

Making predictions from hierarchical models for complex  
longitudinal data, with application to aneurysm growth

ANTHONY R. BRADY<sup>†</sup>

Intensive Care National Audit & Research Centre

Tavistock House, Tavistock Square, London WC1H 9HR, UK

SIMON G. THOMPSON

MRC Biostatistics Unit

Institute of Public Health, Robinson Way, Cambridge CB2 2SR, UK

<sup>†</sup> Correspondence to [tony@icnarc.org](mailto:tony@icnarc.org), tel 020 7388 2856, fax 020 7388 3759

Short-title: Hierarchical models for complex longitudinal data

Key words: hierarchical, longitudinal, multi-level, MCMC, prediction

Supported by a grant from the BUPA Foundation

25 June 2002

## Summary

Longitudinal measurements of markers of disease progression are commonly collected with a view to describing the course of disease, assessing the association with patient characteristics, or predicting when an individual patient may reach some critical threshold. Here we compare hierarchical models for modelling longitudinal marker data fit using maximum likelihood or Markov chain Monte Carlo (MCMC) simulation. Using MCMC, we explore the sensitivity of results to the usual assumptions of multivariate normality, and describe how to make predictions of the time taken to reach a threshold value, or the probability that a patient will exceed the threshold value within a given time period. The methods are illustrated by data from a longitudinal study of aortic aneurysm diameter in 1743 patients with two or more measurement occasions. Interest centres on population average growth, how growth varies between individuals, and predicting when a patient will exceed a diameter of 5.5cm (when surgical repair of the aneurysm is considered). Although model fit was greatly enhanced by using non-normal distributions for within and between subject random effects, both overall population growth and predictions of future outcomes from the models were very similar. Prediction of future aneurysm diameter was generally accurate and coverage of credible intervals close to nominal values. Prediction of time to threshold value was of little practical use because of the high variance associated with such estimates. We recommend estimation of the probability that a patient will exceed the threshold at some future time as a more useful summary measure. MCMC methods of fitting hierarchical models allow modelling assumptions to be easily explored and predictions of not just future outcomes, but functions of future outcomes such as time to threshold value, to be estimated whilst recognising all sources of uncertainty within the model.

# 1 Introduction

Monitoring the extent or severity of disease over time is usually undertaken to inform clinical decision making. For instance patients meeting some threshold level of severity may be offered new or more aggressive treatments. One example of this is for patients with abdominal aortic aneurysm in whom routine ultrasound monitoring of aneurysm diameter is undertaken [1]. Patients with aneurysm diameter  $>5.5\text{cm}$  are usually offered open surgical repair of their aneurysm, because otherwise their risk of aneurysm rupture and death is substantial [2]. It is useful for clinicians involved in such decision making to know what path the severity of disease might take in an average patient. The prediction may also be updated in the light of measurements on a specific patient. This information can be used both to plan the frequency of monitoring and to inform the patient of likely prognosis.

Multilevel models are often used to model longitudinal growth data due to its inherent hierarchical structure [3]. The availability of general purpose software for fitting multilevel models such as MLwin [4] and BUGS [5] has encouraged their use. Here we describe the multilevel modelling of a longitudinal measure of disease severity in both maximum likelihood and Markov chain Monte Carlo (MCMC) frameworks. The flexibility of the MCMC approach allows the impact of different modelling assumptions, such as random effect distributions, to be assessed. The extent to which model based predictions, such as predicting time to aneurysm diameter  $>5.5\text{cm}$ , and their associated uncertainty are sensitive to model specification is also investigated.

## 2 Data on aneurysm diameter over time

The dataset used to illustrate the different modelling approaches comes from the UK Small Aneurysm Trial and Study [6]. All aneurysm patients referred to vascular surgeons at 93 participating hospitals around the UK between August 1991 and November 1995 were eligible

for entry into the Study. Very few patients (N=15) refused to be part of the Study and at least one aneurysm measurement was recorded on 2366 patients. Study patients that met the Trial entry criteria at any time and consented to the Trial [7] were randomised to immediate surgery or surveillance (N=1090). Repeat measurements of aneurysm diameter were taken on Study patients, Trial patients before entry into the Trial, and Trial patients randomised to surveillance. Measurements were scheduled every 6 months for aneurysms less than 5.0cm in diameter and every 3 months for larger aneurysms. Patients were followed up for a mean of 1.4 years and were measured a mean of 4.1 times (range 1 to 24). Approximately one quarter of patients (386 Trial and 237 Study patients) were measured only once at baseline, and are omitted since they contribute no information on aneurysm growth. The distribution of baseline aneurysm diameter for all remaining 1743 patients is shown in Figure 1. There is noticeable left truncation to the distribution, since it is rare to diagnose an aneurysm in a patient with abdominal aortic diameter < 3cm [2], and a long right hand tail.

Observations of patients' aneurysm diameter were censored due to death (16%), surgery (37%) or end of scheduled follow-up (47%). Censoring by end of follow-up was more frequent in Study patients compared to the Trial patients because of resource constraints (Table 1). Aneurysm diameter measurements missing due to death or end of follow-up can arguably be considered as missing completely at random (MCAR) [8] since deaths from aneurysm related causes were a small proportion of total deaths[7] and end of follow-up was largely due to administrative reasons. The censoring due to surgery is an example of data missing at random (MAR) since it usually depended on a previously observed aneurysm diameter > 5.5cm. Likelihood based methods are unbiased when applied to data with a MAR mechanism, but other methods (such as marginal models estimated using generalised estimating equations) are not [9]. A plot of the longitudinal data for 25 randomly chosen individuals is shown in Figure 2. It is apparent that in fact not all patients went for surgery when their aneurysm measured 5.5cm

(for example because they refused or were considered unfit) and continued under surveillance.

### 3 Classical multilevel models

A standard set of repeated measures growth models were applied to the aneurysm data in a classical multilevel modelling framework [10]. These 2-level models follow the general form

$$Y_{ij} = \sum_{h=0}^p \beta_h x_{hij} + \sum_{h=0}^m U_{hi} z_{hij} + R_{ij}$$

where  $Y_{ij}$  is the outcome measure on the  $j$ th occasion ( $j = 1, \dots, n_i$ ) for the  $i$ th patient ( $i = 1, \dots, N$ ),  $x_h$  are explanatory variables in the fixed part of the model,  $z_h$  are explanatory variables in the random part, and intercept terms are created by setting  $x_{0ij} \equiv z_{0ij} \equiv 1$ . The  $z_h$  are usually a subset of the  $x_h$  variables, although this is not a requirement. The level-2 random effects ( $U_{0i}, \dots, U_{mi}$ ) are assumed to have a multivariate normal distribution with zero mean and constant covariance matrix  $\mathbf{\Omega}_2$ . The level-1 residuals  $R_{ij}$  are assumed to be independently normally distributed with zero mean and constant variance  $\sigma_e^2$ . All models were fitted using restricted maximum likelihood estimation (REML) in MLwin [4].

The variance components model contains only the intercept terms  $x_{0ij}$  and  $z_{0ij}$  determining  $\mathbf{\Omega}_2 = (\sigma_{u0}^2)$  (model 1, Table 2). The high intra-person correlation  $\hat{\rho} = 0.83$  ( $\rho = \sigma_{u0}^2 / (\sigma_{u0}^2 + \sigma_e^2)$ ) indicates that most of the variation in aneurysm measurements was due to differences between individuals rather than variation within individuals. There was a large decrease in deviance and  $\sigma_e^2$  when a fixed linear effect of time ( $x_{1ij} = t_{ij}$ , time of aneurysm measurement in years since referral to vascular surgeon) was added to the simple variance components model (model 2, Table 2) demonstrating the importance of this covariate. The residual intra-person correlation coefficient was even higher,  $\hat{\rho} = 0.93$ , than in the variance components model so that nearly all the residual variance was explained by differences between individuals.

Introduction of the linear time effect into the random part of the model ( $z_{1ij} = t_{ij}$ ,

$\mathbf{\Omega}_2 = \begin{pmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u10} & \sigma_{u1}^2 \end{pmatrix}$ ) provided considerable evidence that the linear growth rate varied from patient to patient (deviance difference of 3917 on 2 degrees of freedom comparing model 2

with model 3, Table 2). The estimated standard deviation of the slope random effects was  $\sqrt{0.045} = 0.21$  giving a wide 95% reference range for the linear growth rate of  $-.14$  to  $.70$  cm/year (assuming normality). A moderate positive correlation between intercept and slope random effects of  $0.46$  ( $\sigma_{u01}/(\sigma_{u0}\sigma_{u1})$ ) indicated that patients with larger aneurysms at presentation

tended to experience more rapid growth. Investigation of non-linearity was by introduction

of polynomial terms in  $t_{ij}$ . Setting  $x_{2ij} = z_{2ij} = t_{ij}^2$ ,  $\mathbf{\Omega}_2 = \begin{pmatrix} \sigma_{u0}^2 & \sigma_{u01} & \sigma_{u02} \\ \sigma_{u10} & \sigma_{u1}^2 & \sigma_{u12} \\ \sigma_{u20} & \sigma_{u21} & \sigma_{u2}^2 \end{pmatrix}$  provided

evidence of non-linearity in the growth curves (deviance difference of 500 on 4 degrees of freedom comparing model 3 with model 4, Table 2) with an average trend of accelerating growth.

Unfortunately both level-1 and level-2 residuals from these models displayed considerable non-normality. The level-1 residuals (Figure 3) had high kurtosis with more values in the tails of the distribution compared to a normal distribution. Inspection of some of the longer series with fitted curves from model 4 (Figure 4) showed that aneurysm growth sometimes apparently occurred in spurts. Fitting this step function with a smooth growth curve model leads to a symmetric distribution of residuals with heavier tails than a normal distribution. The level-2 residuals for the intercept (Figure 5 top) reflected the distribution in Figure 1. The level-2 residuals for the slopes (Figure 5 bottom) were reasonably normal apart from an over-abundance of large slopes.

## 4 Flexible hierarchical models

Hierarchical modelling estimated by MCMC allows the sensitivity of the model to distributional assumptions to be easily explored. This may be important here due to the non-normality

apparent in the level-1 and level-2 error distributions of the classical multilevel models fitted. Bayesian estimation of the models that follow was achieved by Gibbs sampling in WinBUGS [5]. Posterior distributions of selected nodes (parameters, functions of parameters or missing data) were monitored for 20,000 iterations following a burn-in of 5000 and were summarised by medians and standard deviations. Convergence was assessed by visual inspection of the Markov chain output. Model fit was compared using the Deviance Information Criteria (DIC), which is approximately equivalent to Akaike's information criteria in models with negligible prior information [11]. The effective number of parameters measure  $p_D$ , used to penalise the deviance in the calculation of DIC, is also reported since it gives an impression of model complexity.

Bayesian analysis requires specification of prior distributions for all unknown parameters. Due to the large amount of data available for analysis we wanted the priors to be non-informative. Therefore for model 3 ( $x_{1ij} = z_{1ij} = t_{ij}$ ) we used vague Normal priors  $N(0, 10^6)$  for  $\beta_0$  and  $\beta_1$ . A Wishart( $R, \rho$ ) prior was specified for  $[\mathbf{\Omega}_2]^{-1}$ , the population precision matrix of the regression coefficients. The degrees of freedom,  $\rho$ , was set at the minimum value 2 (the rank of  $\mathbf{\Omega}_2$ ) to represent vague prior knowledge [12]. The scale matrix  $R = \begin{pmatrix} 0.5 & 0.07 \\ 0.07 & 0.05 \end{pmatrix}$  represents our prior guess at the covariance matrix  $\mathbf{\Omega}_2$ , for which values were obtained from classical model 3. Finally the prior for  $\sigma_e^2$ , the variance of the level-1 residuals  $R_{ij}$ , was specified as inverse gamma( $10^{-3}, 10^{-3}$ ) so as to give approximately uniform support for  $\log(\sigma_e^2)$ .

Fitting model 3 in the Bayesian framework with these vague priors produced virtually identical results to those obtained in the classical paradigm (Table 3). The effective degrees of freedom  $p_D$  was estimated as 2796, 20% less than the 3486 degrees of freedom that would be required to estimate intercepts and slopes individually for each of the 1743 subjects in the dataset. This gives some impression of the degree of shrinkage introduced by assuming intercepts and slopes are drawn from a multivariate normal distribution.

In model 5 the assumption of a Normal distribution for the level-1 errors was replaced by a

$t$ -distribution to allow for the kurtosis observed in Figure 3. The  $t$ -distribution with  $\nu$  degrees of freedom is here modified by a scale parameter  $\tau$ , so that the variance is  $\sigma_e^2 = (\nu/(\nu - 2))/\tau$ . We specified a gamma( $10^{-3}, 10^{-3}$ ) prior for  $\tau$  and a uniform prior for  $\nu$  over the range (2.5, 1000). A uniform prior for  $\nu$  leads to a non log-concave full conditional distribution. WinBUGS uses a slice-sampling algorithm with an adaptive phase of 500 iterations for this type of density [5]. The posterior median of  $\nu$  was 3.0 (95% credible interval 2.8 to 3.3) suggesting that a  $t$ -distribution with extremely heavy tails was supported by the data. The DIC was reduced by 1488 (Table 3), evidence that the level-1  $t$ -distribution provided a much better fit to the data. The variance of the linear growth slopes ( $\sigma_{u1}^2$ ) was reduced by 20% as a result of the reduced influence of outlying aneurysm measurements. Specifying a uniform prior for  $1/\nu$  as suggested by Gelman and Meng [13] made very little difference to the posterior distribution of  $\nu$  from model 5.

All models so far have made the assumption of multivariate normality of random intercepts and slopes. This is undesirable in this context since the intercept, which represents the size of aneurysm at presentation, is not normally distributed (Figure 1). We prefer to treat the intercepts,  $\beta_{0i} = \beta_0 + U_{0i}$ , as free parameters with Uniform(2,11) prior distributions. The association between intercept  $\beta_{0i}$  and slope  $\beta_{1i} = \beta_1 + U_{1i}$  is represented by  $\lambda_1$  in the regression-type relation

$$\beta_{1i} = \lambda_1 (\beta_{0i} - 4.28) + \delta_i$$

where  $\delta_i \sim N(\mu_\delta, \sigma_\delta^2)$ . The  $-4.28$  is added simply for convenience, so that  $\mu_\delta$  is the mean slope when the intercept is 4.28 (the mean in previous models).  $\mu_\delta$  and  $\lambda_1$  were given vague Normal priors  $N(0, 10^6)$ , and the prior for  $\sigma_\delta^2$  was inverse gamma( $10^{-3}, 10^{-3}$ ).  $\sigma_{u0}^2$ ,  $\sigma_{u1}^2$  and  $\sigma_{u01}$  are no longer parameters of the model but can be estimated by monitoring the empirical variances and covariances of  $\beta_{0i}$  and  $\beta_{1i}$ . The posterior medians of these variance distributions are reported. Fitting this model slightly increased the estimated variance of the intercepts (model 6, Table 3), probably because the previous normal distribution assumption was restricting some of the larger

intercepts. The variance of the slopes and the correlation between intercepts and slopes were both reduced slightly in this model. A 95% reference range for linear growth rate under this model was -.10 to .61 cm/year, narrower than that obtained from classical model 3. The normal distribution assumption for the random effect component of the slopes,  $\delta_i$ , appears reasonable in this model (Figure 6).

Model 6 was extended to include a quadratic term in  $t_{ij}$  ( $x_{2ij} = z_{2ij} = t_{ij}^2$ ). Correlation between this quadratic term ( $\beta_{2i} = \beta_2 + U_{2i}$ ) and both intercept and slope was introduced by the relation

$$\beta_{2i} = \lambda_2 (\beta_{0i} - 4.28) + \lambda_3 \delta_i + \eta_i$$

where  $\eta_i \sim N(\mu_\eta, \sigma_\eta^2)$ .  $\mu_\eta$ ,  $\lambda_2$  and  $\lambda_3$  were given vague Normal priors  $N(0, 10^6)$ , and the prior for  $\sigma_\eta^2$  was inverse gamma( $10^{-3}, 10^{-3}$ ). Model fit assessed by DIC was improved considerably with the introduction of the quadratic terms, despite an increase in the effective number of parameters,  $p_D$ , of approximately 500 (model 7, Table 3).

## 5 Prediction

Although it is possible to make predictions of future outcome measurements  $Y_{ij}$  in the classical framework there are several limitations to this approach. Firstly, the impact of distributional assumptions about the level-2 random effects and level-1 residuals on model predictions cannot easily be assessed. Secondly, inference on predictions from the classical models would not take into account uncertainty in estimating the variance components, such as  $\sigma_{u0}^2$ ,  $\sigma_{u01}$ ,  $\sigma_{u1}^2$  and  $\sigma_e^2$ . Predictions from Bayesian modelling naturally incorporate this uncertainty and should lead to better coverage of nominal credible intervals. Thirdly, within the Bayesian framework it is simple to make predictions and inferences about other quantities of interest, such as time taken to reach a threshold value of  $Y$  or probability of exceeding the threshold within a given time. Such an exercise would not be at all straightforward in the classical setting.

## 5.1 $Y_{ij}$ at future time

To assess predictions made by the candidate models in Table 3, a new dataset was created of 500 randomly chosen patients whose measurement series were truncated at a random occasion,  $j_i^*$  ( $1 \leq j_i^* < n_i$ ). All aneurysm measurements taken after occasion  $j_i^*$  ( $Y_{ij}$  for  $j > j_i^*$ ) were set to missing. Each candidate model was applied to this dataset using the `cut()` function in WinBUGS [personal communication - David Spiegelhalter, MRC Biostatistics Unit, Cambridge, UK], which allows logical nodes (such as the missing outcomes) in one dataset to be derived from the model fitted to another dataset. Prediction mean squared error ( $PMSE$ ) was calculated from the real outcomes  $Y_{ij}$  observed after  $j_i^*$  and their predicted values  $\hat{Y}_{ij}$  from the candidate model being assessed:

$$PMSE = \frac{\sum_{i=1}^{500} \sum_{j>j_i^*} (Y_{ij} - \hat{Y}_{ij})^2}{\sum_{i=1}^{500} (n_i - j_i^*)}$$

Nominal 95% credible intervals were obtained from the 2.5 and 97.5 percentiles of the posterior distributions of  $\hat{Y}_{ij}$  for  $j > j_i^*$  (denoted  $\hat{Y}_{ij}^{2.5}$  and  $\hat{Y}_{ij}^{97.5}$ ). Coverage of these nominal intervals was assessed by calculating the proportion that contained the true realised value  $Y_{ij}$ :

$$Coverage = \frac{\sum_{i=1}^{500} \sum_{j>j_i^*} I(\hat{Y}_{ij}^{2.5} \leq Y_{ij} \leq \hat{Y}_{ij}^{97.5})}{\sum_{i=1}^{500} (n_i - j_i^*)}$$

where  $I(cond)$  denotes an indicator that is 1 if the expression  $cond$  is true and 0 otherwise.

The 500 patients chosen to enter the prediction dataset had a total of 2572 aneurysm diameter measurements, and 1310 of these were set to missing. The standard error of  $PMSE$  was in the region .006 to .008 and the standard error on the coverage was around 0.6%.  $PMSE$  was almost invariant to the choice of model (Table 3) although coverage of nominal 95% credible intervals was slightly improved by inclusion of a quadratic random effect (model 7). This was probably because the quadratic model reduced bias in predictions of late measurements (Figure 7).  $PMSE$  was observed to increase when further ahead predictions were attempted in all

models ( $j \gg j_i^*$ ) and when there were less measurements before truncation (small  $j_i^*$ ) - results not shown.

## 5.2 Time to threshold value of $Y_{ij}$

The predicted time taken to reach some threshold value of  $Y_{ij}$  was investigated by creating some fictional patients with different baseline outcomes. In the aneurysm example there is particular interest in the threshold value  $Y_{ij} = 5.5\text{cm}$  when a patient will be considered for surgical repair of their aneurysm. We created 6 fictional patients with baseline aneurysm measurements  $Y_{i1}$  of 4.0cm, 4.5cm and 5.0cm (Table 4). Three of the patients had a second measurement at 6 months - patients 4 and 5 are “fast growers” (growth of 0.5cm over 6 months) whilst patient 6 is a “slow grower” (no growth over 6 months). Using the  $\text{cut}()$  function mentioned above, the estimated time in years taken for the expected aneurysm diameter to reach 5.5cm,  $t_i^{5.5}$ , was calculated from the intercept and slope of the linear model 6 for each of these 6 patients:  $t_i^{5.5} = (5.5 - \beta_{0i}) / \beta_{1i}$ . Note that negative values of  $t_i^{5.5}$  refer to patients who will never reach 5.5cm (because their slope,  $\beta_{1i}$ , is negative) so that we interpret  $t_i^{5.5} < 0$  as  $t_i^{5.5} = \infty$ . The uncertainty associated with  $t_i^{5.5}$  can be evaluated by monitoring it during the MCMC estimation of model 6. The 95% credible interval was extremely wide for  $t_i^{5.5}$ , especially when  $Y_{i1}$ , and hence  $\beta_{1i}$  (through correlation with  $\beta_{0i}$ ), was smaller (Table 4). This is because of the sensitivity of  $t_i^{5.5}$  to values of  $\beta_{1i}$  around 0. Uncertainty was reduced with the availability of a second aneurysm measurement 6 months after baseline (Table 4), but the variability is still too great to make such estimates of practical value.

One alternative is to estimate the probability that each fictional patient will exceed the threshold value at a given future time,  $t^*$ . For the expected aneurysm size and threshold value 5.5cm, this probability is  $\Pr(\beta_{0i} + \beta_{1i}t^* > 5.5) = E[I(0 < t_i^{5.5} < t^*)]$  and can be estimated from the mean of the indicator function  $I(0 < t_i^{5.5} < t^*)$  evaluated from a MCMC run of model 6. The probability that a measurement at time  $t^*$  exceeds 5.5cm can be estimated by creating

a dummy measurement occasion at time  $t^*$  with  $Y$  set to missing. Predictions of this missing node can be monitored and the proportion exceeding 5.5 calculated. The first method evaluates the predicted mean of  $Y$  (corresponding to the ‘true’ aneurysm measurement exceeding 5.5cm) while the second evaluates the predicted value of  $Y$  (corresponding to the observed aneurysm measurement exceeding 5.5cm). Table 5 shows results for the predicted mean and value of  $Y_i$  for our 6 fictional patients under model 6. The probabilities of exceeding 5.5cm are only slightly closer to 0.5 when the predicted value, rather than the predicted mean, is used to monitor passing the threshold. This results from the small amount of measurement error ( $\sigma_e^2$ ); if  $\sigma_e^2$  were large, the probabilities for the predicted value would all be close to 0.5. Under model 6, the estimated proportion of patients with a true aneurysm measurement in excess of 5.5cm at 1 year was 0.21 for patients with a baseline diameter of 5.0cm (Patient 3, Table 5). However, the estimated proportion of such patients with an observed aneurysm measurement in excess of 5.5cm at 1 year was greater at 0.24. These proportions were higher under model 3, being 0.27 and 0.30 respectively, due to the larger estimate of  $\beta_1$  in model 3 compared to model 6 (Table 3). The sensitivity of the proportion exceeding a threshold value to model choice should be borne in mind when interpreting the results from these models.

## 6 Discussion

Hierarchical models are suitable for modelling repeated measurement data. However the assumption of normality for the within and between subject residuals inherent in most classical approaches to analysis may not be satisfied in practice. Adopting MCMC as a tool for estimation allows a flexibility that can address some of the deficiencies of such conventional modelling. In our case, the within subject residuals were better modelled by a heavy-tailed  $t$ -distribution than a normal distribution. Further, the distribution of between subject intercepts could be left unspecified while still allowing random slopes across individuals that were correlated with the

intercepts. This improved the normality of the random linear slopes. Attempting such modelling in a classical framework can be difficult, for example leading to problems of convergence [14]. While one would not expect such changes in modelling the distribution of residuals necessarily to affect the estimates of fixed and random parameters [10], there is a greater potential for effect on standard errors and on the predictions derived from the model.

Alternative approaches for non-normal continuous outcome data include transforming the data to approximate normality [15], or substituting normal scores [16]. One method for deriving growth curve norms combines these two approaches, by calculating normal scores after using a Box-Cox transformation whose parameters change smoothly as a function of time [17]. However with such data transformation approaches there is no guarantee that normality can be achieved at both within and between subject levels simultaneously. Moreover there are some advantages, in terms of interpretability, of models that relate directly to the outcome measured on its original scale.

Both the classical and MCMC models described in this paper could have been extensively elaborated. Covariate effects could have been modelled as either fixed terms, for example as interactions with the parameters  $\beta_0$  and  $\beta_1$ , or as random terms, for example allowing the between or within subject variances to depend on baseline covariates. Different models for curvature over time, such as fractional polynomials [18], could have been explored, and between and within subject variability allowed to be functions of time [10]. However, rather than pursue such elaborations here, our aim was to discuss models where the conventional assumptions of normality could be relaxed. Moreover, if a large number of possible complex models are considered, and even if those selected fit better by criteria of deviance or DIC, the problems of over-fitting becomes more acute [19] and the need to validate predictions on an external data set becomes greater [20].

We note that our use of MCMC in this paper is as an estimation tool, rather than being

essentially ‘Bayesian’, since the priors used were intended to be non-informative. Nevertheless one important advantage of Bayesian modelling pertains, that the uncertainty in predictions takes into account the full uncertainty in all parameters of the model. The same is not true of conventional classical modelling, where the estimated variance parameters are assumed known [14]. Although this limitation of conventional classical modelling can be overcome, for example by parametric bootstrapping or simulation [21], these methods are indirect and computationally time-consuming. Allowing for full parameter uncertainty will be of particular importance in small data sets where the variance parameters are imprecisely estimated.

MCMC methods enable predictions (with associated uncertainty) of non-standard quantities to be undertaken, such as of the time to reach a given threshold value. Also predictions can be made for hypothetical individuals with any number of measurements at arbitrary times. In our example the credible intervals for the predictions obtained had close to nominal coverage, although the prediction error as represented by the PMSE did not improve with model fit as assessed by the DIC. A similar situation was reported in another case study using classical modelling [14]. In our example of aneurysm growth, the times of reaching the threshold of interest were very imprecisely estimated. The difficulty in estimating predicted time to event for individuals has been noted previously in the context of survival times [22]. The probability of exceeding the threshold value at a given time was a more useful summary of future prognosis, especially for considering appropriate re-measurement intervals. For example, there is little value in re-screening patients with aneurysm diameter less than 4.5cm sooner than one year because the probability that their aneurysm will have reached the threshold value of 5.5cm is practically zero, even for fast growers. It should be noted that the time origin for the predictions from this data set is the time of referral to a vascular surgeon. Use of the predictions for other time origins would not necessarily be valid.

An entirely different approach, a multi-state model, might be thought a more direct and

convenient way of estimating the time until reaching a threshold value. For example, a set of states defined by ranges of aneurysm diameter could have been declared, including an ‘absorbing’ state for a diameter over 5.5cm. The transitions between adjacent states could be modelled directly using the observed aneurysm diameters, for example using a Markov assumption [23]. More appropriately, the presence of measurement error could be acknowledged using a ‘hidden’ Markov model for the true underlying states [24]. However, in either case there is some arbitrariness in dividing up a continuous outcome scale into separate categories or states. More importantly, as is shown by the presence of substantial between subject variability in both slopes and curvature, a simple Markov assumption whereby the rate of transition to subsequent states just depends on the last state rather than the history before that time would be inappropriate, and a more complex dependence would have to be modelled.

## Acknowledgements

We are grateful to Professor Janet Powell and Louise Brown for providing the aneurysm growth data, and to Dr David Spiegelhalter for advice. This work was funded by the BUPA Foundation.

## References

- [1] G. M. Grimshaw and J. M. Thompson. The abnormal aorta: a statistical definition and strategy for monitoring change. *European Journal of Vascular and Endovascular Surgery*, 10:95–100, 1995.
- [2] A. B. M. Wilmink and C. R. G. Quick. Epidemiology and potential for prevention of abdominal aortic aneurysm. *British Journal of Surgery*, 85:155–162, 1998.
- [3] L. M. Sullivan, K. A. Dukes, and E. Losina. An introduction to hierarchical linear modelling. *Statistics in Medicine*, 18:855–888, 1999.

- [4] J. Rasbash, W. Browne, H. Goldstein, M. Yang, I. Plewis, D. Draper, M. Healy, and G. Woodhouse. *A User's Guide to MLwiN, Version 2.0*. Institute of Education, London, 1999.
- [5] D. J. Spiegelhalter, A. Thomas, and N. G. Best. *WinBUGS version 1.3 user manual*. MRC Biostatistics Unit, 2000.
- [6] UK Small Aneurysm Trial Participants. The UK Small Aneurysm Trial: Design, methods and progress. *European Journal of Vascular and Endovascular Surgery*, 9:42–48, 1995.
- [7] UK Small Aneurysm Trial Participants. Mortality results for randomised controlled trial of early elective surgery or ultrasonographic surveillance for small abdominal aortic aneurysms. *Lancet*, 352:1649–1655, 1998.
- [8] D. B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.
- [9] P. J. Diggle, K. Y. Liang, and S. L. Zeger. *Analysis of longitudinal data*. Oxford University Press, Oxford, 1994.
- [10] H. Goldstein. *Multilevel statistical models*. Edward Arnold, London, 1995.
- [11] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van der Linde. Bayesian measures of model complexity and fit. 2002, in press.
- [12] D. J. Spiegelhalter, A. Thomas, N. G. Best, and W. R. Gilks. *BUGS 0.5 examples volume 2 (version ii)*. MRC Biostatistics Unit, 1996.
- [13] A. Gelman and X. Meng. *Model checking and model improvement*, pages 189–201. Chapman and Hall, London, 1996.
- [14] K. Tilling, J. A. C. Sterne, and C. D. A. Wolfe. Multilevel growth curve models with covariate effects: application to recovery after stroke. *Statistics in Medicine*, 20:685–704, 2001.

- [15] P. Royston. Calculation of unconditional and conditional reference intervals for foetal size and growth from longitudinal measurements. *Statistics in Medicine*, 14:1417–1436, 1995.
- [16] H. Pan and H. Goldstein. Multi-level models for longitudinal growth norms. *Statistics in Medicine*, 16:2665–2678, 1997.
- [17] T. J. Cole and P. J. Green. Smoothing reference centile curves: the lms method and penalized likelihood. *Statistics in Medicine*, 11:1305–1319, 1992.
- [18] P. Royston and D. G. Altman. Regression using fractional polynomials of continuous covariates : Parsimonious parametric modelling. *Applied Statistics*, 43:429–467, 1994.
- [19] F. E. Harrell, K. L. Lee, and D. B. Mark. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387, 1996.
- [20] D. G. Altman and P. R. Royston. What do we mean by validating a prognostic model? *Statistics in Medicine*, 19:453–473, 2000.
- [21] R. M. Turner, R. Z. Omar, M. Yang, H. Goldstein, and S. G. Thompson. A multilevel framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine*, 19:3417–3432, 2000.
- [22] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18:2529–2545, 1999.
- [23] R. Kay. A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics*, 42:855–865, 1986.

- [24] G. A. Satten and I. M. Longini. Markov chains with measurement error: estimating the ‘true’ course of a marker in the progression of human immunodeficiency virus diseases. *Applied Statistics*, 45:275–295, 1996.

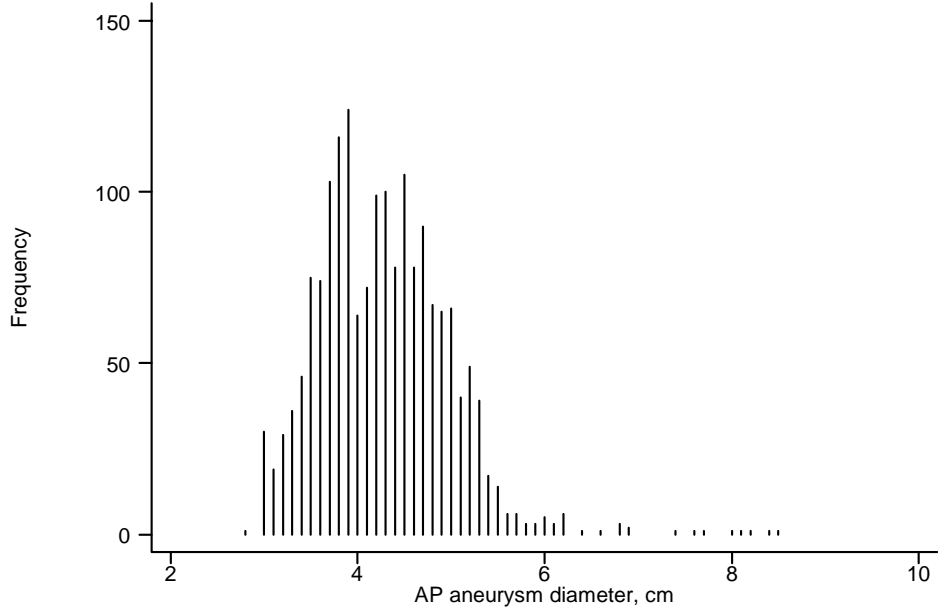


Figure 1: Distribution of baseline aneurysm diameter in 1743 patients

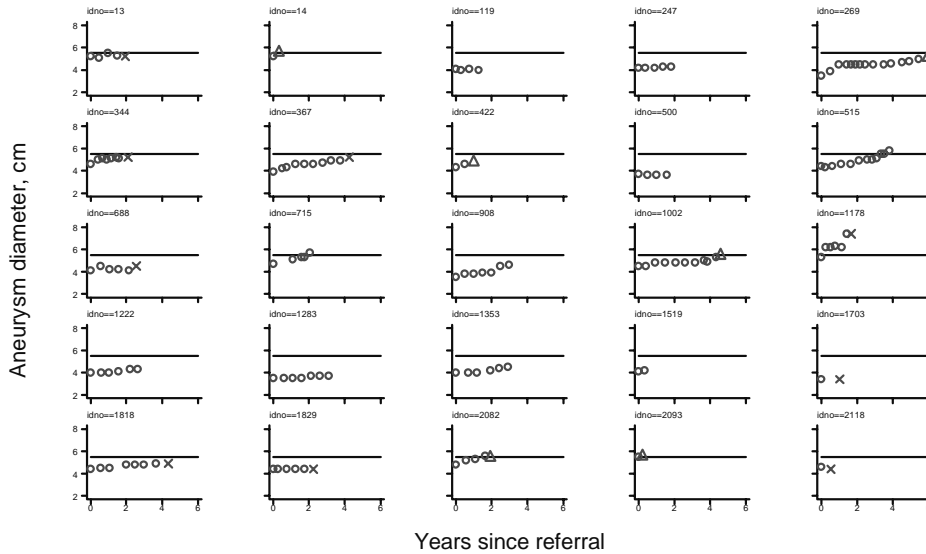


Figure 2: Aneurysm growth profiles of 25 randomly chosen patients ( $\Delta$  denotes series terminated by surgery,  $\times$  denotes series terminated by death). Horizontal line drawn at 5.5cm.

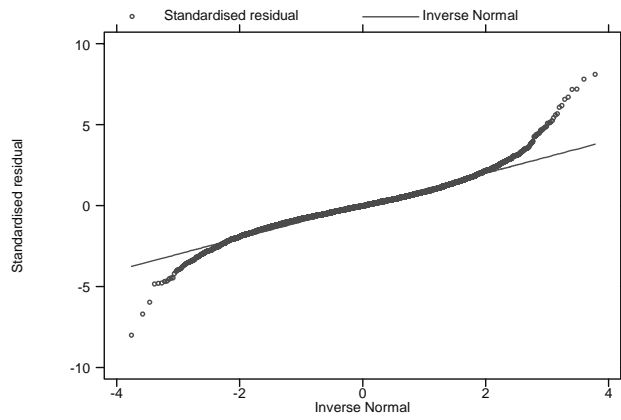


Figure 3: Normal plot of standardised level-1 residuals from classical multilevel model 4

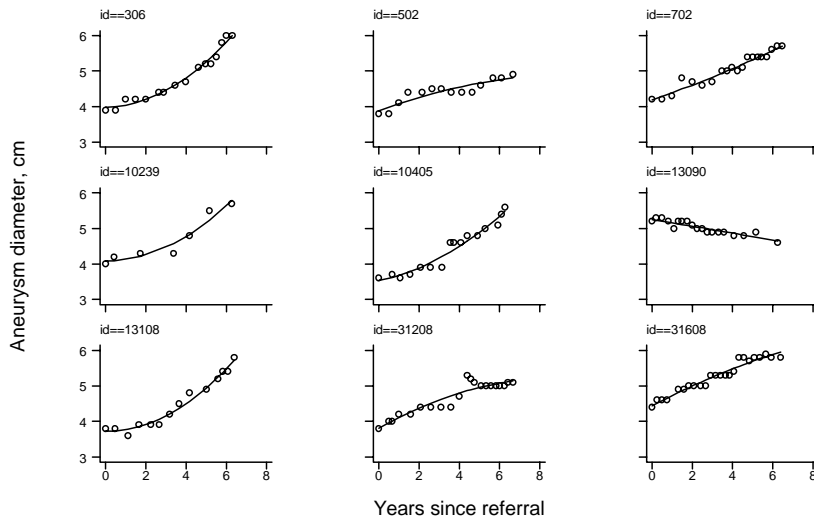


Figure 4: Observed and predicted aneurysm diameter for 9 patients with longest series

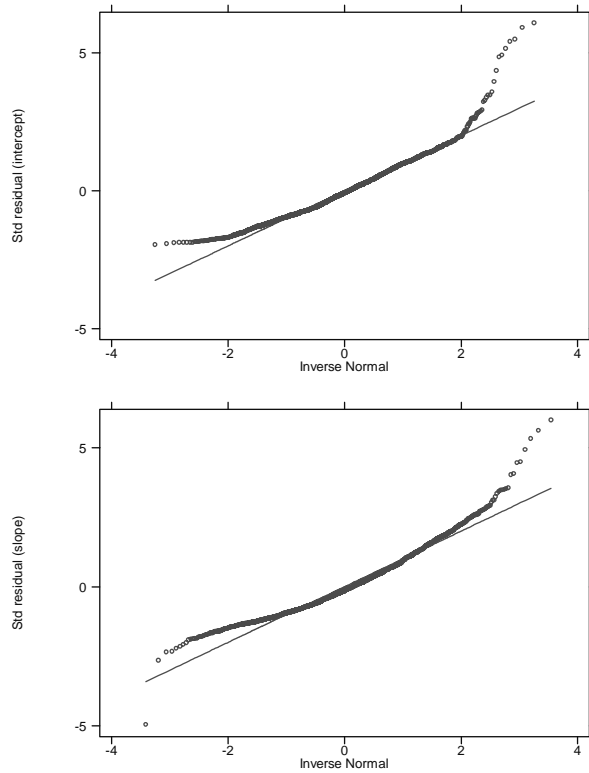


Figure 5: Normal plot of standardised level-2 residuals from classical multilevel model 4 for intercepts (top) and slopes (bottom)

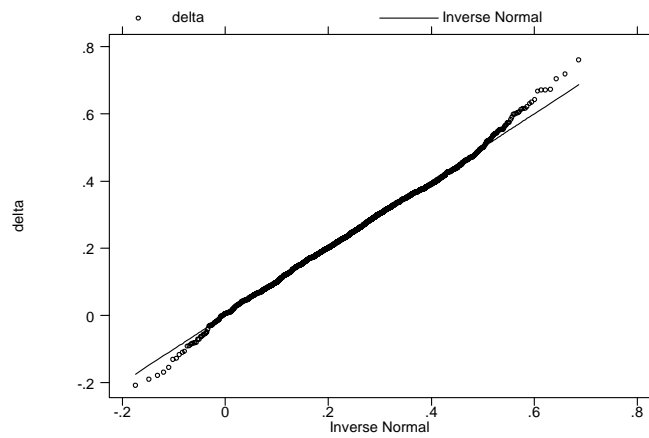


Figure 6: Normal plot of posterior medians of  $\delta_i$ , model 6

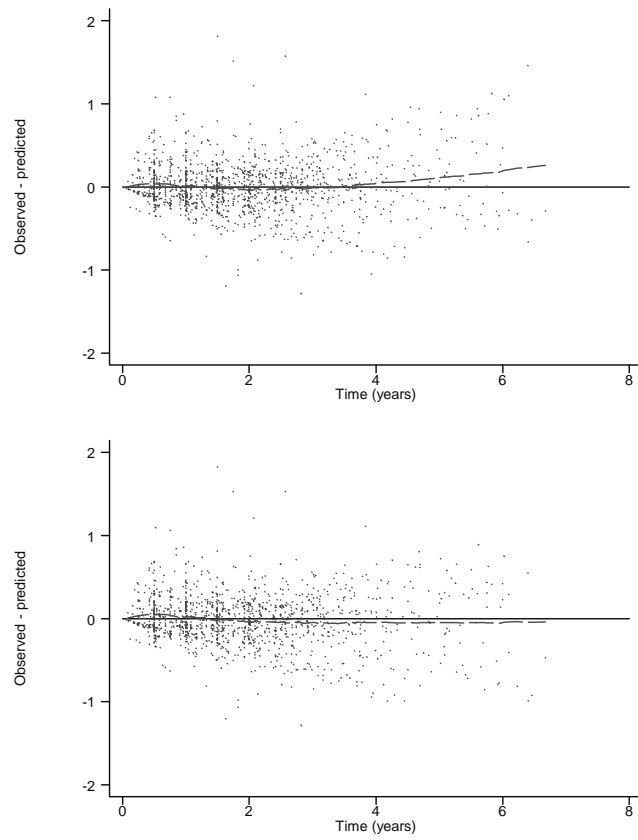


Figure 7: Prediction error for model 6 (top) and model 7 (bottom) by time, with Loess running-line smoother (dashed line)

Table 1: Summary of aneurysm diameter data collected during UK Small Aneurysm Trial and Study (patients with 2 or more aneurysm measurements only)

	<b>Trial</b>	<b>Study</b>	<b>Total</b>
Patients, N	704	1039	1743
Aneurysm diameter at baseline, cm, mean (SD)	4.4 (0.5)	4.2 (0.8)	4.3 (0.7)
Duration of follow-up, years, mean (SD)	2.2 (1.6)	1.7 (1.0)	1.9 (1.3)
Aneurysm measurements per patient, mean (SD)	6.4 (4.0)	4.4 (2.1)	5.2 (3.7)
Reason for termination of aneurysm measurement, N (%)			
Surgery	474 (67%)	174 (17%)	648 (37%)
Death	83 (12%)	200 (19%)	283 (16%)
End of follow-up	147 (21%)	665 (64%)	812 (47%)

Table 2: REML estimates (standard errors) from classical multilevel models

<b>Model</b> (see text)	<b>1</b>	<b>2</b>	<b>3</b>	<b>4*</b>
Fixed part				
$\beta_0$	4.51 (0.02)	4.29 (0.02)	4.27 (0.02)	4.28 (0.02)
$\beta_1$		0.23 (0.002)	0.28 (0.006)	0.26 (0.008)
$\beta_2$				0.0084 (0.002)
Random part				
$\sigma_{u0}^2$	0.53 (0.02)	0.59 (0.02)	0.48 (0.02)	0.48 (0.02)
$\sigma_{u1}^2$			0.045 (0.002)	0.073 (0.004)
$\sigma_{u01}$			0.068 (0.004)	0.070 (0.006)
$\sigma_{u2}^2$				0.0022 (0.0002)
$\sigma_e^2$	0.11 (0.002)	0.045 (0.0007)	0.019 (0.0004)	0.015 (0.0003)
Deviance	11046	4726	809	309

\* Estimates of covariances  $\sigma_{u02}$  and  $\sigma_{u12}$  omitted from table

Table 3: MCMC posterior medians (standard deviations) from Bayesian hierarchical models.

$\rho_{01}$  is correlation between  $\beta_0$  and  $\beta_1$ .

Model (see text)	<b>3</b>	<b>5</b>	<b>6</b>	<b>7*</b>
Fixed part				
$\beta_0$	4.27 (0.02)	4.28 (0.02)	4.28 (0.003)	4.29 (0.003)
$\beta_1$	0.28 (0.006)	0.26 (0.006)	0.26 (0.003)	0.23 (0.006)
$\beta_2$				0.011 (0.002)
Random part				
$\sigma_{u0}^2$	0.48 (0.02)	0.49 (0.02)	0.51 (0.004)	0.51 (0.004)
$\sigma_{u1}^2$	0.045 (0.002)	0.036 (0.002)	0.033 (0.001)	0.047 (0.003)
$\sigma_{u2}^2$				0.010 (0.001)
$\sigma_e^2$	0.019 (0.0004)	0.023 (0.002)	0.023 (0.0009)	0.020 (0.001)
$\rho_{01}$	0.46 (0.02)	0.46 (0.03)	0.40 (0.03)	0.29 (0.04)
$p_D$	2796	2978	2991	3488
DIC	-7536	-9024	-8981	-10158
$PMSE$	0.104	0.109	0.101	0.101
<i>Coverage</i>	93.3%	92.4%	92.4%	93.7%

\* Values of  $\sigma_{u02}$  and  $\sigma_{u12}$  omitted from table

Table 4: Estimated time (years) for expected aneurysm diameter to reach 5.5cm by aneurysm measurement at baseline and 6 months (results from model 6).

<b>Patient, <math>i</math></b>	$Y_{i1}$	$Y_{i2}$	$t_i^{5.5}$	<b>95% CI</b>
1	4.0		6.55	(2.58 to $\infty$ )
2	4.5		3.61	(1.52 to $\infty$ )
3	5.0		1.52	(0.50 to $\infty$ )
4	4.0	4.5	3.69	(2.03 to 39.83)
5	4.5	5.0	2.04	(1.20 to 9.90)
6	5.0	5.0	2.31	(1.04 to $\infty$ )

Table 5: Estimated probability of exceeding threshold aneurysm diameter 5.5cm by time since baseline (results from model 6).

<b>Patient, <math>i</math></b>	$Y_{i1}$	$Y_{i2}$	<b>Predicted mean</b>					<b>Predicted value</b>				
			<b>1 yr</b>	<b>2 yr</b>	<b>3 yr</b>	<b>4 yr</b>	<b>5 yr</b>	<b>1 yr</b>	<b>2 yr</b>	<b>3 yr</b>	<b>4 yr</b>	<b>5 yr</b>
1	4.0		0.00	0.00	0.06	0.20	0.34	0.00	0.01	0.07	0.20	0.34
2	4.5		0.00	0.11	0.38	0.56	0.67	0.01	0.13	0.38	0.56	0.67
3	5.0		0.21	0.66	0.82	0.88	0.91	0.24	0.66	0.81	0.87	0.90
4	4.0	4.5	0.00	0.02	0.30	0.57	0.72	0.00	0.03	0.31	0.57	0.71
5	4.5	5.0	0.00	0.48	0.79	0.89	0.93	0.02	0.48	0.78	0.88	0.93
6	5.0	5.0	0.02	0.39	0.66	0.77	0.83	0.05	0.40	0.65	0.77	0.82