Continuous	ML	—	0.525 (0.453, 0.600)	1.411 (1.342, 1.484)
Binary[a]	MCMC	87.20 (65.57, 94.03)	—	12.08 (5.97, 34.43)
Binary	ML	86.40 (62.00, 99.9)	—	13.60 (0.001, 38.00)

Note: [a]The sum of parameter estimates for the binary case are constrained to equal one.

race for first child was 84.7% European American and 14.2% African–American.

Estimates of fetal (V_F) and maternal (V_M) genetic effects and the contribution of the shared environment (V_C) were obtained using the MCMC approach implemented in WinBUGS. A series of models were fitted following a program of a 10,000 iteration burn-in period followed by 50,000 subsequent samplings (see Table 2). ML estimates using the structural equation modeling program Mx (Neale et al., 1999) were also obtained for comparison. Nested models were compared to the full model that included both genetic sources and the effect of the shared environment. Results for the continuous outcome indicated a lack of support for the contribution of the shared environment. This is evident from the zero lower bound of this parameter yielding a highly skewed posterior distribution of Vc (not shown). Removal of this parameter is also indicated by the lack of significant difference in model fit reflected in the likelihood-ratio test of the ML method (p value = 0.574). Of the remaining models containing only genetic effects, the model including only maternal genetic effects performed best by DIC and AIC. Estimates of the variances due to the random effects show close agreement between the Bayesian and ML approaches. The results for the binary outcome differed in that the best fitting model contained only the effects of fetal genes. The parameter estimates and 95% confidence intervals for the best fitting models for both outcomes are listed in Table 3, but we note these models were preferred over the model including both the fetal and maternal genetic contributions based on very small changes in the AIC. The greater ambiguity of the results for the binary outcome reflects the marked loss in the amount of information associated with dichotomizing a continuous outcome. Even these relatively large samples do not permit the effects of maternal and fetal effects to be estimated with adequate precision for them to be resolved unequivocally. Model summary statistics and paramter estimates were similar when fitting models that included only the predominant racial classification.

Discussion

This study introduces a flexible MCMC approach for the estimation of fetal and maternal genetic effects

using pregnancy histories from the offspring of female and male twin pairs. The models handle data on multiple offspring from the same pregnancy to increase the information from each cousinship and enable more precise estimation of fetal and maternal genetic effects and the shared environment effect. The structural model was fitted efficiently using widely distributed Mx software for linear structural models (Neale et al., 1999) and similar parameter estimates and confidence intervals were obtained. The flexibility of MCMC for our application, however, is the relative ease with which the same basic structural model can be generalized to include other kinds of outcomes such as binary traits. Alternative mathematical models for the distribution of the outcome variables are easily implemented by changing the link function relating liability to outcome. Because of the widespread use of the multifactorial threshold model in kinship studies we implemented this model in our application by expressing the probability of a binary outcome as a probit function of the normal liability of each pregnancy given the random effects of maternal and fetal genotype and shared sibling environment. Other link and distribution functions may be used to explore other definitions of premature birth, including time-to-event (length of gestation), multiple categories (e.g., grades of 'severity' of outcome) or counts of multiple symptoms or events.

The application of this approach to a set of Virginia COT data suggested that the contribution of both maternal and fetal genetic effects to gestational age and did not support the contribution of shared environmental factors. Model results differed by outcome with 27.1% of the continuous trait variance accounted for by maternal genetic factors while fetal genetic factors explained 87.2% of liability to risk of preterm birth when applied to the binary outcome. As noted previously, the selection of the optimal model for both outcome measures over the maternal/fetal combined genetic model was based on small differences in model summary statistics. In addition, the confidence intervals around the fetal genetic variance component were wide and approached the upper limit of the parameter bounds, which suggests that it was difficult for either method to estimate this parameter using the current sample. Increased variability in parameters of the random effects model is seen for the

binary outcome because substantial information is lost in dichotomizing a truly continuous phenotype.

The lack of strong support for a fetal genetic effect in this study and absence in others (Kistka et al., 2008) is curious since birth timing has been described as a culmination of series of physiologic and anatomic changes in both mother and fetus (Chaudhari et al., 2008; Lye et al., 2007). Indeed, genetic disorders affecting the fetus including certain forms of Ehlers-Danlos syndrome are known to predispose to early birth resulting from preterm premature rupture of the fetal membranes (PPROM), whereas Ehlers-Danlos syndrome in the mother with an unaffected fetus does not predispose to preterm birth (Anum, Hill et al., 2009). Association studies have also indicated that certain genetic variants carried by the fetus increase risk of PPROM (Anum, Springel et al., 2009). It is possible that fetal contributions to preterm birth are population-specific. PPROM is a more common cause of preterm birth in African–Americans than non-Hispanic whites (Plunkett et al., 2008; Shen et al., 2008). Consequently, the smaller proportion of African–Americans in the Virginia COT sample may have obscured the fetal contribution.

In the case of COT designs, the most direct methods of increasing power to detect genetic effects are to increase the number of twin pairs or to include pregnancy history on additional family members, for instance, by the inclusion of avuncular relationships and/or the offspring of sibling pairs. In addition, the models presented in this study provide a framework to include multiple births per mother, which, as opposed to pooling offspring measures, facilitates the inclusion of covariates in genetic models that could help to uncover genetic effects. For instance, when available, covariate measures can be used to test the effect of gene–environment interaction by examining the modulation of the contribution of genetic variation by the level of salient environmental attributes.

In MCMC, random effects ('liabilities') are assigned values conditional upon the data that are as much parameters of the model as the components of variance and regression coefficients. If covariates can be specified, the random effects for missing subjects are estimated conditional on the data available from other members of the family and covariate values, on the assumption that missing values are sampled from the same distribution as valid observations. Although this is typically regarded as a convenient way of handling missing data, it also has implications for counseling and follow-up since risks to future pregnancies may be estimated, together with their confidence intervals, by including such putative pregnancies as additional 'missing' observations. Furthermore, the estimates of the individual components of the random effects (e.g., the effect of the maternal and/or fetal genotypes) provide an informed basis for sampling (or excluding) families for subsequent follow-up to identify those that are likely to be most informative for genotyping

with respect to the candidate genes that are hypothesized to affect either the maternal or fetal components of liability.

Although MCMC methods have been applied to the analysis of pedigree data (Brock et al., 2007), their application to data on twins and their relatives is relatively recent. Benchmark studies (L. Eaves & Erkanli, 2003) show that, for problems that are tractable with Mx, such as fitting a simple structural model to multivariate normal data, ML and MCMC yield comparable answers, as also observed in this application. In such cases the ML algorithm in Mx is far faster and easier to use and remains the algorithm of choice. However, the advantage quickly reverses for a wide range of models that cannot currently be implemented in ML packages because of the excessive demands of repeated numerical integration. A wide range of applications to real and simulated twin data are now published that illustrate the flexibility of MCMC. These include: genetic survival analysis of the timing of menarche in twins (Do et al., 2000); nonlinear latent growth curve models (L. Eaves & Erkanli, 2003); analysis of genetic and environment components of the relative timing of pubertal change in multiple indicators of puberty; 'genetic' IRT models for large numbers of multi-category items (L. Eaves et al., 2005; van den Berg et al., 2006); the interaction and correlation of genetic liability and exposure to life events in the etiology of adolescent depression (L. Eaves et al., 2003); and the genetic control of developmental change in multivariate indictors of childhood fears (L. J. Eaves & Silberg, 2008). The above examples illustrate the flexibility of MCMC to tackle a range of hitherto intractable problems in the analysis of twin data.

The structural model for this application is relatively simple, assuming that genetic effects do not depend on sex, that mating is random and that there are no correlations between successive pregnancies apart from the effects of the maternal genotype and the correlation between offspring for the effects of the fetal genotype. Each of these assumptions may be modified by altering the structural model appropriately. Heath et al. (1985), Truett et al. (1994), Maes et al. (1997), and Eaves et al. (L. Eaves, et al., 1999; L. J. Eaves, et al., 1986), provide more general models for the correlations between relatives in the kinships of twins in which it is possible to allow for effects such as assortative mating, sex-limited gene expression, carry-over effects from previous pregnancies, developmental change, nonadditive genetic effects and other shared nongenetic influences. Although MCMC is astonishingly flexible, the approach is not a panacea and may share many of the limitations that characterize numerical algorithms that maximize the likelihood. Typically, these problems result from multicollinearity, under-identification, or less than optimal model parameterization that results in slow convergence ('mixing') of the MCMC algorithm.

In conclusion, we show how MCMC methods can be applied to the full pregnancy history from COT samples. Although further study is warranted with larger samples, we anticipate the flexibility of these methods to handle increasingly complex models to be of particular relevance for the study of birth outcomes.

Acknowledgments

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317–332.

Anderson, L. S., Beverly, W. T., Corey, L. A., & Murrelle, L. (2002). The Mid-Atlantic Twin Registry. *Twin Research, 5*, 449–455.

Anum, E. A., Hill, L. D., Pandya, A., & Strauss, J. F., 3rd (2009). Connective tissue and related disorders and preterm birth: Clues to genes contributing to prematurity. *Placenta, 30*, 207–215.

Anum, E. A., Springel, E. H., Shriver, M. D., & Strauss, J. F., 3rd (2009). Genetic contributions to disparities in preterm birth. *Pediatric Research, 65*, 1–9.

Behrman, R. E., & Butler, A. S. (2007). *Preterm birth: Causes, consequences, and prevention*. Washington, D. C.: Academy Press.

Brock, G. N., Weeks, D. E., Sobel, E., & Feingold, E. (2007). A hierarchical model for estimating significance levels of non-parametric linkage statistics for large pedigrees. *Genetic Epidemiology, 31*, 417–430.

Chaudhari, B. P., Plunkett, J., Ratajczak, C. K., Shen, T. T., DeFranco, E. A., & Muglia, L. J. (2008). The genetics of birth timing: Insights into a fundamental component of human development. *Clinical Genetics, 74*, 493–501.

Clausson, B., Lichtenstein, P., & Cnattingius, S. (2000). Genetic influence on birthweight and gestational length determined by studies in offspring of twins. *BJOG: An International Journal of Obstetrics and Gynaecology, 107*, 375–381.

Do, K. A., Broom, B. M., Kuhnert, P., Duffy, D. L., Todorov, A. A., Treloar, S. A., & Martin, N. G. (2000). Genetic analysis of the age at menopause by using estimating equations and Bayesian random effects models. *Statistics in Medicine, 19*, 1217–1235.

Eaves, L., & Erkanli, A. (2003). Markov Chain Monte Carlo approaches to analysis of genetic and environmental components of human developmental change and G × E interaction. *Behavior Genetics, 33*, 279–299.

Eaves, L., Erkanli, A., Silberg, J., Angold, A., Maes, H. H., & Foley, D. (2005). Application of Bayesian inference using Gibbs sampling to item-response theory modeling of multi-symptom genetic data. *Behavior Genetics, 35*, 765–780.

Eaves, L., Heath, A., Martin, N., Maes, H., Neale, M., Kendler, K., Kirk, K., & Corey, L. (1999). Comparing the biological and cultural inheritance of personality and social attitudes in the Virginia 30,000 study of twins and their relatives. *Twin Research, 2*, 62–80.

Eaves, L., Silberg, J., & Erkanli, A. (2003). Resolving multiple epigenetic pathways to adolescent depression. *Journal of Child Psychology and Psychiatry, 44*, 1006–1014.

Eaves, L. J., Last, K. A., Young, P. A., & Martin, N. G. (1978). Model-fitting approaches to the analysis of human behaviour. *Heredity, 41*, 249–320.

Eaves, L. J., Long, J., & Heath, A. C. (1986). A theory of developmental change in quantitative phenotypes applied to cognitive development. *Behavior Genetics, 16*, 143–162.

Eaves, L. J., & Silberg, J. L. (2008). Developmental-genetic effects on level and change in childhood fears of twins during adolescence. *Journal of Child Psychology and Psychiatry, 49*, 1201–1210.

Eaves, L. J., Silberg, J. L., & Maes, H. H. (2005). Revisiting the children of twins: Can they be used to resolve the environmental effects of dyadic parental treatment on child behavior? *Twin Research and Human Genetics, 8*, 283–290.

Feng, R., Zhou, G., Zhang, M., & Zhang, H. (2009). Analysis of Twin Data Using SAS. *Biometrics, 65*, 584–589.

Fulker, D. W. (1973). A biometrical genetic approach to intelligence and schizophrenia. *Social Biology, 20*, 266–275.

Galway, N. W. (2006). *Introduction to mixed modelling*. Chichester: John Wiley & Sons, Ltd.

George, A. W., Wijsman, E. M., & Thompson, E. A. (2005). MCMC multilocus lod scores: Application of a new approach. *Human Heredity, 59*, 98–108.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Boca Raton: Chapman and Hall/CRC.

Haley, C. S., Jinks, J. L., & Last, K. (1981). The monozygotic twin half-sib method for analysing maternal effects and sex-linkage in humans. *Heredity, 46*, 227–238.

Heath, A. C., Kendler, K. S., Eaves, L. J., & Markell, D. (1985). The resolution of cultural and biological inheritance: Informativeness of different relationships. *Behavior Genetics, 15*, 439–465.

Himes, K. P., & Simhan, H. N. (2008). Genetic susceptibility to infection-mediated preterm birth. *Infectious Disease Clinics of North America, 22*, 741–753, vii.

Joreskog, K. G., & Sorbom, D. (1996). *LISREL 8: User's reference guide*. Chicago: Scientific Software International.

Kistka, Z. A., DeFranco, E. A., Ligthart, L., Willemsen, G., Plunkett, J., Muglia, L. J., & Boomsma, D. I.

(2008). Heritability of parturition timing: An extended twin design analysis. *American Journal of Obstetrics and Gynecology, 199*, 43 e41–45.

Lange, K., Weeks, D., & Boehnke, M. (1988). Programs for pedigree analysis: MENDEL, FISHER, and dGENE. *Genetic Epidemiology, 5*, 471–472.

Lunn, D. J., Whittaker, J. C., & Best, N. (2006). A Bayesian toolkit for genetic association studies. *Genetic Epidemiology, 30*, 231–247.

Lye, S. J., Tsui, P., Dong, X., Mitchell, J., Dorogin, A., MacPhee, D., Oldenhoff, A., Langille, B. L., Challis, J. R. G., & Shynlova, O. (2007). *Myometrial programming: A new concept underlying the regulation of myometrial function during pregnancy*. Oxon: Informa UK Ltd.

Maes, H. H., Neale, M. C., & Eaves, L. J. (1997). Genetic and environmental factors in relative body weight and human adiposity. *Behavior Genetics, 27*, 325–351.

Martin, N. G., & Eaves, L. J. (1977). The genetical analysis of covariance structure. *Heredity, 38*, 79–95.

Menon, R. (2008). Spontaneous preterm birth, a clinical dilemma: Etiologic, pathophysiologic and genetic heterogeneities and racial disparity. *Acta Obstetricia et Gynecologica Scandinavia, 87*, 590–600.

Muthen, L. K., & Muthen, B. O. (1998). *Mplus: User's guide*. Los Angeles, CA: Muthen and Muthen.

Nance, W. E., & Corey, L. A. (1976). Genetic models for the analysis of data from the families of identical twins. *Genetics, 83*, 811–826.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. M. (1999). *Mx: Statistical modeling*. Richmond: Department of Psychiatry, Virginia Commonwealth University.

Neale, M. C., & Cardon, L. R. (1992). *Methodology for Genetic Studies of Twins and Families*. Dordrecht: Kluwer Academic Publishers.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Hoboken, New Jersey: John Wiley & Sons.

Pawitan, Y., Reilly, M., Nilsson, E., Cnattingius, S., & Lichtenstein, P. (2004). Estimation of genetic and environmental factors for binary traits using family data. *Statistics in Medicine, 23*, 449–465.

Plunkett, J., Borecki, I., Morgan, T., Stamilio, D., & Muglia, L. J. (2008). Population-based estimate of sibling risk for preterm birth, preterm premature rupture of membranes, placental abruption and pre-eclampsia. *BMC Genetics, 9*, 44.

Posthuma, D., & Boomsma, D. I. (2000). A note on the statistical power in extended twin designs. *Behavior Genetics, 30*, 147–158.

Rabe-Hesketh, S., Skrondal, A., & Gjessing, H. K. (2008). Biometrical modeling of twin and family data using standard mixed model software. *Biometrics, 64*, 280–288.

Shen, T. T., DeFranco, E. A., Stamilio, D. M., Chang, J. J., & Muglia, L. J. (2008). A population-based study of race-specific risk for preterm premature rupture of membranes. *American Journal of Obstetrics and Gynecology, 199*, 373 e1–7.

Silberg, J. L., & Eaves, L. J. (2004). Analysing the contributions of genes and parent–child interaction to childhood behavioural and emotional problems: A model for the children of twins. *Psychological Medicine, 34*, 347–356.

Sobel, E., & Lange, K. (1996). Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics, 58*, 1323–1337.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). WinBUGS User Manual Available from *http://www.mrc-bsu.cam.ac.uk/bugs*

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society: Series B, 64*, 583–640.

Spielgelhalter, D. J., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS version 1.4 User Manual*. Cambridge: MRC Biostatistics Unit.

Treloar, S. A., Macones, G. A., Mitchell, L. E., & Martin, N. G. (2000). Genetic influences on premature parturition in an Australian twin sample. *Twin Research, 3*, 80–82.

Truett, K. R., Eaves, L. J., Walters, E. E., Heath, A. C., Hewitt, J. K., Meyer, J. M., Silberg, J., Neale, M. C., Martin, N. G., & Kendler, K. S. (1994). A model system for analysis of family resemblance in extended kinships of twins. *Behavior Genetics, 24*, 35–49.

van den Berg, S. M., Beem, L., & Boomsma, D. I. (2006). Fitting genetic models using Markov Chain Monte Carlo algorithms with BUGS. *Twin Research and Human Genetics, 9*, 334–342.

Visscher, P. M., Benyamin, B., & White, I. (2004). The use of linear mixed models to estimate variance components from data on twin pairs by maximum likelihood. *Twin Research, 7*, 670–674.

Appendix A

```
#===================================================================#
# Estimation of maternal/fetal genetic, shared/unique environmental factors to birth outcomes.
# Comments are to the right of '#'.
# Outcome is continuous trait (e.g., gestational age); Binary trait can be specified as indicated.
# '.m' is a maternal parameter.
# '.p' is a paternal (fetal in this context) parameter.
#===================================================================#
model;
{
# Specify priors
mu ~ dnorm(39,0.001) #overall mean
s.m ~ dunif(0.0,10) # s.d. of random maternal effects
s.f ~ dunif(0.0,10) # s.d. of random fetal ('paternal') effects; 'cousin effects'
s.ec ~ dunif(0,10) # s.d. of random shared environmental effects
s.e ~ dunif(0,15) # s.d. of random within sibship environmental effects
v.m <- s.m*s.m
v.f <- s.f*s.f
v.ec <- s.ec*s.ec
v.e <- s.e*s.e

# Specify covariances between offspring of twins
c.mzm <- 0.25*v.f #MZ paternal
c.dzm <- 0.125*v.f #DZ paternal
c.mzf <- 0.25*v.f + v.m #MZ maternal
c.dzf <- 0.125*v.f + 0.5*v.m #DZ maternal
c.dzo <- 0.125*v.f #DZ opp sex
v.w <- 0.5*v.f + v.e #Within sibship deviations
csib <- v.m + 0.5*v.f + v.ec #Covariance of siblings
r.mzm <- csib - c.mzm
r.dzm <- csib - c.dzm
r.mzf <- csib - c.mzf
r.dzf <- csib - c.dzf
r.dzo <- csib - c.dzo

# Precision of cousin effects
tau.c.mzm <- 1/c.mzm
tau.c.dzm <- 1/c.dzm
tau.c.mzf <- 1/c.mzf
tau.c.dzf <- 1/c.dzf
tau.c.dzo <- 1/c.dzo

# Precision of between sibship differences within cousinships
tau.r.mzm <- 1/r.mzm
tau.r.dzm <- 1/r.dzm
tau.r.mzf <- 1/r.mzf
tau.r.dzf <- 1/r.dzf
tau.r.dzo <- 1/r.dzo

# Precision of within sibship deviations
tau.w <- 1/v.w

#Simulate twin liabilities: DZ.m
for (pair in 1:NDZ.m) {
mucous[pair] ~ dnorm(mu,tau.c.dzf) # simulate cousin means
for (twin in 1:2) {
mupair[pair,twin] ~ dnorm(mucous[pair],tau.r.dzf) }} #simulate means of sibpair
#Simulate twin liabilities: MZ.m
for (pair in (NDZ.m+1):(NDZ.m+NMZ.m)) {
```

Timothy P. York, Jerome F. Strauss III, Michael C. Neale, and Lindon J. Eaves

```
mucous[pair] ~ dnorm(mu,tau.c.mzf)
for (twin in 1:2) {
mupair[pair,twin] ~ dnorm(mucous[pair],tau.r.mzf) }}

#Simulate twin liabilities: DZ.p
for (pair in (NDZ.m+NMZ.m+1):(NDZ.m+NMZ.m+NDZ.p)) {
mucous[pair] ~ dnorm(mu,tau.c.dzm)
for (twin in 1:2) {
mupair[pair,twin] ~ dnorm(mucous[pair],tau.r.dzm) }}

#Simulate twin liabilities: MZ.p
for (pair in (NDZ.m+NMZ.m+NDZ.p+1):(NDZ.m+NMZ.m+NDZ.p+NMZ.p)) {
mucous[pair] ~ dnorm(mu,tau.c.mzm)
for (twin in 1:2) {
mupair[pair,twin] ~ dnorm(mucous[pair],tau.r.mzm) }}

#Simulate twin liabilities: DZO
for (pair in (NDZ.m+NMZ.m+NDZ.p+NMZ.p+1):(NDZ.m+NMZ.m+NDZ.p+NMZ.p+NDZ.o)) {
mucous[pair] ~ dnorm(mu,tau.c.dzo)
for (twin in 1:2) {
mupair[pair,twin] ~ dnorm(mucous[pair],tau.r.dzo) }}

# Compute likelihood for every offspring
# For the binary case an additional loop is needed to compute subject specific endorsement probabilities.
# Binary model assumes sharp threshold at 'a' such that p[pair,twin,child]=1 if x[pair,twin,child]>a, else p=0.
for (pair in 1:N){
  for (twin in 1:2){
    for (child in 1:NCHILD){
      y[pair, twin, child] ~ dnorm(mupair[pair, twin],tau.w) }}}

} #end of model
```