STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:2547–2564
Published online 24 October 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2718

Infant growth modelling using a shape invariant model with random effects

Ken J. Beath*, †

Department of Statistics, Macquarie University, NSW 2109, Australia

SUMMARY

Models for infant growth have usually been based on parametric forms, commonly an exponential or similar model, which have been shown to fit poorly especially during the first year of life. An alternative approach is to use a non-parametric model, based on a shape invariant model (SIM), where a single function is transformed by shifting and scaling to fit each subject. In the model a regression spline is used as the function, with log transformation of the data and a simplification of the SIM, obtained from the relationship with the exponential model. All subjects are fitted as a nonlinear mixed effects model, allowing the variation in the parameters between subjects to be determined. Methods for the inclusion of covariates in growth models based on SIM are developed, with parameters for time independent covariates included in the model by varying either the shape, the size parameter or the growth parameter and time-dependent co-variates included by transforming the time axis, to either increase or decrease the growth rate dependent on the co-variate, similar to methods used for accelerated failure-time models. The model is used to fit weight data for 602 infants, measured from 0 to 2 years as part of the Childhood Asthma Prevention Study (CAPS) trial, and to determine the effect of breastfeeding on infant weight. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: shape invariant model; growth curve; nonlinear mixed effects; time-dependent covariates

1. INTRODUCTION

Early methods of analysing growth data were based on polynomials, as described in Wishart [1], which in many cases require higher order polynomials to provide adequate fit, with the resulting coefficients not having a physical interpretation. More recent models are based on parametric forms, a number of which have been described. Karlberg [2] describes a three phase model for growth, with the first phase of infancy lasting until age 3 or 4 years described by an exponential

*Correspondence to: Ken J. Beath, Department of Statistics, Macquarie University, NSW 2109, Australia.
†E-mail: kbeath@efs.mq.edu.au

Received 4 October 2005
Accepted 4 September 2006

model

$$Y = a_1 + b_1\{1 - \exp(-c_1 t)\} \tag{1}$$

A similar model, with an additional linear term, proposed by Jenss and Bayley [3] is

$$Y = a_0 + a_1 t - b_1\{1 - \exp(c_0 + c_1 t)\} \tag{2}$$

This was fitted to data by Berkey [4] and found to fit poorly, based on systematic variation of the residuals, especially in the first year of life. The model of Count [5]

$$Y = a + bt + c \log t \tag{3}$$

was also considered by Berkey and found to be inferior to the exponential model. However, this model was selected from several by Milani *et al*. [6] to construct growth curves for the first 3 years of life, but excluding measurements earlier than 3 months. Wingerd [7] investigated a number of functions for growth in the first 2 years of life and found that all fitted poorly, concluding that a model with more than three parameters was required. The main difficulty with the use of a parametric form is that it assumes that the underlying biological process to the growth may be represented by a simple functional form.

In contrast to the extensive literature on models for infant growth to describe the variation between individuals, less extensive are approaches to modelling variation in individual growth curves. A simplistic method is to use a polynomial with covariates either adding to the growth or as interactions with time, a method used by Geva *et al*. [8] for determining the effect of prenatal alcohol exposure. This has the disadvantage that polynomials are only appropriate for small data sets and the model does not seem biologically plausible. An alternative method is to compare the various subgroups to standard growth curves, by calculating the standardized difference between the actual and predicted growth and relating this to the covariates. This model was used by Ong *et al*. [9] for growth related to maternal smoking and other factors and by the European Collaborative Study [10] in children born to HIV-infected mothers. Another methodology for comparing groups is the transformed time scale method of Rao [11], which was applied by Wingerd [7] to determining factors in growth up to 2 years. A baseline group is chosen and the age scale transformed so that the growth lies along a straight line. Other groups are then fitted with the same transformed time scale and compared to the baseline group, either graphically or using regression techniques. The disadvantage of these models is that they are only applicable to categorical and time-independent covariates, and it is difficult to determine the way in which the covariates affect the growth in an easily interpretable way. For this we require a parametric model.

Parametric models for growth have a number of features in common with survival models. Covariates may be time independent, for example, sex or factors relating to parenting, where the same covariate is to apply over the period of observation; or time dependent, for example, dietary or environmental factors, where the covariate changes during the observation period. Inclusion of time-independent covariates can be accomplished by including them as part of the parameters of the model, by replacing a parameter in the model by a linear function of the covariates. This method was used by Berkey and Laird [12] for modelling the effect of gender and protein intake on the growth of children with a two stage process, first fitting a nonlinear function for each child and then modelling the effect of the covariates on the coefficients.

DOI: 10.1002/sim

Inclusion of time-dependent covariates is more difficult as we generally require two features:

1. The effects of covariates should be cumulative over time, that is, any relative increase or decrease in growth due to a change in a covariate should be maintained after the covariate returns to its original value. This relative change should also be a function of the duration of the covariate change.
2. In general, it is expected that growth will be monotonic. This is not necessarily true for some outcomes, for example, infant weight, where short-term loss in weight may occur, but over sufficiently long intervals this will be true.

Wang and Jackson [13], describe a method where the growth velocity is modelled as a function of time, where time-dependent covariates are included as a function multiplying the growth velocity. The growth velocity is then integrated to obtain the growth. In their application growth is allowed to decrease, but can be easily modified to allow only positive growth. An alternative method, used to model the infant growth data, is to transform the time axis, based on the value of the covariate during the period. In this way the growth is increased or decreased, while still resulting in increased growth during the period. This is similar to the method for survival analysis suggested by Cox and Oakes [14] for time-dependent covariates in the accelerated failure-time model. The main difference between the methods is in the way it affects the features of the curve. Modifying the growth velocity, maintains the features at the same time points, whereas transforming the time axis will shift the features. Provided the effect of the covariates is small, the two methods will produce similar results.

2. DATA

Data for the analysis were collected during the CAPS (Mihrshahi *et al.*) [15]. This study was designed to measure the effect of two interventions, dietary supplementation with omega-3 fatty acids and dust mite reduction, to reduce the incidence of atopy and asthma in infants at high risk, based on family history, of developing asthma. Babies of low birth weight and poor health were excluded from the study.

Pregnant women were recruited to the study from ante-natal clinics. Demographic data were collected 4 weeks before birth, with infant data collected at 1 month, then quarterly for the first year and then twice yearly. At birth and each visit, for a maximum of 8 visits, the infant's weight, height and head circumference were recorded along with a series of questionnaires on the home environment and health.

Subjects in the trial were recruited from Westmead and Liverpool Hospitals, Sydney, Australia with recruitment from September 1997 to December 1999.

Data were available for the first two years of the study. Of the 616 women recruited there were 602 subjects eligible for the study, with 599 having birth weight available. At the end of the two years, 544 subjects remained with weight data available for 528. One observation was deleted due to errors in the data collection. Out of the 602 subjects 302 (50.2 per cent) were male and 300 (49.8 per cent) female with mother's birthplace in Australia or New Zealand 444 (73.8 per cent), Asia 73 (12.1 per cent), other 84 (14.0 per cent) and unknown 1 (0.2 per cent).

3. MODEL

The model used is based on a shape invariant model (SIM) incorporating a function consisting of a natural cubic spline, which is fitted as a non linear mixed effects model. Based on the relationship

DOI: 10.1002/sim

between the SIM and exponential model, a log transform is used. The following sections define each component of the model.

3.1. Natural cubic splines

A spline is a piecewise polynomial, defined by the location of the knots joining each polynomial and the degree n of the polynomials, with the spline and its first $n-1$ derivatives constrained to be continuous at each knot. The splines used in this analysis are cubic ($n=3$), which are used to give a smooth shape to the function. Natural or restricted splines are further constrained to be linear beyond the boundary knots, which allows fitting of data points beyond the boundaries. The use of natural cubic splines has been examined for modelling data in Durrleman and Simon [16] and for longitudinal data in Beacon *et al*. [17].

3.2. Shape invariant models

The concept of SIM was first introduced by Lawton *et al*. [18]. The basis of the model is that a population has a common characteristic curve or function, which by shifting and scaling can be made to have the form of any individual curve. The mathematical form of a SIM is

$$f(\alpha_{0i}, \alpha_{1i}, \beta_{0i}, \beta_{1i}; t) = \alpha_{0i} + \mathrm{e}^{\alpha_{1i}} g\left(\frac{t-\beta_{0i}}{\mathrm{e}^{\beta_{1i}}}\right) \tag{4}$$

where $g(t)$ represents the characteristic shape of the response curve and $\alpha_{0i}, \alpha_{1i}, \beta_{0i}, \beta_{1i}$ are the parameters for the SIM. Parameters β_{0i} and α_{0i} are the shift parameters for the x- and y-axis, respectively, and β_{1i} and α_{1i} are the corresponding scale parameters. Exponentiation is used to constrain the scale parameters to be greater than zero, preventing problems with identification of these parameters and the function.

The function $g(t)$ may either be a parametric or non-parametric function. Lawton *et al*. [18] proposed and used a linear spline model to fit models for spirometry, spectrophotometric and optical density data. The modelling process involved simultaneous fitting of the response curve and separate parameters for each individual. It was suggested in the paper that the linear spline could be replaced by a cubic spline and this is the approach taken for this analysis.

Shape invariant modelling was first applied to human growth data by Stützle *et al*. [19]. The method used was to fit a model consisting of a nonlinear function plus a spline function to correct for the errors in the parametric form. Gasser *et al*. [20] fitted curves for pubertal growth using a SIM and kernel smoothing.

A difficulty of using a spline for SIM is that a spline is only specified over a limited range, usually the range of the data. However, for a SIM the shifting and scaling transforms the data beyond the bounds of the original data. Allowing the bounds to change during the fitting algorithm will usually cause the fitting algorithm to fail. Lindstrom [21] used a modified version of the `nlme` package [22] from the `R` language [23] to perform the fitting, with the modification allowing the boundary and internal knots to change position only at the start of each iteration. Altman and Villarreal [24] used an iterative algorithm consisting of a linear mixed effects step to estimate the spline function using penalized splines, followed by a nonlinear mixed effects step to estimate the shift and scale parameters. The inverse shift and scale operation is then performed to obtain the shape of the function and the process repeated until convergence is achieved.

DOI: 10.1002/sim

More recent methods of fitting SIM are based around smoothing splines (Ke and Wang [25]), where the spline is defined by a large number of knots and penalizing fits that are not smooth within the fitting algorithm. The model has the disadvantage of larger computation time, and was not able to be fitted to the infant growth data. Wang *et al*. [26] extended the model to fit periodic functions. An extended model is described by Brumback and Lindstrom [27] where the SIM transformation for the time axis is replaced by a more general smooth transform.

The method used for the infant data is fixed boundary and internal knots, with the boundary knots chosen to be slightly outside the range of the data. The boundaries can then be expanded if required to cover the range of the transformed data. By the use of natural splines, when the transformed data extend beyond the boundaries of the spline then a linear relationship is fitted. The internal knots were located at the defined time points for the visits, and individual knots were then removed to produce a more parsimonious model. Attempts to reduce the number of knots below the final number resulted in convergence failure.

A simplification of the SIM for infant data can be obtained by noting that the exponential model (1) and its log transform are parametric forms of a SIM. Starting with the exponential model, the form of the log transformed outcome can be found as:

$$Y = \exp(\alpha_0) * (1 - \exp(-\exp(-\beta_1) * (t - \beta_0))) \tag{5}$$

$$\begin{aligned} \log Y &= \alpha_0 + \log(1 - \exp(-\exp(-\beta_1) * (t - \beta_0))) \\ &= \alpha_0 + g\left(\frac{t - \beta_0}{e^{\beta_1}}\right) \end{aligned} \tag{6}$$

Where

$$g(x) = \log(-\exp(-x))$$

This is a form of SIM except that α_1 is set to zero, corresponding to no scaling of $\log Y$. The transformation of $\log Y$ is simply a shift, corresponding to a multiplication of Y by $\exp(\alpha_0)$. For individuals with the same β_1 this corresponds to a constant relative difference in growth over time. The parametric function $g(x)$ is replaced by a natural cubic spline to complete the model. This model has the advantage for fitting because it is closer to a linear model, and the log transform stabilizes the variance and reduces the positive skewness (Cole and Green [28]) that commonly occurs at older ages. As an alternative to the log transform a Box–Cox transformation [28, 29] may be suitable. Negative skewness may also occur at younger ages, especially for head circumference, which can be explained by the variation in alignment of the growth curve relative to the time origin. For each infant, birth does not necessarily occur at the same point in the growth curve, resulting in greater variation in the younger ages, which the β_0 parameter in the SIM allows for.

3.3. Nonlinear mixed-effect models

During early research on growth curves the methodology was to fit a curve to each subject separately and the parameter estimates were then summarized for the population. This has a disadvantage that each subject may be observed for only a small number of times, resulting in large variability of

DOI: 10.1002/sim

the parameter estimates. Lawton *et al*. [18] applied a similar method to the SIM, fitting a common curve but with individual transformation parameters. More recent investigations of growth using parametric nonlinear equations, for example, Susman *et al*. [30] have used a nonlinear mixed-effect model. This allows for parameters which are fixed or constant across a population and random effects which vary across a population, and are described by normal distributions which describe the variability among subjects. For the infant growth data, the proposed model has the coefficients of the spline function as fixed effects, resulting in a single spline function for all subjects; and the parameters of the SIM (α_0, β_0 and β_1) as random effects.

The nonlinear mixed effects model is described in Pinheiro and Bates [22]. From the SIM model (6) the corresponding nonlinear effects model is

$$\log\, y_{ij} = \alpha_{0i} + g\left(\frac{t_j - \beta_{0i}}{e^{\beta_{1i}}}\right) + \varepsilon_{ij} \tag{7}$$

where

$$\begin{pmatrix} \alpha_{0i} \\ \beta_{0i} \\ \beta_{1i} \end{pmatrix} \sim \mathrm{N}(0, \Psi)$$

and

$$\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$$

From the nonlinear mixed-effect model predicted curves for subjects and population may be obtained. For population prediction each of the random effects is set to zero and the predicted curve is obtained from the fixed effects. For individual predicted curves a best linear unbiased prediction (BLUP) of each of the random effects for a subject is obtained and used to predict the individual curve. The effect of BLUP is to shrink the predictions towards zero, which produces better predictions [31] than using the individual parameter estimates from each subject.

Inference for random effects involve a boundary condition, as we are testing for the variance of the random effect to be zero, and so cannot be performed using likelihood ratio tests. There are several methods available, of which the Akaike information criterion (AIC) (Sakamoto *et al.* [32]) was chosen, as it is simple and fast to compute compared to other methods, such as simulation. Models were fitted as nonlinear mixed effects models using the `nlme` package with the `R` language. The `nlme` package uses an approximation described by Lindstrom and Bates [33], to fit the nonlinear mixed effects algorithm, avoiding integration of the marginal likelihood. This has shown to be an efficient and accurate method by Pinheiro and Bates [34].

3.4. Growth velocity

As well as the growth, the growth velocity is often of interest. This may be obtained by direct differentiation of the spline function or by numerical differentiation. Numerical differentiation used as implementation is easier and produces accurate results, provided the interval between function evaluations is sufficiently small.

DOI: 10.1002/sim

4. COVARIATE MODELS

4.1. Time-independent covariates

For a SIM there are two ways that a covariate can be incorporated into the model, either through the parameters of the SIM or by modifying the shape function. The usual method will be to modify the parameters of the SIM, as this is the underlying foundation of the model, that all subjects have the same functional form that is shifted and scaled for an individual. The size α_0 of a subject can be modified by allowing the size parameter to be the sum of a random effect and a linear function of the covariates. The growth rate β_1 could be similarly replaced, however, it is simpler to introduce a parameter to modify the time directly. For a single covariate x_i with parameter α for the effect of the covariate on size and β for the effect on growth rate the model is

$$\log Y_{ij} = \alpha_{0i} + \alpha x_i + g\left(\frac{e^{\beta x_i} t_j - \beta_{0i}}{e^{\beta_{1i}}}\right) + \varepsilon_{ij} \tag{8}$$

The parameters can be transformed by exponentiation to a form that is easily interpretable. The effects of a unit increase in covariate x_i on the subject's size and growth rate are e^{α} and e^{β}, respectively. Based on biological considerations, covariates can be chosen to affect either the size or the growth rate of a subject or can be included for both. Alternatively, a categorical covariate may be included by using a different shape function for each level, in which case it would not be included as a parameter for either the size or growth rate. The model for multiple covariates is

$$\log Y_{ij} = \alpha_{0i} + \boldsymbol{\alpha}_i^T x_{\alpha i} + g\left(\frac{\eta_i t_j - \beta_{0i}}{e^{\beta_{1i}}}\right) + \varepsilon_{ij} \tag{9}$$

where $\boldsymbol{\alpha}_i$ is the parameters affecting the size for subject i, $x_{\alpha i}$ the covariates relating to $\boldsymbol{\alpha}$ for subject i, η_i the $e^{\boldsymbol{\beta}^T x_{\beta i}}$, $\boldsymbol{\beta}$ the parameters affecting the rate of growth and $x_{\beta i}$ the covariates relating to $\boldsymbol{\beta}$ for subject i.

4.2. Time-dependent covariates

We assume that a covariate such as breastfeeding will have an effect on the subject's growth rate only related to its current value. However, the increased or decreased growth rate during the period will produce an increase or decrease in size which will persist for the remainder of the subjects' life. To achieve a change in growth rate, the time scale is varied in response to the covariate. As this is cumulative, the time will be related to the cumulative effect of previous covariate values. A covariate is assumed to have a constant value within each interval. While continuous covariates may be used, the computations are simplified when values are binary. For the SIM the growth at time t_{ij} at the jth visit for subject i is given by

$$\log Y_{ij} = \alpha_{0i} + g\left(\frac{t_{ij} - \beta_{0i}}{e^{\beta_{1i}}}\right) + \varepsilon_{ij} \tag{10}$$

Given the covariate x_{ik} corresponding to the intervals $k = 1, \ldots, j$ for subject i and the parameter γ we can determine the growth as

$$\log Y_{ij} = \alpha_{0i} + g\left(\frac{\sum_{k=1}^{j} e^{\gamma(x_{ik} - \bar{x})}(t_{ik} - t_{i(k-1)}) - \beta_{0i}}{e^{\beta_{1i}}}\right) + \varepsilon_{ij} \tag{11}$$

DOI: 10.1002/sim

where $\bar{x}$ is the mean of x_{ik} over i and k. This is required for improved numerical stability in the fitting algorithm. This may be fitted using standard software, however, requiring that the summation be calculated at each function evaluation. Covariates are often categorical, for which a simpler form is available. For the simplest case of a binary variable then x_{ik} can be defined to have the values of either 0 or 1. Define $T_{0ij} = \sum_{k:x_{ik}=0} (t_{ik} - t_{i(k-1)})$ as the cumulative time to t_{ij} when x_{ik} is equal to 0 and $T_{1ij} = \sum_{k:x_{ik}=1} (t_{ik} - t_{i(k-1)})$ as the corresponding cumulative time when x_{ik} is equal to 1. We can then group the terms corresponding to x_{ik} equal to 0 and 1. Therefore

$$\sum_{k=1}^{j} e^{\gamma(x_{ik}-\bar{x})}(t_{ik} - t_{i(k-1)}) = \sum_{k:x_{ik}=0} e^{\gamma(x_{ik}-\bar{x})}(t_{ik} - t_{i(k-1)}) + \sum_{k:x_{ik}=1} e^{\gamma(x_{ik}-\bar{x})}(t_{ik} - t_{i(k-1)})$$

$$= e^{-\gamma\bar{x}} T_{0ij} + e^{\gamma(1-\bar{x})} T_{1ij}$$

and

$$\log Y_{ij} = \alpha_{0i} + g\left(\frac{e^{-\gamma\bar{x}} T_{0ij} + e^{\gamma(1-\bar{x})} T_{1ij} - \beta_{0i}}{e^{\beta_{1i}}}\right) + \varepsilon_{ij} \tag{12}$$

This requires for each time point only the cumulative time for which the covariate has each of its values. This technique can be easily extended to covariates with more than 2 levels and to combinations of covariates.

For r binary time-dependent covariates, there are $R = 2^r$ possible patterns of covariates, with a cumulative time corresponding to each during which the covariate combination applies. Each time period is then transformed dependent on the covariates, and the effective time experienced by a subject calculated as the sum of the transformed time. The transformed time for the effect of the time-dependent covariates alone is

$$t_{\gamma i} = (e^{X_{\gamma i}\gamma})^{\mathrm{T}} \mathbf{T}_i \tag{13}$$

where γ is the parameter affecting the rate of growth for time-dependent parameters, $\mathbf{T}_i$ the vector of length R describing the total cumulative time for each covariate combination and, $X_{\gamma i}$ the $R \times r$ design matrix.

4.3. Complete model

The effect of time-dependent and time independent covariates can be incorporated into a model. For the size of each subject, this is the sum of the effects of each covariate on size plus an additional random effect.

The transformed time for subject i at time j will be determined by both the time-dependent and time independent covariates as

$$t'_{ij} = e^{\boldsymbol{\beta}^{\mathrm{T}} x_{\beta i}} (e^{X_{\gamma i}\boldsymbol{\gamma}})^{\mathrm{T}} \boldsymbol{T}_{ij} \tag{14}$$

The model for all covariates, with α_{0i}, β_{0i}, β_{1i} as previously defined is

$$\log Y_{ij} = \alpha_{0i} + \boldsymbol{\alpha}^{\mathrm{T}} x_{\alpha i} + g\left(\frac{t'_{ij} - \beta_{0i}}{e^{\beta_{1i}}}\right) + \varepsilon_{ij} \tag{15}$$

DOI: 10.1002/sim

5. MODEL FITTING

Models were fitted as nonlinear mixed effects models using the `nlme` [22] package and the `R` language [23]. Prior to fitting the models, the covariates were centred to have zero mean, and initial parameter estimates for the fixed effects obtained by first fitting a nonlinear least-squares model using `nls` [23]. This was found necessary to produce correct fitting. After fitting, profile likelihood plots were produced to verify that the algorithm had found the correct maximum. Example code for fitting the models is contained in the Appendix.

6. RESULTS

In this paper the results for weight are presented. Models for height and head circumference were also fitted and gave similar results, with only the model for weight requiring a random effect for β_1. The model was fitted with boundary knots at -30 and 1200 days and internal knots at 30, 90, 180 and 540 days, corresponding to the time points at which visits were to occur. Raw data are plotted for all patients in Figure 1, with five randomly selected subjects shown with connecting lines, showing the appropriateness of the log transform for weights.

6.1. Males and females separately

Models were fitted for each random effect singly, each combination of two random effects and all three random effects and the AIC obtained from `nlme` are shown in Table I. Based on the AIC an appropriate model for both the male and female subjects will contain all three random effects.

Plots of residuals against age and normal quantile plots for the fitted model are shown in Figure 2. The plot of residuals *versus* age shows an improved fit compared to those for an exponential model (1), shown in Figure 3, but with severe kurtosis. Variograms [35] were used to check for autocorrelation between observations, with no adjustment for autocorrelation being required.

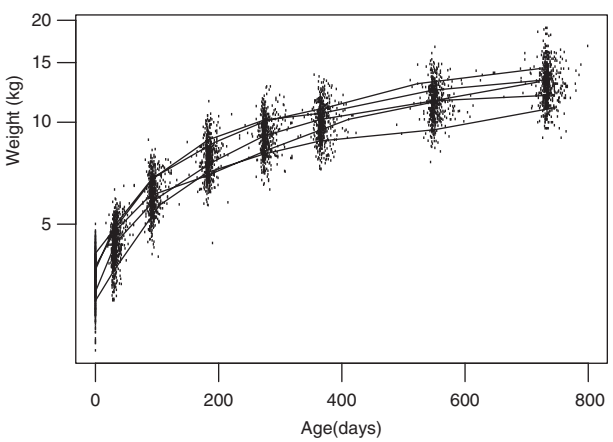

Figure 1. Weight *versus* age.

DOI: 10.1002/sim

Table I. AIC for random effects models.

Random effects	Male	Female
None	−794.1	−759.0
α_0	−4552.2	−4735.6
β_0	−3558.6	−3772.3
β_1	−4098.2	−4445.7
β_0, β_1	−4997.4	−5248.8
α_0, β_0	−5192.6	−5403.7
α_0, β_1	−4928.9	−5249.2
$\alpha_0, \beta_0, \beta_1$	−5452.2	−5635.5

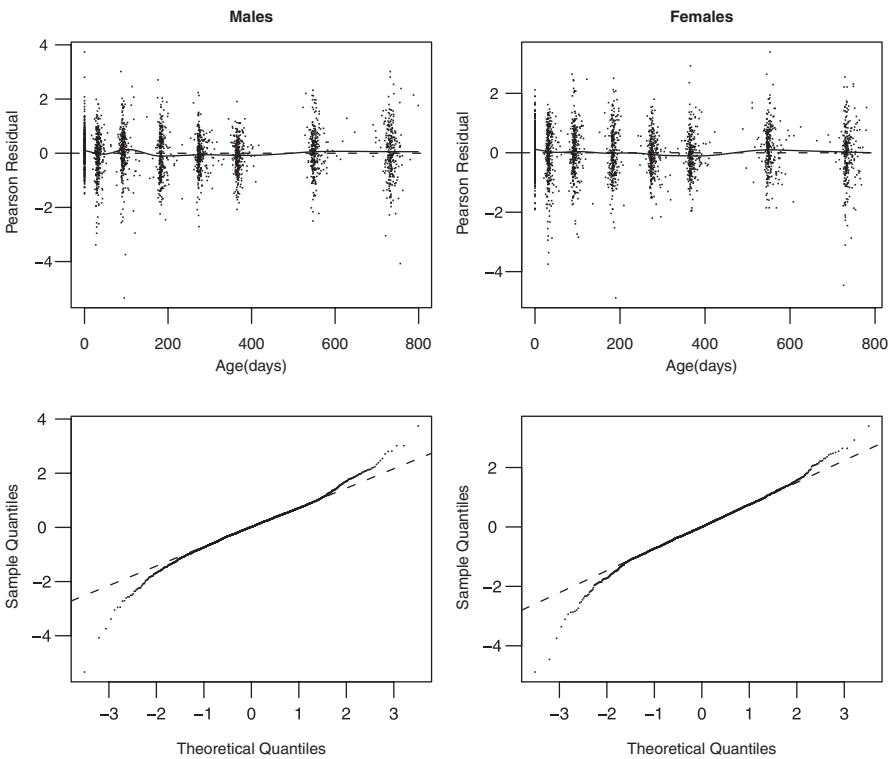

Figure 2. Residual plots for SIM model.

For 12 randomly selected subjects the fitted curve and observed data are shown in Figure 4. Only minor deviations of the fitted curve from the observed data are present, with possible reasons including measurement difficulties, variation in weight due to feeding and short-term illness.

Table II gives the standard deviations of the random effects for each gender. Variation in females appears to be greater, especially for β_1.

Population plots are obtained by setting each of the random effects to zero, and are shown for weight and the derived weight velocity in Figure 5. As expected males are heavier than females

DOI: 10.1002/sim

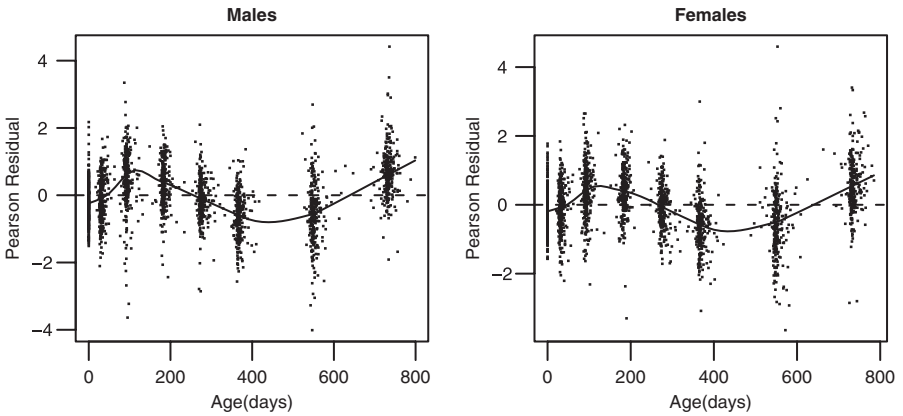

Figure 3. Residual plots for exponential model.

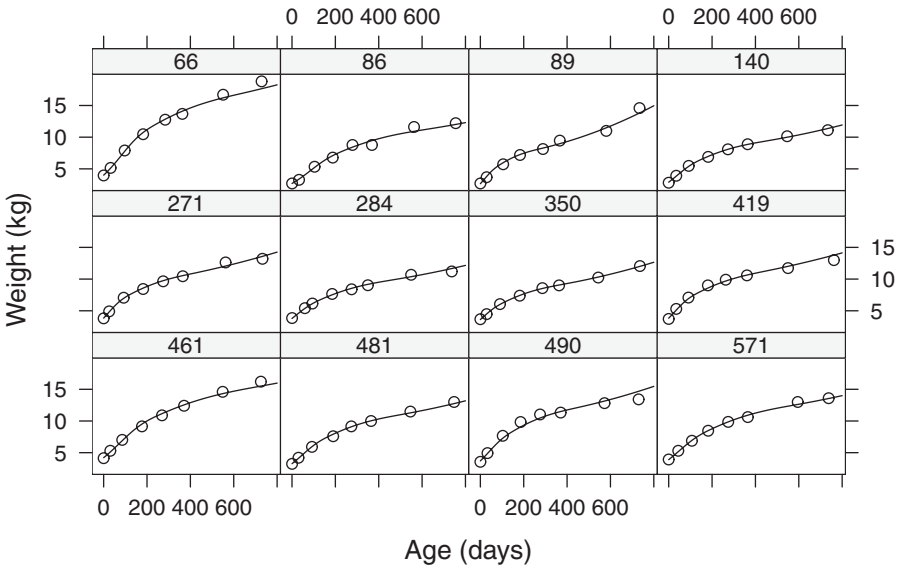

Figure 4. Observed and predicted weight *versus* age by subject.

Table II. Estimated standard deviation of random effects from final model.

	Male			Female		
Random effect	SD	Correlation		SD	Correlation	
α_0	0.18	α_0	β_0	0.20	α_0	β_0
β_0	22	0.72		24	0.79	
β_1	0.34	0.77	0.66	0.38	0.81	0.70
Residual	0.043			0.040		

DOI: 10.1002/sim

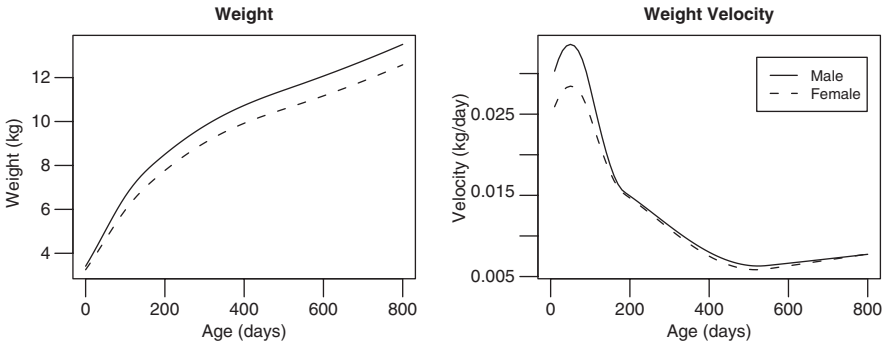

Figure 5. Population predicted weight and weight velocity *versus* age.

Table III. AIC for combined models.

Shape	Random effects	AIC
Common	Common	−10 984.6
Common	Individual	−10 980.4
Individual	Common	−11 097.3
Individual	Individual	−11 090.5

Table IV. Descriptive statistics for breastfeeding and wheeze.

	Breastfed			Wheeze	
Visit	Yes	No	Unknown	Yes	No
1 month	386 (64%)	185 (31%)	30 (5%)	83 (14%)	518 (86%)
3 months	280 (48%)	236 (40%)	69 (12%)	101 (17%)	484 (83%)
6 months	227 (40%)	336 (59%)	9 (2%)	151 (26%)	421 (74%)
9 months	160 (29%)	400 (71%)	0 (0%)	153 (27%)	407 (73%)
1 year	104 (19%)	455 (81%)	0 (0%)	132 (24%)	427 (76%)
18 months		Not available		136 (25%)	417 (75%)
2 years		Not available		103 (19%)	441 (81%)

at all ages. Of more interest is the weight velocity, which is initially high but rapidly decreases, a feature that has been previously observed [36]. The velocity is initially higher in males than females but after 6 months the growth velocity is similar for both sexes.

6.2. Males and females combined

Models were fitted to compare the models for combining the two genders. Models were fitted for the combinations of separate shape and random effects and are shown in Table III. This shows that individual shape curves are required but with a common random effects distribution.

DOI: 10.1002/sim

6.3. Combined model with covariates

At each visit it was recorded whether the infant was currently breastfeeding. As information was not available for when breastfeeding was stopped within an interval, it was assumed that if a baby was breastfed at a visit, then breastfeeding continued for at least half the subsequent interval. This was different from wheeze which was recorded if it occurred within the interval, rather than at the time of the visit. To correct for the different time periods between visits, the presence of wheeze was pooled across 6 month periods. Descriptive statistics for breastfeeding and wheeze are shown in Table IV. Models were fitted with and without the assumptions, and no substantial difference in the results was found.

A model was fitted for breastfeeding and wheeze allowing for an effect of mother's birthplace on size and growth, with a separate shape function for each sex and common random effect distribution, with random effects for all three SIM parameters. For each covariate, a model with sex interaction was fitted and if significant at the 0.05 level was included in the final model. This was only necessary for breastfeeding. The results for this model are shown in Table V. There was found to be a significant effect of breastfeeding on growth but not of wheeze, with mother's

Table V. Results for model for weight.

		Relative growth (95 per cent CI)		p value
Growth rate				
Breastfeeding				
Males	Yes	0.80	(0.76, 0.83)	<0.0001
	No	1.00		
Females	Yes	0.85	(0.82, 0.89)	<0.0001
	No	1.00		
Wheeze				
Yes		1.011	(0.99, 1.04)	0.4
No		1.00		
Mother's birthplace				
Asia		1.09	(1.01, 1.18)	0.02
Other		1.06	(0.99, 1.14)	0.08
Australia		1.00		
Size				
Mother's birthplace				
Asia		0.94	(0.91, 0.97)	0.0003
Other		1.00	(0.98, 1.03)	0.8
Australia		1.00		

Table VI. Estimated standard deviation of random effects for model with covariates.

	No covariates			Covariates		
Random effect	SD	Correlation		SD	Correlation	
α_0	0.19	α_0	β_0	0.18	α_0	β_0
β_0	23	0.76		21	0.75	
β_1	0.37	0.80	0.68	0.39	0.79	0.67
Residual	0.041			0.041		

DOI: 10.1002/sim

birthplace affecting both size and growth. Table VI shows the random effects standard deviations for the models with and without covariates. As expected the standard deviations for both the random effects and residual are similar but slightly smaller in the model with covariates. Figure 6 shows the residuals for the covariate model.

The effect of breastfeeding produced reduced growth rate. This reduction was greater for males with a relative growth of 0.81 (95 per cent CI 0.77, 0.84) and for females 0.85 (95 per cent CI 0.81, 0.88), and may result from males being larger and as a result having greater energy requirements. A comparison is shown in Figure 7 for infants breastfed for the full first year *versus* those for the first 6 months only, which shows the reduced growth rate due to breastfeeding resulting in reduced weight at the end of the first year. The results are in agreement with the WHO Working Group on Infant Growth [37] who found that breastfeeding resulted in lower weight gain. This is important as the increased growth during the time when the infant is not breastfed is assumed to be maintained in later life, which is in agreement with the results of Hediger *et al.* [38] and

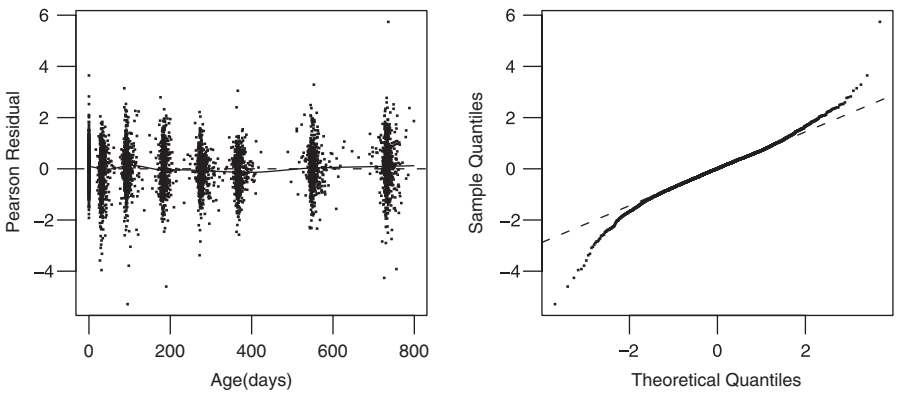

Figure 6. Residuals for covariate model.

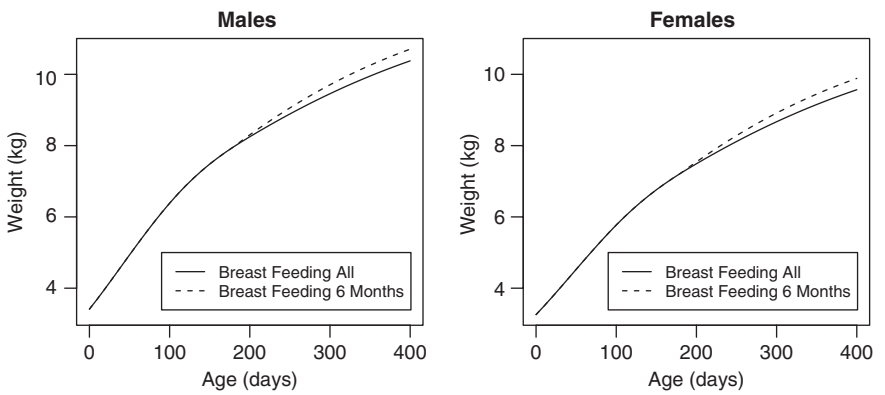

Figure 7. Population weight *versus* age for subjects breastfed for full first year *versus* first 6 months only.

DOI: 10.1002/sim

Gillman *et al*. [39] who found that breastfeeding was associated with reduced risk of being overweight in later life.

7. DISCUSSION

The analysis demonstrates that the SIM using a natural cubic spline and random effects for the shape invariant parameters is a useful model for infant growth data. The fit of the models was good with some systematic departures from the data. The residuals showed severe kurtosis, probably due to the errors being from a mixed distribution. A reasonable explanation is that there are two sources of error, the usual day-to-day variation and more extreme variation caused by illness or changes in feeding habits. Models require only three parameters to adjust the base curve to fit each individual subject, the same number as the exponential and fewer than the model of Jenss and Bayley, and the log transform of the outcome was shown to be appropriate for the data. Importantly, the model can be fitted with standard nonlinear mixed effects software.

One difficulty with using a spline function is to choose the number and location of the knots. For this application knots were located at the time points of the observed data, and attempts to remove knots beyond the final set resulted in convergence failure. Changing the location of the boundary knots did not have any appreciable affect on the estimated curves. The approach of using fixed boundary knots appears appropriate for infant data, due to the only moderate transformation of the time scale, but for data requiring a greater range of transformation, alternative methods would be required.

An important feature of the SIM is that the parameters have a simple biological interpretation. From the model definition: α_0 is the size of the infant at the transformed time origin, β_1 scales the time axis and so determines the growth rate and β_0 is an alignment parameter which allows for variation of the birth date with respect to the growth curve. For two infants, with the same growth rate β_1, any difference in size will be maintained during the infants growth, however, as α_0 is an estimate of the size at the time origin, it is possible for an infant to be smaller, reflected in a smaller α_0, but be larger at 2 years age due to a greater growth rate β_1.

The model was shown to fit well, and is useful for both simple and covariate analysis of growth data. The advantage of this method is that the parameter estimates for the model are easily interpretable, with parameters being interpreted either as a change in size or rate of growth of the subject. In comparison, for methods where the covariates are included as linear functions within the parameters of a nonlinear function, it may be difficult to obtain a physical interpretation of how the covariate affects growth. The methods used should be applicable to parametric nonlinear functions as well as the spline based functions used in the analysis, and may be fitted using standard nonlinear mixed-effect software.

APPENDIX

The following code has been tested with `R` version 2.3.1 with `nlme` version 3.1-73 on Windows, Linux and MacOS X. Starting values for the SIM were obtained by fitting models without random effects using `nls`. Convergence failure usually indicates fitting a model with unnecessary random effects, and can be improved by changing the `nlme` options to increase the number of `nlm`, PNLS and total iterations, reduce the pnlsTol, or by initially fitting models with simpler random

DOI: 10.1002/sim

effects structures to obtain starting values. It is important that the covariates be centred, by either subtracting the mean prior to fitting or including the mean in the fitting, as in the examples.

A.1. Simple model

```
fitnlme <- function(age,s1,s2,s3,s4,s5,salpha0,sbeta0,sbeta1){
   splinecoefs <- as.matrix(cbind(s1,s2,s3,s4,s5))
   as.vector(salpha0+t(matrix(rep(1,5),ncol=5) %*%
       t(splinecoefs*as.matrix(ns((age-sbeta0)/exp(sbeta1),
      knots=myknots,Boundary.knots=mybounds)))))
}

sim.male.weight <-
   nlme(logWt~fitnlme(Agedays,s1,s2,s3,s4,s5,alpha0,beta0,beta1),
      data=capsMales,
      fixed = s1+s2+s3+s4+s5+alpha0 ~ 1,
      random = alpha0+beta0+beta1 ~ 1 | Id,
      start = c(inits1,inits2,inits3,inits4,inits5,inits0)
)
```

A.2. Simple model with individual shape and random effects

```
# create indicators for random effects
capsAll$Male <- ifelse(capsAll$Sex==1,1,0)
capsAll$Female <- ifelse(capsAll$Sex==2,1,0)
sim.all.weight4 <-
   nlme(logWt~fitnlme(Agedays,s1,s2,s3,s4,s5,alpha0,beta0,beta1),
      data=capsAll,
      fixed = s1+s2+s3+s4+s5+alpha0 ~ Sex,
      random = list(Id=pdBlocked(list(
         pdSymm(alpha0+beta0+beta1~Male-1),
         pdSymm(alpha0+beta0+beta1~Female-1)))),
      start = c(inits1,0,inits2,0,inits3,0,inits4,0,inits5,0,inits0,0)
)
```

A.3. Covariate model

Includes a single time-dependent covariate for breastfeeding, the other covariates have been omitted to simplify the code.

```
fitnlme3 <- function(agebf,agenobf,sex,s1,s2,s3,s4,s5,
     deltas1,deltas2,deltas3,deltas4,deltas5,
      salpha0,sbeta0,sbeta1,ssex,sbfmale,sbffemale) {
   mys1 <- ifelse(sex==0,s1,s1+deltas1)
   mys2 <- ifelse(sex==0,s2,s2+deltas2)
   mys3 <- ifelse(sex==0,s3,s3+deltas3)
   mys4 <- ifelse(sex==0,s4,s4+deltas4)
   mys5 <- ifelse(sex==0,s5,s5+deltas5)
   splinecoefs <- as.matrix(cbind(mys1,mys2,mys3,mys4,mys5))
     newage <- agebf*exp((1-meanbfmale)*sbfmale*sex+
               (1-meanbffemale)*sbffemale*(1-sex))+
               agenobf*exp(-meanbfmale*sbfmale*sex-meanbffemale*sbffemale*(1-sex))
   as.vector(ssex*sex+salpha0+t(matrix(rep(1,5),ncol=5) %*%
       t(splinecoefs*as.matrix(ns((newage-sbeta0)/exp(sbeta1),
      knots=myknots,Boundary.knots=mybounds)))))
}
```


DOI: 10.1002/sim

```
sim.weight6 <-
   nlme(logWt~fitnlme3(daysbf,daysnobf,mySex,s1,s2,s3,s4,s5,
      deltas1,deltas2,deltas3,deltas4,deltas5,
      alpha0,beta0,beta1,ssex,sbfmale,sbffemale),
      data=capsWt,
      fixed = s1+s2+s3+s4+s5+deltas1+deltas2+deltas3+deltas4+deltas5+
         alpha0+ssex+sbfmale+sbffemale ~ 1,
      random = alpha0+beta0+beta1 ~ 1 | Id,
      start = c(s1=inits1,s2=inits2,s3=inits3,s4=inits4,s5=inits5,
      deltas1=initdeltas1,deltas2=initdeltas2,deltas3=initdeltas3,
      deltas4=initdeltas4,deltas5=initdeltas5,alpha0=inits0,ssex=initssex,
      sbfmale=initsbfmale,sbffemale=initsbffemale)
)
```

ACKNOWLEDGEMENTS

Jenny Peat for supplying the data, and the investigators and nurses of the CAPS study for data collection. Gillian Heller, Malcolm Hudson and the referees for their comments and suggestions.

REFERENCES

1. Wishart J. Growth-rate determinations in nutrition studies with the bacon pig, and their analysis. *Biometrika* 1938; **30**:16–28.
2. Karlberg J. On the modelling of human growth. *Statistics in Medicine* 1987; **6**:185–192.
3. Jenss RM, Bayley N. A mathematical method for studying the growth of a child. *Human Biology* 1937; **9**(4):556–563.
4. Berkey CS. Comparison of two longitudinal growth models for preschool children. *Biometrics* 1982; **38**(1): 221–234.
5. Count EW. Growth patterns of the human physique: an approach to kinetic anthropometry. *Human Biology* 1943; **15**(1):1–32.
6. Milani S, Bossi A, Marubini E. Individual growth curves and longitudinal growth charts between 0 and 3 years. *Acta Paediatrica Scandinavia Supplement* 1989; **350**:95–104.
7. Wingerd J. The relation of growth from birth to 2 years to sex, parental size and other factors, using Rao's method of the transformed time scale. *Human Biology* 1970; **42**(1):105–131.
8. Geva D, Goldschmidt L, Stoffer D, Day NL. A longitudinal analysis of the effect of prenatal alcohol exposure on growth. *Alcoholism: Clinical and Experimental Research* 1993; **17**(6):1124–1129.
9. Ong KKL, Preece MA, Emmett PM, Ahmed ML, Dunger DB. Size at birth and early childhood growth in relation to maternal smoking, parity and infant breast-feeding: longitudinal birth cohort study and analysis. *Pediatric Research* 2002; **52**(6):863–867.
10. Newell ML, Borja MC, Peckham C. European collaborative study height, weight, and growth in children born to mothers with HIV-1 infection in Europe. *Pediatrics* 2003; **111**(1):e52–e60.
11. Rao CR. Some statistical methods for comparison of growth curves. *Biometrics* 1958; **14**(1):1–17.
12. Berkey CS, Laird NM. Nonlinear growth curve analysis: estimating the population parameters. *Annals of Human Biology* 1986; **13**(2):111–128.
13. Wang Y-G, Jackson CJ. Growth curves with time-dependent explanatory variables. *Environmetrics* 2000; **11**: 597–605.
14. Cox DR, Oakes D. *Analysis of Survival Data*. Chapman & Hall: London, 1984.
15. Mihrshahi S, Peat JK, Webb K, Tovey RE, Marks GB, Mellis CM, Leeder SR. The childhood asthma prevention study (CAPS): design and research protocol of a randomized trial for the primary prevention of asthma. *Controlled Clinical Trials* 2001; **22**(3):333–354.
16. Durrleman S, Simon R. Flexible regression models with cubic splines. *Statistics in Medicine* 1989; **8**:551–561.
17. Beacon HJ, Thompson SG, England PD. The analysis of complex patterns of longitudinal binary response: an example of transient dysphagia following radiotherapy. *Statistics in Medicine* 1998; **17**:2551–2561.

DOI: 10.1002/sim

18. Lawton WH, Sylvestre EA, Maggio MS. Self modeling nonlinear regression. *Technometrics* 1972; **14**(3):513–532.
19. Stützle W, Gasser T, Molinari L, Largo RH, Prader A, Huber PJ. Shape-invariant modelling of human growth. *Annals of Human Biology* 1980; **7**(6):507–528.
20. Gasser T, Kneip A, Ziegler P, Largo R, Prader A. A method for determining the dynamics and intensity of average growth. *Annals of Human Biology* 1990; **17**(6):459–474.
21. Lindstrom MJ. Self-modelling with random shift and scale parameters and a free-knot spline shape function. *Statistics in Medicine* 1995; **14**:2009–2021.
22. Pinheiro JC, Bates DM. *Mixed-Effects Models in S and S-Plus*. Springer: New York, 2000.
23. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2004.
24. Altman NS, Villarreal JC. Self-modelling regression for longitudinal data with time-invariant covariates. *Canadian Journal of Statistics* 2004; **32**(3):251–268.
25. Ke C, Wang Y. Semiparametric nonlinear mixed-effects models and their applications. *JASA* 2001; **96**:1272–1281.
26. Wang Y, Ke C, Brown MB. Shape-invariant modelling of circadian rhythms with random effects and smoothing spline ANOVA decompositions. *Biometrics* 2003; **59**(4):804–812.
27. Brumback LC, Lindstrom MJ. Self modeling with flexible, random time transformations. *Biometrics* 2004; **60**(2):461–470.
28. Cole TJ, Green PJ. Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine* 1992; **11**:1305–1319.
29. Cole TJ, Freeman JV, Preece MA. British 1990 growth reference centiles for weight, height, body mass index and head circumference fitted by maximum penalized likelihood. *Statistics in Medicine* 1998; **17**:407–429.
30. Susman EP, Murphy JR, Zerbe GO, Jones RH. Using a nonlinear mixed model to evaluate three models of human stature. *Growth, Development and Aging* 1998; **62**:161–171.
31. Robinson, GK. That BLUP is a good thing: the estimation of random effects. *Statistical Science* 1991; **6**(1):15–51.
32. Sakamoto Y, Ishiguro M, Kitagawa G. *Akaike Information Criterion Statistics*. Reidel: Dordrecht, Holland, 1986.
33. Lindstrom MJ, Bates DM. Nonlinear mixed effects models for repeated measures data. *Biometrics* 1990; **46**(3):673–687.
34. Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* 1995; **4**(1):12–35.
35. Diggle PJ, Heagarty P, Liang K-Y, Zeger SL. *Analysis of Longitudinal Data*. Oxford University Press: New York, 2002.
36. Gasser T, Kneip A, Binding A, Prader A, Molinari L. The dynamics of linear growth in distance, velocity and acceleration. *Annals of Human Biology* 1991; **18**(3):187–205.
37. Anderson MA, Dewey KG, Frongillo E, Garza C, Haschke F, Kramer M, Whitehead RG, Winichagoon P, Deonis M. An evaluation of infant growth: the use and interpretation of anthropometry in infants. *Bulletin of the World Health Organization* 1995; **73**(2):165–174.
38. Hegider ML, Overpeck MD, Kuczmarski RJ, Ruan WJ. Association between infant breastfeeding and overweight in young children. *JAMA* 2001; **285**(19):2453–2460.
39. Gillman MW, Rifas-Shiman SL, Camargo CA, Berkey CS, Frazier AL, Rockett HRH, Field AE, Colditz GA. Risk of overweight among adolescents who were breastfed as infants. *JAMA* 2001; **285**(19):2461–2467.

DOI: 10.1002/sim