METHODOLOGY

Covariance components models for longitudinal family data

Paul R Burton,1* Katrina J Scurrah,1,2 Martin D Tobin1 and Lyle J Palmer3

Accepted 1 March 2005

A longitudinal family study is an epidemiological design that involves repeated measurements over time in a sample that includes families. Such studies, that may also include relative pairs and unrelated individuals, allow closer investigation of not only the factors that cause a disease to arise, but also the genetic and environmental determinants that modulate the subsequent progression of that disease. Knowledge of such determinants may pay high dividends in terms of prognostic assessment and in the development of new treatments that may be tailored to the prognostic profile of individual patients. Unfortunately longitudinal family studies are difficult to analyse. They conflate the complex within-family correlation structure of a cross-sectional family study with the correlation over time that is intrinsic to longitudinal repeated measures. Here we describe an approach to analysis that is relatively straightforward to implement, yet is flexible in its application. It represents a natural extension of a Gibbs-sampling-based approach to the analysis of cross-sectional family studies that we have described previously. The approach can be applied to pedigrees of arbitrary complexity. It is applicable to continuous traits, repeated binary disease states, and repeated counts or rates with a Poisson distribution. It not only supports the analysis of observed determinants, including measured genotypes, but also allows decomposition of the correlation structure, thereby permitting conclusions to be drawn about the effect of unobserved genes and environment on key features of disease progression, and hence to estimate the heritability of these features. We demonstrate the efficacy of our methods using a range of simulated data analyses, and illustrate its practical application to longitudinal blood pressure data measured in families from the Framingham Heart Study.

Keywords Longitudinal, family studies, MCMC, Gibbs sampling, Bayesian, genetic epidemiology

1 Biostatistics and Genetic Epidemiology, Department of Health Sciences and Department of Genetics, Institute of Genetics, University of Leicester, UK.
2 Department of Physiology and Centre for Genetic Epidemiology, the University of Melbourne, Australia.
3 Western Australian Institute for Medical Research and UWA Centre for Medical Research, University of Western Australia, Australia.

* Corresponding author. University of Leicester, Department of Health Sciences, 22–28 Princess Road West, Leicester LE1 6TP, UK. E-mail: paul.genepi@ntlworld.com

A conventional longitudinal study involves the repeated evaluation of one or more measurable characteristics (phenotypes or traits) in a series of unrelated individuals. Such studies are used widely. The repeated measurements provide for reduced error, increased statistical power and, critically, a means to study the pattern and determinants of, systematic changes in a phenotype of interest over time.1–3 In contrast, a cross-sectional family study enrolls groups of relatives rather than unrelated individuals and typically involves a single phenotypic assessment in each of the study participants in each family. This permits the study of phenotypic similarities and differences amongst close relatives and allows one to disentangle the genetic and environmental contributions to the aetiology of the disease or trait under study.4–7 Both types of studies involve the analysis of complex correlated data: within-subject phenotypic correlations over time in a longitudinal study, and within-family phenotypic correlations in a cross-sectional family study. Crucially, the correlated data structures fundamental to both types of study permit the investigation of latent, or unobserved, aetiological determinants of a trait as well as those reflected in variables that may be measured directly.

By combining the features of longitudinal studies in individuals and cross-sectional studies in families, longitudinal family studies increase the power to resolve the genetic and environmental determinants of traits associated with complex diseases and
provide a powerful and flexible method to study the corresponding determinants of change in such traits over time.

The scientific power and the analytic challenge of longitudinal family studies derive from the residual correlation that exists between the phenotypes of individuals in the same family, and between repeated measurements in the same individual over time. Sometimes this correlation (or covariance) is of prime interest in its own right. In other situations, it is a nuisance that must be taken into appropriate account in order to obtain valid statistical inferences. In either case, the correlation structure of longitudinal family data conflates all the complexity of cross-sectional family data with that of longitudinal data in individuals. This means that the analysis of such data is difficult and ‘there is a relative paucity of methods’ with which it may be approached.

This dearth of applicable methods is a serious problem, and the importance of the methodological challenge that it implies was underlined by the decision to make longitudinal family data from the Framingham Heart Study the focus of Genetic Analysis Workshop 13 (GAW13) in New Orleans, in November 2002. At that workshop, many different approaches to the analysis of longitudinal family data were investigated. An important conclusion was that: although ‘Well-selected cross-sectional data (e.g., first or last visit) provided good power for detecting some genes.’ summaries of longitudinal data (e.g., means, slopes) were generally most effective for finding genes, particularly those that affected trends in outcome over time. This suggests that future investigations based on longitudinal family data hold considerable promise. It also underscores the priority that must be given to the development of appropriate analytic methods and supporting software. Therefore, contemporary experience with longitudinal family studies is rather limited, and any contribution that methods for the analysis of longitudinal family data may potentially make to scientific progress lies primarily in the future.

Nevertheless, in order to fully realise this potential, it is already clear that a good analytic model for longitudinal familial data should satisfy a number of desirable criteria. First, it should be able to take appropriate account of the correlation structure. Second, it should be able to deal easily with repeated measurements on a trait with any one of a variety of common functional forms. These include binary disease states, continuous traits with a normal distribution, and phenotypes that are reflected in counts with a Poisson distribution. Third, such a model should be applicable to family data in the broadest sense including unrelated individuals, sibships, nuclear families, and complex multigenerational pedigrees. Fourth, the model should be tractable: the necessary mathematics should readily be implemented in appropriate software, and that software should be usable by research groups other than those that originally developed the model.

Keeping these criteria in mind, a number of different approaches to analysis have been proposed, with varying degrees of success and applicability. One of the simplest is to apply some form of data reduction procedure to resolve the repeated measurements in each individual to generate a single summary measure. In a second step, the summary measures may be analysed using a conventional method for cross-sectional family data. As an example, the quantitative change in a phenotype between the first and last measurements can be used as a summary measure in each individual. Alternatively, a regression model can be fitted for each individual and its intercept or slope, or sometimes an appropriate regression residual, may be used as an adjusted phenotype. However, such two-step approaches are often less efficient than a single model that makes use of all the data simultaneously. Furthermore, two-stage procedures often fail to take proper account of uncertainty in the value of the summary measures when fitting the second stage model.

Multilevel models and marginal models that assume a simple hierarchical correlation structure are often used to analyse repeated measures data in individuals, but it is difficult to generalize these approaches to fully utilize the additional information contained in family data. For example, multilevel models can be used for nuclear family data, but the model specification is hard to generalize to multigenerational pedigrees. Furthermore, when the outcome is non-normal, multilevel marginal models fitted using iterative generalized least squares (IGLS) can introduce bias. On the other hand, marginal models fitted using generalized estimating equations (first order GEEs) are useful for normal, binary, and Poisson repeated measures data when fixed effects are of primary interest, but first order GEEs do not produce efficient estimates of the covariance parameters themselves.

The topic has long been of interest to genetic epidemiologists. More than 20 years ago, Lange and colleagues described very general covariance components models for quantitative phenotypes in pedigrees that could be extended to multivariate traits and thus, in principle, to longitudinal data. These could be fitted with user modifiable Fortran 77 code in the FISHER Program. Unfortunately, although the modification to the scoring algorithm that was required to extend the models to non-linear problems was straightforward to formulate, the software implementation was restricted to problems involving multivariate normal or multivariate t distributions. Structural equation models designed for the cross-sectional analysis of quantitative family data and twins can also be extended to longitudinal data. But using standard software, it is again difficult to generalize the models to non-linear problems involving non-normal phenotypes and also to extended pedigrees.

In this paper, we propose an approach to the analysis of longitudinal family data that is based on a natural extension of the generalized linear mixed models (GLMMs) that we have already developed for cross-sectional family data. These models are appropriate for continuous and binary traits, for Poisson distributed counts or rates and for other phenotypes with any one of the many distributions that are available in the flexible Markov chain Monte Carlo (MCMC) environment in which the models are fitted (see below). They may be used both in nuclear families and extended pedigrees. The models are fitted using a MCMC-based approach, usually Gibbs sampling, within the general purpose software WinBUGS. By sampling from the joint posterior distribution of unknown parameters, this approach avoids high dimensional integration and thereby surmounts many of the practical difficulties encountered when fitting GLMMs for non-multivariate normal data when analytic likelihood estimation methods are used.

This paper describes the extension of GLMMs for cross-sectional family data to longitudinal family data.
In section 1 we motivate the extended model. In section 2 we test out the proposed methods on simulated data and section 3 describes a real data analysis based on the Framingham Heart Study. Section 4 is a general discussion.

1 Methods

The construction of GLMMs for cross-sectional data from randomly ascertained families has been described in detail elsewhere. A brief description is presented here before the extension to longitudinal data is outlined. We assume throughout that families are non-informatively ascertained, i.e. that they are obtained as a population-based sample, with no deliberate excess or under-sampling on the basis of the trait of interest.

Assume that a phenotypic measurement \( y_{ij} \) has been obtained from each member \( j \) of family \( i \), and that, conditional on covariates and random effects, the random error in each observed phenotype follows a distribution \( f(.) \) (e.g. normal, binomial, or Poisson). Then the model may be written:

\[
g(\mu_{ij}) = (\beta_0 + \beta_i) + x_{ij}^T \beta
\]

where \( x_{ij} \) is a transposed vector of observed covariate values; \( \beta \) is a corresponding vector of unknown regression coefficients; \( \beta_0 \) is the regression intercept; \( \beta_i \) is a random component to the intercept specific to the \( j \)th member of the \( i \)th family that represents the sum of a series of multivariate normal random effects (see below); \( g(.) \) is an appropriate link function; \( \theta \) (if required) is a ‘nuisance parameter’ such as the error variance of a normal distribution; \( \theta_0 \) is a vector that contains the random effects of all subjects in the \( i \)th family: its \( j \)th element is \( \theta_{0ij} \); \( \theta_0 \) is a covariance matrix reflecting the within-family covariance of the elements of \( \theta_0 \). The right-hand side of the first line of the model, \( (\beta_0 + \beta_i) + x_{ij}^T \beta \), is often called the linear predictor and may be denoted \( \mu_{ij} \).

If \( y_{ij} \) has a normal distribution, \( g(.) \) is usually taken to be the identity function (i.e. the linear predictor directly predicts \( y_{ij} \)), \( \theta \) is the residual error variance (\( \sigma^2_{e} \)), and model (1) is an extended form of conventional multiple regression. If \( y_{ij} \) is binary, \( g(.) \) is usually the logit function, \( \theta \) is not required and model (1) is an extended logistic regression model. Given a binary trait, \( g(.) \) can also be taken to be the probit function. If \( y_{ij} \) is a censored survival time (or a rate or count) and analysis is to be based on an extended Poisson regression model, as described in Scurrah et al., \( g(.) \) is the logarithmic function and \( \theta \) is again not required.

In the typical cross-sectional family models that we have described we use three variance components to parameterize the within-family covariance (or correlation) structure (\( \sigma^2_{A}, \sigma^2_{C}, \) and \( \sigma^2_{E} \)). The variance component \( \sigma^2_{A} \) denotes additive polygenic variance. That is, the variation that arises from the effect of individual alleles at multiple loci with no genetic interactions. Covariance consequent upon a shared nuclear family environment is denoted by \( \sigma^2_{C} \). The additional covariance between siblings due to a common sibship environment, not shared with parents, is denoted \( \sigma^2_{E} \). In traditional family designs, \( \sigma^2_{E} \) is completely confounded with polygenic dominance (\( \sigma^2_{D} \)) and, here, we choose to model only \( \sigma^2_{E} \). Nevertheless, models incorporating \( \sigma^2_{E} \) (e.g. in twin family studies, when \( \sigma^2_{A} \) and \( \sigma^2_{D} \) can be disentangled) or \( \sigma^2_{D} \) (maternal line effects) can be fitted if appropriate.

Therefore the total random effect \( \beta_{0ij} \) for each individual can be viewed as the sum of three distinct random effects (each associated with one of the three variance components):

\[
\beta_{0ij} = A_{ij} + C_{ij} + C_{Eij}
\]

where the \( A_{ij} \) terms reflect additive polygenic variance (variance, \( \sigma^2_{A} \)), the \( C_{ij} \) terms represent the effect of common nuclear family environment (variance, \( \sigma^2_{C} \)) and the \( C_{Eij} \) terms represent common sibling environmental effects (variance, \( \sigma^2_{E} \)). These are shared appropriately between relatives to build the complex within-family correlation structure that is required.

We have described the construction of such models using Gibbs sampling in WinBUGS both in nuclear families and in extended pedigrees. Under the standard nuclear family parameterization, and as is biologically appropriate, the conditional covariance between parents (on the scale of the linear predictor) is \( \sigma^2_{C} \) that between a parent and child is \( \sigma^2_{C} + \sigma^2_{A} \) and the covariance between two full siblings is \( 2\sigma^2_{C} + \sigma^2_{A} \). In extended pedigrees, the manner in which the random effects are generated also leads naturally to biologically appropriate covariances between more distant relatives; for example, grandparent-grandchild (\( \sigma^2_{C} \)), uncle-niece (\( \sigma^2_{C} \)) and first cousins (\( \sigma^2_{A} \)). It is also straightforward to ensure appropriate covariances between monozygous twins and half siblings.

We have demonstrated that the GLMMs generate approximately consistent estimates both of conventional regression coefficients (fixed effects) and of components of variance. We have also shown that individual random effects (specifically the \( A_{ij} \)) can successfully be used as adjusted phenotypes in subsequent linkage analyses.

A series of working WinBUGS files, with a variety of different purposes, may be downloaded from our website: http://www.pw.le.ac.uk/research/HCG/gebugs.html. The three files for the longitudinal family study analyses were all constructed in WinBUGS 1.429 and should definitely be able to run on that platform.

It is now straightforward to extend the GLMMs for cross-sectional family studies to incorporate genetic and environmental contributions to the longitudinal change in a phenotype over time. If \( T_{ijk} \) is the time (sometimes age) of the \( k \)th measurement in the \( j \)th individual in the \( i \)th family, model (1) can be extended:

\[
g(\mu_{ijk}) = (\beta_0 + \beta_i + \beta_T) + x_{ijk}^T \beta + (\beta_T + \beta_{fijk}) \times T_{ijk}
\]

The transposed vector of covariates, \( x_{ijk} \), is now indexed by time as well as family and subject. It can therefore incorporate ‘time-dependent’ covariates that can change in their observed values with time. The term \( \beta_T \) estimates the average increase in response (on the scale of the linear predictor) associated with a one-unit increase in time. That is, \( \beta_T \) reflects the overall rate of change.
longitudinal change in the phenotype of interest. The subject level random effect, \( \beta_{ji} \), is analogous to \( \beta_{0ij} \) (see above). It is a composite random effect, specific to the \( j \)th member of the \( i \)th family. It reflects polygenic and shared environmental contributions to the rate of phenotypic change over time (\( \beta_{ji} = A_{ji} + C_{ji} + G_{ji} \)). \( \mathbf{P}_{ji} \) is a family level random effects vector with \( j \)th element \( \beta_{ji} \). \( \mathbf{V}_T \) is a covariance matrix reflecting the within-family covariance of the elements of \( \beta_{ji} \) and is parameterized by variance components \( \tau^2_A, \tau^2_C, \) and \( \tau^2_G \). The generation of the \( \mathbf{P}_{ji} \) random effects is so organized that \( \mathbf{V}_T \) has the same overall structure as \( \mathbf{V}_0 \) (see the WinBUGS files for longitudinal family data on our website http://www.prw.le.ac.uk/research/HCG/gebugs.html). However, the actual magnitudes of the variance components \( \tau^2_A, \tau^2_C, \) and \( \tau^2_G \) are assumed here to be unrelated to the magnitudes of those that parameterize \( \mathbf{V}_0 \) (i.e., \( \sigma^2_A, \sigma^2_C, \) and \( \sigma^2_G \)). Nevertheless, just as the size of \( \tau^2_C \) indicates the extent to which additive polygenic effects influence the mean value of the phenotype (an intercept effect), the estimated size of \( \tau^2_A \) indicates the influence of additive polygenic effects on the rate of change of the phenotype over time (a slope effect).

Model parameters may be interpreted as described previously.8,25 Interpretation of the fixed regression coefficients depends in the usual way on the chosen link function. Covariances and variances are also interpreted on the scale of the linear predictor. Thus, in a binary model with a logit(.) link, a fixed regression coefficient \( \beta_{ij} \) estimates the increase in the log(odds) of a positive phenotype associated with a one-unit increase in the value of covariate \( X \). Similarly, the covariance between the log(odds) of a positive phenotype in two full siblings is \( \sigma^2_A + \sigma^2_C + \sigma^2_G \), while the corresponding covariance of the rate of linear increase in log(odds) between two siblings is \( \frac{1}{2} \tau^2_A + \tau^2_C + \tau^2_G \). In normal models that incorporate an unshared residual error term (\( \sigma^2_R \)) one can also calculate the narrow sense heritability of the phenotype:6: \( h^2_N = \frac{\sigma^2_A}{\sigma^2_R} \). Furthermore, even if the phenotype is not normally distributed, a \( \tau^2_A \) term can usually be added to reflect unshared residual variance (on the scale of the linear predictor) in the gradient from subject to subject. This allows one to calculate, on the scale of the linear predictor, the narrow sense heritability of the rate of linear change in the phenotype over time:

\[ h^2_N = \frac{\tau^2_A}{\tau^2_A + \tau^2_C + \tau^2_G} \]

Because a Bayesian approach is used, prior distributions must be specified for all unknown parameters. For reasons we have described previously,8,25 vague priors were used for all model parameters. For example, the priors for fixed effects were specified as flat normal distributions: normal(mean = 0, variance = 104). The priors for precision parameters (inverse variances) were generally given Pareto(0, 0.01) prior distributions. This specification is precisely equivalent to placing a uniform prior on each variance component in the range 0–100.34 Given the sample sizes of the analyses described in this paper, there is likely to be little sensitivity to the choice of priors among the class that can reasonably be viewed as vague.35 A relatively informal approach was taken to the assessment of convergence for the simulated datasets. Visual inspection of trace plots and densities was performed and plots of the Gelman–Rubin statistic29,36 were used to determine whether the parallel chains had converged.

2 Simulated data analysis

The models described in section 1 were applied to a range of datasets that were simulated using S-PLUS 2000.37 The primary purpose of the simulation work that will be described in this section was to investigate whether a correctly specified model of the type motivated in section 1 would generate estimates that were consistent and that 95% credible intervals (in this setting, a Bayesian equivalent to 95% confidence intervals) had appropriate coverage. A range of different simulated ‘target’ values was used, both for the regression coefficients and for the variance components. Four primary sets of simulation scenarios were constructed. These may be categorized by whether or not they involve a continuous normally distributed trait or a binary trait, and whether the simulated pedigrees were nuclear families or extended pedigrees. Unless otherwise stated, all MCMC analyses were based upon the running of three parallel chains. The first 1000 iterations of each chain were discarded as a burn-in, and formal inferences about the posterior distribution were based on appropriate summaries of iterations 1000–10000. All iterations in the stated range were analysed, despite the autocorrelation between adjacent iterations, since for the conventional types of summary in which we were interested: ‘there is no real advantage in thinning except to reduce storage requirements’.29

2.1 A normal trait in nuclear families

Simulation was based upon model (2) with a normally distributed response variable and an identity link:

\[ y_{ijk} \sim \text{normal}(\mu_{ijk}, \sigma^2_e) \]

\( y_{ijk} \) denotes the full linear predictor on the right-hand side of model (2). In addition to the fixed parameters reflecting the overall intercept (\( \beta_0 \)) and slope (\( \beta_2 \)) there are two observed covariate vectors—\( \text{bvar} \) (binary) and \( \text{qvar} \) (continuous)—with coefficients \( \beta_{bvar} \) and \( \beta_{qvar} \). The covariance matrices (\( \mathbf{V}_0 \) and \( \mathbf{V}_T \)) associated with the random effect vectors for the regression intercept (\( \beta_{qi} \)) and slope (\( \beta_{2i} \)) are parameterized with terms \( \sigma^2_A, \sigma^2_C, \sigma^2_G, \) and \( \tau^2_A, \tau^2_C, \tau^2_G \), respectively.

Table 1 summarizes the combined MCMC results from 250 independently simulated datasets based on a common simulation model of this form. Each dataset was based on the simulation of 50 nuclear families containing five members (two parents and three children) with observations made at five equally spaced time-points. A random 10% of individuals and 10% of observations were deleted, and all individuals with less than three observations were also deleted. These random deletions were deliberately invoked in order to break up the otherwise balanced nature of the simulated data. Across 50 typical datasets (2500 families generated) this particular simulation mechanism generated 1468 families with five members and 815, 192, 22, and 3 families, respectively, with four, three, two, and one member. Overall, 2009 families had two parents, 463 had one, and 28 had no parents. Correspondingly, 1814 families had three children, 619 had two children, 62 had one, and 5 had none in the dataset. Across the fifty datasets, there were 11 223 individuals and of these, 6689 had observations at all five time-points, 3751 at four, and 783 at three. No attempt was made to impute missing data.

Subsequently, six sets of simulations were also generated. Each of these six scenarios involved the generation and analysis
of 25 independent datasets with a variety of different target values for each parameter. In addition, the 250 independent datasets simulated under the scenario detailed in Table 1 were arbitrarily divided up into 10 sets of 25 simulations. This implies a total of 16 groups (6 + 10) each consisting of 25 independent datasets. Figure 1 details the relationship between target values and mean MCMC-based estimates for each parameter across the 16 simulation groups. Figure 2 details the corresponding

### Table 1

Summary of MCMC results from 250 independent simulations each based on a normal phenotype in 50 nuclear families of maximum size five people

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Target value</th>
<th>Mean of empirical posterior means</th>
<th>Empirical posterior standard deviation</th>
<th>Ratio</th>
<th>Coverage of 95% credible intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>β₀</td>
<td>140</td>
<td>140.06</td>
<td>0.702</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td>β₂</td>
<td>2</td>
<td>-2.04</td>
<td>0.665</td>
<td>1.02</td>
<td>0.96</td>
</tr>
<tr>
<td>σ₂</td>
<td>25</td>
<td>25.49</td>
<td>4.381</td>
<td>1.02</td>
<td>0.94</td>
</tr>
<tr>
<td>σ₂</td>
<td>5</td>
<td>6.27</td>
<td>5.911</td>
<td>1.25</td>
<td>0.92</td>
</tr>
<tr>
<td>σ₂</td>
<td>10</td>
<td>10.97</td>
<td>3.417</td>
<td>1.10</td>
<td>0.96</td>
</tr>
<tr>
<td>σ₂</td>
<td>10</td>
<td>9.99</td>
<td>0.590</td>
<td>1.00</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Effects on the mean level of response (intercept effects)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Target value</th>
<th>Mean of empirical posterior means</th>
<th>Empirical posterior standard deviation</th>
<th>Ratio</th>
<th>Coverage of 95% credible intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>β₀</td>
<td>-1</td>
<td>-1.01</td>
<td>0.267</td>
<td>1.01</td>
<td>0.96</td>
</tr>
<tr>
<td>τ₂</td>
<td>1</td>
<td>1.11</td>
<td>0.468</td>
<td>1.11</td>
<td>0.97</td>
</tr>
<tr>
<td>τ₂</td>
<td>2</td>
<td>2.25</td>
<td>0.393</td>
<td>1.13</td>
<td>0.94</td>
</tr>
<tr>
<td>τ₂</td>
<td>1</td>
<td>1.08</td>
<td>0.393</td>
<td>1.08</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Effects on the rate of change with time (gradient effects)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Target value</th>
<th>Mean of empirical posterior means</th>
<th>Empirical posterior standard deviation</th>
<th>Ratio</th>
<th>Coverage of 95% credible intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>β₀</td>
<td>100</td>
<td>110</td>
<td>0.702</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td>β₀</td>
<td>120</td>
<td>110</td>
<td>0.702</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td>β₀</td>
<td>140</td>
<td>110</td>
<td>0.702</td>
<td>1.00</td>
<td>0.96</td>
</tr>
</tbody>
</table>

- a Formally, the square root of the mean of the posterior variances.
- b Ratio = empirical mean of posterior means / target value.

Figure 1 Accuracy of MCMC estimates vs target values in simulated data. Mean of estimated posterior means vs target values for each simulated parameter across 16 sets of 25 independent simulations, each consisting of 50 nuclear families including up to 2 parents and 3 children (see text for further details). Dashed line denotes the line of equality: if all parameters in all models were estimated perfectly all of the plotted points (X) would lie on the line. Departures from the line are, of course, to be expected because of random variation. Although there are 16 sets of simulations, a number of the plotted points so precisely overlie one another that <16 plotted points are visible in each frame of the figure.
coverage of 95% credible intervals for each parameter. In both figures, a number of the plotting points so closely overlie one another that less than 16 individual points can be distinguished in each figure. The grouping of the simulations in this manner and the graphical presentation of the results in Figures 1 and 2 not only allow one to examine the extent of bias and coverage at a particular set of target values, but also to determine how that bias and coverage might change as the target values vary.

Overall (see Table 1 and Figures 1 and 2), the means of the posterior means were close to the simulated target values and the mean coverage was close to nominal (95.3%) with no obvious pattern to the small discrepancies from 95%. However, there was a consistent tendency for the estimates of all of the variance components to be slightly raised. The geometric mean of the posterior mean target value ratios for all seven variance-covariance parameters was ~1.10. A positive bias of up to 20–30% in the estimation of variance terms in Gibbs-sampling-based GLMMs has been noted before. 27 The bias tends to be greater in small datasets and if the true variance being estimated is close to zero. 8 If it is viewed as being a concern, it can be mitigated by using the posterior mode as a summary measure rather than the posterior mean. 27,38 That this bias is evident in this particular example and not in the examples described in sections 3.2 and 3.3 is because this was deliberately constructed as a small problem. In fact, it might reasonably be emphasized that despite the fact that a complex GLMM is being fitted to a sample consisting of only 50 nuclear families, the bias in the parameter estimates is very small—and certainly not enough to lead to misleading inferences.

2.2 A normal trait in extended pedigrees

As a familiar and convenient template for the family structures in the simulations involving extended pedigrees, we used the extended pedigree structures that were used in the simulated out-bred population generated for the Genetic Analysis Workshop 12 (GAW12). 39 However, for our purpose, we simulated entirely new exposure and outcome data under the models specified in this paper and made no use of the outcome/exposure data or simulating models that were supplied and analysed as part of GAW12.

Using this template, we generated family structures corresponding to 23 extended pedigrees (n = 1497 individuals). Each pedigree consisted of four or five generations, and the number of identified individuals in each pedigree varied from 37 to 128. Given the structure of these families, we simulated phenotypic and covariate data for the 1000 subjects who were deemed to be ‘alive’ and ‘available for study’. 39 The remaining 497 individuals were identified in the pedigrees but were allocated missing value codes for both the phenotypic and covariate data.

Figure 2 Coverage of MCMC estimates in simulated data. Empirical coverage of the 95% credible interval for the estimate of each simulated parameter across 16 sets of 25 independent simulations, each consisting of up to 50 nuclear families including up to 2 parents and 3 children (see text for further details). Dashed line denotes the line corresponding to a coverage of 95%; if all estimates in all scenarios had had precisely 95% coverage, all of the plotted points (×) would lie on the line. Departures from the line are, of course, to be expected because of random variation. Although there are 16 sets of simulations, a number of the plotted points so precisely overlie one another that <16 plotted points are visible in each frame of the figure.
Simulation was again based upon model (2) with a normally distributed response variable and an identity link. The only change in model structure from the model used to simulate the normal trait in nuclear families (section 2.1) was that there were three observed covariate vectors \( (x_1, x_2, x_3) \) with coefficients \( \beta_1, \beta_2, \) and \( \beta_3 \). The data were simulated so as to reflect a realistic dataset with systolic blood pressure (SBP) as the phenotype of interest. The fixed covariates were generated so as to represent: proportion of ideal body weight (continuous); salt intake above or below the population median (binary), and current smoking (binary). Longitudinal measurements were made at five time points at two yearly intervals; time was zeroed at the time of the central measurement. A random 20% of the measurements were deleted from the dataset.

An annotated S-PLUS script file (normal.extended.pedigrees.example.ssc) that generates a single simulated dataset of the form described in the preceding paragraph may be found at website http://www.prw.le.ac.uk/research/HCG/gebugs.html. The website also includes an annotated copy of a WinBUGS file (normal.extended.pedigrees.example.odc) into which the two datasets written out by the S-PLUS file have already been imported. The WinBUGS file is ready to run in WinBUGS version 1.4.29 (or version 1.4.1). The distribution of key variables in this particular simulated data set is summarized in Box 1. By clicking on the arrows indicating the WinBUGS 'folds'29 for Target Values and Stats under RESULTS in the annotated WinBUGS file, it can be seen that all point estimates are close to simulated target values, and that all 95% credible intervals include the corresponding simulated target values.

As a more rigorous test of the consistency and coverage of point and interval estimates, Table 2 summarizes the combined MCMC results from the analysis of 20 independently simulated datasets of the same form. There was no obvious bias to the posterior means and the coverage was close to nominal.

### 2.3 A binary trait in nuclear families

Simulation was based upon model (2) with a binary trait and a logit link:

\[
Y_{ijk} \sim \text{Bernoulli}(\mu_{ijk}); \quad \logit(\mu_{ijk}/(1 - \mu_{ijk})) = \eta_{ijk}.
\]

The only change in model structure from the model used to simulate a normal trait in nuclear families (section 2.1) was that the model incorporated an additional residual variance term \( \tau^2 \) reflecting random variation from subject to subject in the gradient reflecting the rate at which the log-odds of exhibiting the trait increased with age. An annotated S-PLUS script file (binary.nuclear.families.example.ssc) that generates a single simulated dataset in this form, and an annotated WinBUGS file (binary.nuclear.families.example.odc) into which this dataset has been incorporated can be found on our web site. A comparison of Target Values and Stats under RESULTS in the WinBUGS file demonstrates that point estimates are close to nominal.

### Box 1 Key variables in simulated dataset

| Total number of longitudinal measurements | 4017 |
| Systolic blood pressure                  | mean = 121.9 mm Hg; SD = 9.6 mm Hg; minimum = 87 mm Hg; maximum = 157 mm Hg |
| Weight relative to ideal body weight     | mean = 1.35; SD = 0.39; minimum = 0.7; maximum = 2.0 |
| Salt intake above population median      | yes = 51.4%; no = 48.6% |
| Current smoker                           | yes = 30.4%; no = 69.6% |

### Table 2 Summary of MCMC results from 20 independent simulations based on a normally distributed phenotype simulated in 23 extended pedigrees

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Target value</th>
<th>Mean of empirical posterior means</th>
<th>Empirical posterior standard deviation</th>
<th>Ratio</th>
<th>Coverage of 95% credible intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>120</td>
<td>120.01</td>
<td>0.353</td>
<td>1.00</td>
<td>0.95</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>0.5</td>
<td>0.50</td>
<td>0.010</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>5</td>
<td>5.03</td>
<td>0.301</td>
<td>1.01</td>
<td>1.00</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>5</td>
<td>5.01</td>
<td>0.144</td>
<td>1.00</td>
<td>0.95</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>20</td>
<td>19.64</td>
<td>1.868</td>
<td>0.98</td>
<td>0.90</td>
</tr>
<tr>
<td>( \sigma^2_c )</td>
<td>8</td>
<td>8.36</td>
<td>2.016</td>
<td>1.04</td>
<td>0.90</td>
</tr>
<tr>
<td>( \sigma^2_s )</td>
<td>15</td>
<td>15.15</td>
<td>1.979</td>
<td>1.01</td>
<td>1.00</td>
</tr>
<tr>
<td>( \sigma^2_e )</td>
<td>10</td>
<td>9.94</td>
<td>0.307</td>
<td>0.99</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Effects on the rate of change with time (gradient effects)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Target value</th>
<th>Mean of empirical posterior means</th>
<th>Empirical posterior standard deviation</th>
<th>Ratio</th>
<th>Coverage of 95% credible intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_T )</td>
<td>2</td>
<td>1.99</td>
<td>0.109</td>
<td>1.00</td>
<td>0.95</td>
</tr>
<tr>
<td>( \tau^2_x )</td>
<td>2</td>
<td>2.00</td>
<td>0.292</td>
<td>1.00</td>
<td>0.95</td>
</tr>
<tr>
<td>( \tau^2_c )</td>
<td>0.8</td>
<td>0.78</td>
<td>0.237</td>
<td>0.98</td>
<td>0.95</td>
</tr>
<tr>
<td>( \tau^2_s )</td>
<td>0.4</td>
<td>0.42</td>
<td>0.167</td>
<td>1.06</td>
<td>0.9</td>
</tr>
</tbody>
</table>

\( a \) Formally, the square root of the mean of the posterior variances.

\( b \) Ratio = empirical mean of posterior means \div target value.
Effects on the rate of change with time (gradient effects)

Effects on the mean level of response (intercept effects)

Twenty independently simulated datasets were constructed. Table 4 summarizes the results of the MCMC analysis applied to these datasets. This example exhibited the same small positive bias in the posterior means for the variance terms that was illustrated in the example considered in section 2.1. In fact the bias was slightly larger in this example: the geometric mean of the posterior mean over target value ratios for all seven variance-covariance parameters was ~1.15. The explanation for the bias and the potential solutions are the same as described in section 2.1. That the bias is evident in this particular example emphasizes the fact that binary phenotypes are far less informative than quantitative traits, particularly when one is fitting a complex random effects model.

### 2.5 Overall assessment of model fit and statistical inferences

On the basis of informal scrutiny of the developing MCMC chains and of the plots of the Gelman-Rubin diagnostic, it was clear that MCMC convergence was typically quick (certainly within 1000 iterations) under all scenarios. This is well illustrated in the diagnostic plots included at the bottom of the exemplar WinBUGS files available on our website. As is inevitable in these models, there is substantial autocorrelation from iteration to iteration within a chain; particularly so, for the random effect variances. As a necessary response, chains were therefore run long enough to ensure that any effect that this may have had on statistical inference was minimal.

In relation to the quality of the resultant statistical inferences, across all scenarios, the only issue beyond the small positive bias in the posterior means for the covariance terms (see above) was that there was perhaps a tendency for $\sigma^2_c$ in particular, to be overestimated (by up to 25%). We have no obvious explanation for this particular phenomenon; it is not a prominent feature of our GLMM models in general. We also note that, as a characteristic, it was not shared by $\tau^2_a$; indeed, it was $\tau^2_a$ that was most consistently overestimated amongst the variance

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Target value</th>
<th>Mean of empirical posterior means</th>
<th>Empirical posterior standard deviation$^a$</th>
<th>Ratio$^b$</th>
<th>Coverage of 95% credible intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>0</td>
<td>$-0.012$</td>
<td>0.079</td>
<td>Not defined</td>
<td>1</td>
</tr>
<tr>
<td>$\beta_{b\text{var}}$</td>
<td>1</td>
<td>$1.025$</td>
<td>0.110</td>
<td>$1.03$</td>
<td>0.95</td>
</tr>
<tr>
<td>$\beta_{q\text{var}}$</td>
<td>$-0.5$</td>
<td>$-0.474$</td>
<td>0.176</td>
<td>0.95</td>
<td>1</td>
</tr>
<tr>
<td>$\sigma^2_a$</td>
<td>1.8</td>
<td>$1.714$</td>
<td>0.637</td>
<td>0.95</td>
<td>0.85</td>
</tr>
<tr>
<td>$\sigma^2_c$</td>
<td>0.75</td>
<td>$0.925$</td>
<td>0.375</td>
<td>1.23</td>
<td>0.95</td>
</tr>
<tr>
<td>$\sigma^2_{\alpha_0}$</td>
<td>1</td>
<td>$1.102$</td>
<td>0.351</td>
<td>1.10</td>
<td>1</td>
</tr>
</tbody>
</table>

$^a$ Formally, the square root of the mean of the posterior variances.

$^b$ Ratio = empirical mean of posterior means ÷ target value.

### 2.4 A binary trait in extended families

Simulation was again based upon model (2) with a binary response variable and a logit link. Family composition was based upon the structure of the same set of extended pedigrees used in section 2.2. Data were simulated at five equally spaced time-points with 20% of observations deleted at random.

the simulated targets and that the 95% credible intervals all include their corresponding target values. The single simulated dataset in the example file is based upon 1000 nuclear families, and upon a longitudinal binary phenotype assessed on up to seven equally spaced occasions. A random 20% of the individuals and 20% of observations were deleted, and all individuals with less than three non-deleted assessments were also excluded.

Table 3 summarizes the combined MCMC results from 20 independently simulated datasets based on the same model, but using a different set of target values for the model parameters. Each scenario was based on the simulation of 500 nuclear families each containing five members (two parents and three children) with observations made at five equally spaced time-points. A random 10% of individuals and 25% of observations were deleted, and all individuals with less than three observations were deleted. Across five typical datasets (2500 families generated) this simulation mechanism led to 591 observations were deleted. Across five typical datasets (2500 families generated) this simulation mechanism led to 591

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Target value</th>
<th>Mean of empirical posterior means</th>
<th>Empirical posterior standard deviation$^a$</th>
<th>Ratio$^b$</th>
<th>Coverage of 95% credible intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>0</td>
<td>$-0.012$</td>
<td>0.079</td>
<td>Not defined</td>
<td>1</td>
</tr>
<tr>
<td>$\beta_{b\text{var}}$</td>
<td>1</td>
<td>$1.025$</td>
<td>0.110</td>
<td>$1.03$</td>
<td>0.95</td>
</tr>
<tr>
<td>$\beta_{q\text{var}}$</td>
<td>$-0.5$</td>
<td>$-0.474$</td>
<td>0.176</td>
<td>0.95</td>
<td>1</td>
</tr>
<tr>
<td>$\sigma^2_a$</td>
<td>1.8</td>
<td>$1.714$</td>
<td>0.637</td>
<td>0.95</td>
<td>0.85</td>
</tr>
<tr>
<td>$\sigma^2_c$</td>
<td>0.75</td>
<td>$0.925$</td>
<td>0.375</td>
<td>1.23</td>
<td>0.95</td>
</tr>
<tr>
<td>$\sigma^2_{\alpha_0}$</td>
<td>1</td>
<td>$1.102$</td>
<td>0.351</td>
<td>1.10</td>
<td>1</td>
</tr>
</tbody>
</table>

$^a$ Formally, the square root of the mean of the posterior variances.

$^b$ Ratio = empirical mean of posterior means ÷ target value.

Overall (see Table 3), there was no consistent bias to the parameter estimates and the coverage was again close to nominal.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Target value</th>
<th>Mean of empirical posterior means</th>
<th>Empirical posterior standard deviation$^a$</th>
<th>Ratio$^b$</th>
<th>Coverage of 95% credible intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>0</td>
<td>$-0.012$</td>
<td>0.079</td>
<td>Not defined</td>
<td>1</td>
</tr>
<tr>
<td>$\beta_{b\text{var}}$</td>
<td>1</td>
<td>$1.025$</td>
<td>0.110</td>
<td>$1.03$</td>
<td>0.95</td>
</tr>
<tr>
<td>$\beta_{q\text{var}}$</td>
<td>$-0.5$</td>
<td>$-0.474$</td>
<td>0.176</td>
<td>0.95</td>
<td>1</td>
</tr>
<tr>
<td>$\sigma^2_a$</td>
<td>1.8</td>
<td>$1.714$</td>
<td>0.637</td>
<td>0.95</td>
<td>0.85</td>
</tr>
<tr>
<td>$\sigma^2_c$</td>
<td>0.75</td>
<td>$0.925$</td>
<td>0.375</td>
<td>1.23</td>
<td>0.95</td>
</tr>
<tr>
<td>$\sigma^2_{\alpha_0}$</td>
<td>1</td>
<td>$1.102$</td>
<td>0.351</td>
<td>1.10</td>
<td>1</td>
</tr>
</tbody>
</table>

$^a$ Formally, the square root of the mean of the posterior variances.

$^b$ Ratio = empirical mean of posterior means ÷ target value.

Table 3 summarizes the combined MCMC results from 20 independently simulated datasets based on the same model, but using a different set of target values for the model parameters. Each scenario was based on the simulation of 500 nuclear families each containing five members (two parents and three children) with observations made at five equally spaced time-points. A random 10% of individuals and 25% of observations were deleted, and all individuals with less than three observations were deleted. Across five typical datasets (2500 families generated) this simulation mechanism led to 591 observations were deleted. Across five typical datasets (2500 families generated) this simulation mechanism led to 591
Effects on the mean level of response (intercept effects)

younger subjects but, given that age was included in the model study had potentially many more observations than the observations per phenotyped subject). The older subjects in the 24 174 observations in 2860 individuals (on average, 8.45 interpolated values for drink and cpd. This represented a total of values for SBP, age, height, and weight, and non-missing or was based entirely on those records that had non-missing values bracketing the missing observation. This increased the therefore replaced by the interpolated mean of the observed period of study and each missing value of drink and cpd was one or two distinct values for these variables throughout the examinations each. The majority of individuals exhibited only cohort members were subject to a maximum of 21 and 18 time-points, respectively, in the first cohort, although and ‘cpd’ (number of cigarettes per day) were measured at only terms in age were included in the fixed effects component of variables were treated as continuous. All continuous covariates were centred at or close to their mean. Quadratic and cubic variables were treated as continuous. All continuous covariates were included in the fixed effects component of the model in order to allow for a non-linear increase in SBP with age.

The covariates ‘drink’ (number of standard drinks per week) and ‘cpd’ (number of cigarettes per day) were measured at only 11 and 18 time-points, respectively, in the first cohort, although cohort members were subject to a maximum of 21 examinations each. The majority of individuals exhibited only one or two distinct values for these variables throughout the period of study and each missing value of drink and cpd was therefore replaced by the interpolated mean of the observed values bracketing the missing observation. This increased the amount of useful observations without excluding important covariates from the model. The modelling that we report here was based entirely on those records that had non-missing values for SBP, age, height, and weight, and non-missing or interpolated values for drink and cpd. This represented a total of 24 174 observations in 2860 individuals (on average, 8.45 observations per phenotyped subject). The older subjects in the study had potentially many more observations than the younger subjects but, given that age was included in the model and that our method does not require balanced data, this was not problematic.

3 Real data analysis

We analysed the longitudinal Framingham Family Cohort dataset, provided for the GAW13, which included 4692 individuals in 330 pedigrees.40 The primary response variable of the GLMMs described here was the SBP measured over time. Explanatory covariates for all models included: sex, age, height, weight, number of standard drinks per week, and number of cigarettes per day. Sex (0 = female; 1 = male) was analysed as a binary covariate. Other variables were treated as continuous. All continuous covariates were centred at or close to their mean. Quadratic and cubic terms in age were included in the fixed effects component of the model in order to allow for a non-linear increase in SBP with age.

The covariates ‘drink’ (number of standard drinks per week) and ‘cpd’ (number of cigarettes per day) were measured at only 11 and 18 time-points, respectively, in the first cohort, although cohort members were subject to a maximum of 21 examinations each. The majority of individuals exhibited only one or two distinct values for these variables throughout the period of study and each missing value of drink and cpd was therefore replaced by the interpolated mean of the observed values bracketing the missing observation. This increased the amount of useful observations without excluding important covariates from the model. The modelling that we report here was based entirely on those records that had non-missing values for SBP, age, height, and weight, and non-missing or interpolated values for drink and cpd. This represented a total of 24 174 observations in 2860 individuals (on average, 8.45 observations per phenotyped subject). The older subjects in the study had potentially many more observations than the younger subjects but, given that age was included in the model and that our method does not require balanced data, this was not problematic.

The classes of models fitted in this analysis are equivalent to those described in section 1. Models were run using two parallel MCMC chains for 20 000 iterations after a burn-in of 15 000 iterations. The raw traces of the MCMC chains and the Gelman–Rubin diagnostic plots29,36 for the individual parameters of interest suggested that this length of burn-in and analysis were adequate.

The problem of how to model continuous SBP when some individuals are on blood pressure treatment is difficult.41 Since complex modelling of SBP was not the main aim of this particular analysis, we chose the simplest valid method of adjustment based on known average treatment effects.41–43 This involved adding a constant (10 mm Hg) to each phenotype when an individual was on treatment, to reflect the ‘true’ SBP that might have been observed if the individual had not been on treatment. Although the model is Bayesian, vague priors were used throughout (see section 1), and in consequence the effect of each covariate on the phenotype could be assessed using a pseudo-likelihood-based approach. For example, an empirical posterior mean divided by its empirical standard deviation could be treated as an approximate standardized normal deviate (Z-score), and 95% credible intervals could be interpreted as ~95% confidence intervals.17,44

The results of fitting the variance components GLMM in WinBUGS are shown in Table 5. Both posterior means and 95% credible intervals are presented for each parameter. We deliberately excluded certain intermediate phenotypes (such as cholesterol and BMI) from the model, even though they were clearly related to SBP. This was because they were potentially on causal pathways of interest, and we did not want to model away their contribution to variation in SBP.

The size of $\tau^2$ (relative to the other variance components for the slope) implied a narrow sense heritability ($h^2$) estimate of only 9.0%. Furthermore, a formal test at the null hypothesis

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Target value</th>
<th>Mean of empirical posterior means</th>
<th>Empirical posterior standard deviation</th>
<th>Ratio</th>
<th>Coverage of 95% credible intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>-0.3</td>
<td>-0.311</td>
<td>0.117</td>
<td>1.04</td>
<td>0.9</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.05</td>
<td>0.051</td>
<td>0.005</td>
<td>1.03</td>
<td>0.95</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.5</td>
<td>0.519</td>
<td>0.137</td>
<td>1.04</td>
<td>1.0</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.2</td>
<td>0.209</td>
<td>0.121</td>
<td>1.05</td>
<td>1.0</td>
</tr>
<tr>
<td>$\sigma^2_a$</td>
<td>1</td>
<td>1.241</td>
<td>0.541</td>
<td>1.24</td>
<td>0.95</td>
</tr>
<tr>
<td>$\sigma^2_c$</td>
<td>0.6</td>
<td>0.729</td>
<td>0.292</td>
<td>1.22</td>
<td>0.95</td>
</tr>
<tr>
<td>$\tau^2_e$</td>
<td>1.25</td>
<td>1.22</td>
<td>0.350</td>
<td>0.98</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Effects on the rate of change with time (gradient effects)

\[ \beta_T = 1 \]

\[ \tau^2_T = 0.8 \]

\[ \tau^2_c = 1 \]

\[ \tau^2_e = 0.4 \]

\[ Z = Z_{\text{score}} \]

\[ 95\% \text{ credible intervals} \]

\[ \text{Ratio} = \text{Empirical mean of posterior means} / \text{target value}. \]

\[ \text{Coverage of 95\% credible intervals} \]
additional year of age was a component of the mean rate of change of SBP with each increment of years. Having taken account of these other covariates, the linear regression suggested that the additive genetic effects represent the largest component of variance for SBP at baseline (\(\sigma^2_A\)). In contrast, the variance components estimates for the intercept (\(\sigma^2_0\)) and age (\(\sigma^2_2\)) suggest that additive genetic effects account for a greater part of the variability.

A number of covariates exhibit a significant association with SBP. The Framingham Heart Study results from primary generalized linear mixed model are shown in Table 5. The intercept and age effects can be read from the estimates for the quadratic and cubic fixed effect terms for age. The solid line denotes the modelled profile of mean SBP with age for a male of height 173 cm, who weighs 80 kg, who neither drinks alcohol nor smokes cigarettes and who was recruited into study cohort 2 (the more recently ascertained of the two Framingham cohorts). The dashed line denotes the equivalent profile of SBP with age for a female of height 159 cm weighing 65 kg. Other covariate patterns will shift the curve upwards or downwards.

A non-linear effect was also apparent, since the credible intervals for both the quadratic and cubic fixed effect terms for age exclude 0. The non-linear relationship between mean SBP and age that is implied by the model is illustrated in Figure 3. Separate profiles are presented for males and females. The plots for either sex are presented for subjects that have a weight and height that correspond to the sex-specific means of these two measures across all the 24 174 observations that were included in the analysis. The age range plotted corresponds to the central 90% of the observed distribution of age across all observations in the analysis. In order to keep the illustrative example as simple as possible, the model presented in Table 5 contains no interactions between age and gender. The sex-specific SBP profiles with age in Figure 3 are therefore parallel. Such interactions could easily be added and it would have been possible to model entirely independent relationships between SBP and age in men and in women.

Results changed little when \(\sigma^2_0\) was excluded, when \(\sigma^2_2\) and \(\sigma^2_3\) were excluded, and when \(\sigma^2_C\) and \(\sigma^2_{Cs}\) were excluded. This suggests that the analytic findings were robust to the particular choice of correlation/covariance structure that we adopted. In addition, there was little change when the SBP adjustment (for those on blood pressure treatment) was changed to +5 mm Hg or +20 mm Hg (instead of +10 mm Hg).

### 4 Discussion

Longitudinal family studies have an important role to play in attempts to dissect the aetiological architecture of the complex genetic and environmental factors that influence SBP.
diseases. They will help us to look more closely not only at the aetiological determinants that are responsible for initial disease causation, but also, beyond these, at the genetic and environmental determinants that modulate the subsequent natural history of a disease. Knowledge of such determinants may pay high dividends not only in terms of prognostic assessment, but also in the development of new therapeutic modalities that can be biologically targeted and may, potentially, be tailored to the prognostic profile of individual patients.

There is increasing international investment in large case series for major complex diseases, such as the British Medical Research Council’s National DNA Collections Initiative (http://www.mrc.ac.uk), and in large general population epidemiological studies that are linked to health records, such as UK Biobank46 (http://www.biobank.ac.uk/), Iceland’s deCODE project47,48 (http://sunsite.berkeley.edu/biotech/iceland/), the Estonian Gene Bank Project (http://www.geenivaramu.ee/), CONOR (the cohort of Norway)49 and the Western Australian Data Linkage Project50 (http://www.populationhealth.uwa.edu.au/welcome/research/dlu/linkage). Some of these initiatives are able to delineate some family structure by design (e.g. deCODE, CONOR, and the Western Australian Data Linkage Project) while others (e.g. UK Biobank) are actively investigating the scientific and ethical issues that underpin the possibility of record linkage to identify close relatives after recruitment. Studies such as these are a rich potential source of longitudinal family data and will become more informative over time. The optimal analysis of data from these studies requires that we are able to model changes in a phenotype over time, not only in individuals, but also in relative pairs, nuclear families and in extended families.

The particular approach to the analysis of such data that we describe in this paper generalizes a pre-existing method for cross-sectional family data to longitudinal family data, but we are not the first to consider such a generalization. For example, approaches based on the covariance components models described by Lange and colleagues5,21 or on structural equation modelling22,23 may be viewed as extensions of precisely this type. But, for the reasons we discussed in the beginning, these methods are currently restricted in their potential application by the constraints of available software. Equivalently, de Andrade et al.24 describe an approach to longitudinal family data that extends an earlier method for linkage analysis designed for cross-sectional family data.51 However, their work has a different focus to ours. First, it is aimed specifically at quantitative traits. Second, they have a primary interest in incorporating the effect of a major gene linked to an observed marker. Third, their model is designed to address scientific questions of the form: ‘does heritability vary over time?’ As it is structured, it does not directly permit one to turn the question around in order to assess the impact of unobserved genes and environment on the systematic rate of change of a phenotype over time. This latter is one of the fundamental objectives of our own modelling. In their analysis of the Framingham Heart Study data in GAW13,14 Soler and Blangero52 also extend a conventional variance-components-based approach to linkage analysis to produce a flexible one-step approach for longitudinal family data. But their work again focuses on linkage analysis for continuous traits with a multivariate normal distribution. At the same workshop, Yang et al.53 and MacGregor et al.54 described random regression models which incorporated polynomial coefficients to model the covariance matrix as a smooth function of time.10 These models were again used to estimate age-specific changes in heritability, but not to estimate the heritability of the rate of change of a quantitative phenotype over time.

In contrast to the alternative methods, the approach based on MCMC that we describe extends easily both to non-normal problems and to complex pedigrees.8,23 Furthermore, our models are so structured that they lead naturally to inferences based on the genetic and environmental covariance components associated with the trait mean and with the rate of change of the trait over time. Among the wide variety of different approaches to analysing longitudinal family data that were explored in GAW13,10,12-14 including those referred to above, there is none that is designed to address these particular scientific questions that is both as flexible and as tractable as the GLMMs we describe. However, our models make a number of key assumptions that warrant further discussion.

First, they assume that, conditional on the fixed and random effects—and the genetically and environmentally determined correlation structure that they generate—the repeated observations in an individual are conditionally independent. Should it be viewed as desirable to introduce an additional autoregressive component to the model to account for a tendency for closely spaced observations in an individual to be more similar than measurements that are made at wider intervals3,9,52 the flexibility of the WinBUGS modelling environment is such that it would be straightforward to do so.29 However, such models would be susceptible to poor mixing. They would demand the running of a prolonged series of iterations, and this would come with a heavy computational overhead.

Second, they assume that the \( \beta_{ij} \) are independent of the \( \beta_{ij} \). But realistic scenarios can be imagined under which some determinants affect both the level of response and the rate of change of the response with time.10 Such determinants will introduce a covariance between the \( \beta_{ij} \) and \( \beta_{ij} \). In principle, this is straightforward to incorporate. For example, the additive effect of genes that affect both the intercept and the gradient can be included in model (2) by extending the model for the subject level random effects:

\[
\beta_{ij} = A_{ij} + C_{ij} + Q_{ij}
\]

where \( Q_{ij} \) is a random effect, corresponding to a variance component \( \sigma^2_{Q_{ij}} \), which reflects the additive genetic effect (on the intercept) that is attributable to genes that also affect the slope. The parameter \( \gamma \) is an unknown constant: it allows the average magnitude of the effect of the \( Q_{ij} \) on the intercept to differ from the corresponding effect on the gradient. The model implies a (genetically mediated) covariance of \( \gamma \times \sigma^2_{Q_{ij}} \) between \( \beta_{ij} \) and \( \beta_{ij} \). Posterior distributions for \( \gamma \) and \( \sigma^2_{Q_{ij}} \) are obtained from the MCMC analysis, as also are posterior distributions for the individual \( Q_{ij} \). Equivalent extensions can be implemented to incorporate covariances arising from shared environmental determinants that affect both the level of response and the rate of change of response over time. Although such extensions are straightforward to implement, they impair the mixing of the MCMC process—and

\[
\begin{align*}
\beta_{ij} &= A_{ij} + C_{ij} + Q_{ij} + \gamma \times Q_{ij} \\
\end{align*}
\]
models must therefore be run longer. Our website http://www.prw.le.ac.uk/research/HCG/gebugs.html contains a working example of a GLMM (longitudinal.covariance.b0.b7To.d0d) that models a simulated scenario that is precisely equivalent to that considered above. In this particular case, the response variable is normally distributed and the dataset consists of five stacked simulations, each involving 500 nuclear families. As simulated, all subjects are observed at five separate time-points; but 10% of subjects and 10% of observations are deleted at random. Analysis is based on a burn-in of 5000 iterations and an analysis of 50,000 iterations in each of three chains. Detailed results for all parameters may be obtained from the website. The simulated values for the four key parameters $\sigma^2_0$, $\gamma$, $\sigma^2_\epsilon$, and $\gamma$ were: 25, 5, 4, and 0.4, respectively. The corresponding estimates (95% credible intervals) from the MCMC analysis averaged across the five separate simulations were: 23.1 (13.2–31.5); 4.22 (2.56–5.73); 21.1 (12.7–31.0); and 0.48 (0.30–0.70). Results from each of the five individual simulations were very similar to one another. All the simulated target values fell within the 95% credible intervals of the estimates. It is clear from inspecting the history of the MCMC traces (see website) that the mixing of the traces for the key parameters was very slow and that an analysis run shorter than 50,000 iterations would have been inappropriate. We are considering ways to reparameterize the model to improve the mixing.

Third, the models described in this paper explicitly assume that families are ascertained randomly and although we have developed approaches to deal with non-random ascertainment when fitting MCMC-based GLMMs of this type, such models require careful interpretation. To date, randomly ascertained longitudinal family data have been unusual, but this is gradually changing. In addition to the large national initiatives referred to above, population-based studies in progress that include a longitudinal family component include the Framingham Heart Study, the Busselton Heart Health study and studies using The Netherlands Twin Register. Several other large longitudinal studies based primarily on the population-based recruitment of individuals are considering extensions to permit an enhanced ability to undertake family-based research. These include ALSPAC (http://www.alspac.bris.ac.uk/) and the proposed US National Children’s Study (http://nationalchildrensstudy.gov/). In addition, many cross-sectional family studies are likely to collect longitudinal data in a way that families are ascertained randomly and although we have developed approaches to deal with non-random ascertainment when fitting MCMC-based GLMMs of this type, such models require careful interpretation. To date, randomly ascertained longitudinal family data have been unusual, but this is gradually changing. In addition to the large national initiatives referred to above, population-based studies in progress that include a longitudinal family component include the Framingham Heart Study, the Busselton Heart Health study and studies using The Netherlands Twin Register. Several other large longitudinal studies based primarily on the population-based recruitment of individuals are considering extensions to permit an enhanced ability to undertake family-based research. These include ALSPAC (http://www.alspac.bris.ac.uk/) and the proposed US National Children’s Study (http://nationalchildrensstudy.gov/). In addition, many cross-sectional family studies are likely to collect longitudinal data in the future. These include the Western Australian Twin Child Health (WATCH) Study.

Fourth, the models in this paper all focus on the determinants of a linear trend in the value of a trait over time. But the method easily extends to more complex time trends. For example, if the linear predictor in model (2) is re-expressed in the form:

$$g(\mu_{ik}) = (\beta_0 + \beta_1 b_{ik}) + x'_{ik} \mathbf{B} + (\beta_2 + \beta_3 b_{ik}) \times U_{ik}$$

then $U_{ik}$ may be any function of time. This includes linear time ($T_{ik}$ itself), as a special case, but also allows for non-linear and cyclical functions too. This generalization allows for the possibility, for example, of studying the genetic and environmental determinants of diurnal and seasonal variations in a trait.

One of the greatest challenges in fitting extended GLMMs is the limited amount of relevant information that is intrinsic to the data. Binary phenotypes in particular may not always provide enough information to fit as complex a model as may be desired. But this is a fundamental feature of the underlying science rather than a problem with our particular approach itself. Nevertheless, this is why we illustrated the analysis of a binary phenotype in 500 nuclear families, while the equivalent simulations for a normal phenotype were able to be based on only 50 nuclear families (see comments in section 2.1). Furthermore, this emphasizes the scientific need for large sample sizes when one wishes to fit sophisticated models (particularly those with latent variables) to data relating to a complex disease. Another relevant limitation is the computer time required to fit the GLMMs. As an approximate guide, using a typical high end PC (Pentium 4 CPU 3 GHz; 1GB memory) 10,000 iterations of the model based on the simulated normal phenotype fitted to the extended pedigrees (1497 subjects in 23 pedigrees with phenotypic information available on 1000 subjects each with, on average, four longitudinal assessments) took 2–3 mins to compile and initialize and ~4 h to run. As computers become faster the processing speed will inevitably become a lesser concern.

In the light of all the alternatives, the particular approach to analysis that we describe in this paper offers a number of distinct benefits. (i) It makes full use of the information contained in the repeated observations for each individual, without first requiring the information to be summarized in any way. In other words, using the terminology adopted by Gauderman et al., it is a ‘joint’ rather than a ‘two-step’ method. (ii) It allows the inclusion of all pedigree members, without requiring decomposition of the pedigrees into smaller groups of relatives. (iii) Genetic effects can be modelled with complete flexibility. Conventional regression covariates may be used to reflect the direct effect of measured genotypes, as well as interactions between genotypes at different loci and/or between genotypes and measured environmental determinants. Variance components can be used to model and estimate the effect of unmeasured genes. It is then straightforward to add a segregation component involving an unobserved major gene. It is theoretically possible to incorporate genetic linkage by modifying the Mendelian segregation probabilities of the simple segregation model to reflect estimates of IBD sharing based on linked markers. As an alternative, linkage analysis can instead be based on the $A_{ij}$ and $A_{ij}$ random effects (see models (1) and (2)) as continuous phenotypes (already adjusted for confounders and non-genetic components of variance) that reflect the effect of additive genetic effects on the aetiology and progression of the phenotype of interest. Sampled values for each of these parameters are generated at every iteration of the MCMC process, thus permitting the construction of a posterior distribution for each $A_{ij}$ and $A_{ij}$. Linkage analysis can then be based on the posterior means of these parameters in each individual subject. From the specific perspective of a linkage analysis, this second approach is, of course, a ‘two-step’ model. However, the longitudinal and genetic elements of the full covariance model are fitted jointly, and from all other perspectives our approach is really a ‘one-step’ model. Furthermore, we are working to incorporate genetic linkage directly into the GLMMs. In practice, the theory is straightforward, but there are problematic restrictions imposed by the updating procedures in WinBUGS that have to...
be circumvented. (iv) The results obtained from our simulated data analyses suggest that our approach generates accurate point estimates with approximately nominal coverage of credible intervals. Furthermore, the analysis of the Framingham Heart Study data indicates that our extended variance components models appear to perform well in practice. Not only are the models relatively straightforward to fit, but they also permit inferences to be expressed in a manner that is intuitive from a bio-clinical perspective. In the case of the Framingham data, the results supported the existence of a large additive genetic component to SBP itself but a smaller (non-significant) additive genetic component to the rate of increase in SBP with age. There was strong evidence of important shared environmental effects. All this is consistent with the complex, multifactorial susceptibility to the development and natural history of hypertension that would be anticipated. (v) Precisely, the same methods can also be used in situations involving family-based data and a phenotype that is cross-sectional from an epidemiological perspective, but consists of many repeated measures. For example, 24 hour ambulatory blood pressure assessment generates multiple blood pressure measurements over a single day and a night. (vi) Randomly missing data can be accommodated with ease, and there is no need for balanced data; repeated observations may occur at irregular intervals that can vary from subject to subject. (vii) Formally, the GLMMs can easily be extended to the wide range of other distributions supported within the WinBUGS modelling environment. This means that the same models and methods that can be used to analyse the determinants of change in a normally distributed trait over time can also be used to analyse repeated evaluations of a binary trait or of a Poisson distributed rate or count, with minimal changes to the WinBUGS code. (viii) Parameter interpretation is straightforward, and the models may be fitted with readily available freeware (WinBUGS). (ix) The Bayesian approach provides flexibility and clear, intuitive answers, and incorporates uncertainty about each set of parameters into estimates of the uncertainty about every other set of parameters. While software programs such as ASREML or Proc Mixed in SAS may be used to efficiently analyse specific types of longitudinal data (e.g. normally distributed data) our method is more generally applicable and more flexible. (x) When fitted using Gibbs sampling in WinBUGS, the models we describe are highly tractable not only in the sense that they can be easily implemented in software that is freely available, but also in the sense that, because that particular software encodes a general purpose modelling environment widely used throughout the world, our methods could be easily used by research groups other than our own.

In conclusion, the GLMMs described in this paper do not solve all problems in the analysis of longitudinal family data and the models themselves require further study. For example, it is important to investigate the consequences of model misspecification. However, we concur with the sentiment: 'Commonly used genetic software programs are not designed for longitudinal data analysis and there is a clear need to develop integrated programs'. When scientific interest focuses on gene–phenotype associations and on formally decomposing the genetic and environmental covariance components for the intercept and for the slope of a trait, the models that we have described fulfil the need for an integrated approach, and appear to work very well. Furthermore, they provide a sound platform upon which other key developments can now be based.

**Acknowledgements**

The methodological research programme in Genetic Epidemiology at the University of Leicester is supported by MRC Cooperative Grant no. G9806740, by Program Grant no. 0033209 from the National Health and Medical Research Council of Australia, and by Leverhulme Research Interchange Grant no. F/07134/K. The University’s research programme in the genetics of hypertension is supported, in part, by a Wellcome Trust Thematic Programme Grant in Functional Genomics (no. 066780/Z/01/Z). M.D.T. has been funded by a Medical Research Council Training Fellowship in Health Services and Health of the Public Research. We thank the Framingham Heart Study investigators who kindly provided longitudinal data to GAW13. The Framingham Heart Study is supported by the NHLBI Framingham Heart Study Contract N01-HC-25195. We are grateful to Dr Jean McCluer and the organizers of GAW12 and GAW13. GAW is funded by NIGMS grant GM31575.

**KEY MESSAGES**

- Longitudinal family data provide unique opportunities to extend understanding of the aetiological architecture of complex diseases.
- They enable investigation of the genetic and environmental determinants of the progression of a disease, or a disease-related trait, over time.
- A class of generalized linear mixed models (GLMMs) that was originally developed for the analysis of cross-sectional data in families is extended to longitudinal family data.
- The GLMMs are straightforward to fit in the general purpose Markov chain Monte Carlo (MCMC) modelling software WinBUGS.
- The use of such models is illustrated on simulated data and on a real data problem based on the Framingham Heart Study, and a link is provided to working WinBUGS files on our web site.
References

Commentary: Models for longitudinal family data

W James Gauderman* and David V Conti

Cohort studies will become increasingly important in understanding the aetiology of complex human traits. While the longstanding approach of analysing cross-sectional data to identify genetic and/or environmental factors for disease or quantitative traits has resulted in some success, there have been many inconclusive results and far too few replications. There are recognized explanations that are often put forth for this, including low power and heterogeneity across study samples. However, a reason that is not often cited is that a single cross-sectional examination of data may not capture the essential aetiological mechanisms. For example, a specific variant genotype might cause an increase in a trait value that cumulates as a person ages. That is, a specific gene may affect the trajectory of the trait over time. Thus, two studies, one of young-aged subjects and the other of older-aged subjects, would likely come to different conclusions with respect to that locus owing to the different part of the gene–age trajectory that was examined. In a similar manner, different genes may also act at different time periods in the disease process, such as disease initiation or...