Research Article

Statistics in Medicine

Received 8 November 2013, Accepted 16 December 2014 Published online 23 January 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6411

A multivariate Bayesian model for embryonic growth

Sten P. Willemsen,[a,b*†] Paul H. C. Eilers,[a]
Régine P. M. Steegers-Theunissen[b,c] and Emmanuel Lesaffre[a,d]

Most longitudinal growth curve models evaluate the evolution of each of the anthropometric measurements separately. When applied to a 'reference population', this exercise leads to univariate reference curves against which new individuals can be evaluated. However, growth should be evaluated in totality, that is, by evaluating all body characteristics jointly. Recently, Cole *et al.* suggested the Superimposition by Translation and Rotation (SITAR) model, which expresses individual growth curves by three subject-specific parameters indicating their deviation from a flexible overall growth curve. This model allows the characterization of normal growth in a flexible though compact manner. In this paper, we generalize the SITAR model in a Bayesian way to multiple dimensions. The multivariate SITAR model allows us to create multivariate reference regions, which is advantageous for prediction. The usefulness of the model is illustrated on longitudinal measurements of embryonic growth obtained in the first semester of pregnancy, collected in the ongoing Rotterdam Predict study. Further, we demonstrate how the model can be used to find determinants of embryonic growth. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: Bayesian modeling; multivariate statistics; growth curves

1. Introduction

The statistical analysis of longitudinal growth curves is an active area of research, with a wide range of applications in medicine, veterinary medicine, biology and other fields. Many models have been suggested in the literature to fit such data. When applied to a normal population, the fitted growth curves can be used to define 'normal growth'. It is important to detect whether subjects exhibit normal growth. For example, a good reference of normal fetal growth during pregnancy can be essential in the management of obstetrical care [1]. In addition, abnormal growth patterns can be an indication for disease later in life [2, 3]. Many growth models have a parametric, non-linear nature and are applied to specific application areas. Examples are the Gompertz curve, the logistic curve or the Brody curve; see, for example, [4].

To relax the parametric nature of the growth curve and hence to broaden the applicability of the growth curve methodology, Beath [5] suggested a flexible approach to model the longitudinal evolution of the growth patterns. This model was later extended by Cole *et al*. [6] who called it the 'Superimposition by Translation and Rotation' (SITAR) model. This is a shape invariant model with a single-fitted 'prototype' curve. Individual curves are obtained from the prototype curve by horizontal and vertical shifts and shrinking or stretching the age scale. However, the model is univariate in nature and therefore cannot completely capture what normal and abnormal growth is when considering multiple dimensions. Multivariate growth models have been suggested in the past, but they either lack flexibility or are difficult to interpret; see, for example, [7–9]. In this paper, the SITAR model is generalized to multiple dimensions and is referred to as the multivariate SITAR (MSITAR) model.

To illustrate the MSITAR model, we use data from an ongoing periconceptional cohort study, called the *Rotterdam Predict study*, conducted in the Department of Obstetrics and Gynaecology and the

[a] *Department of Biostatistics, Erasmus University Medical Center, Rotterdam, the Netherlands*
[b] *Department of Obstetrics and Gynaecology, Erasmus University Medical Center, Rotterdam, the Netherlands*
[c] *Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam, the Netherlands*
[d] *I-BioStat, KU Leuven, Leuven, Belgium*
[*] *Correspondence to: Sten P. Willemsen, Department of Biostatistics, Erasmus University Medical Center, Rotterdam, the Netherlands.*
[†] *E-mail: s.willemsen@erasmusmc.nl*

Department of Clinical Genetics at Erasmus University Medical Center (Erasmus MC), Rotterdam, the Netherlands. In this longitudinal study, human embryonic growth in the first trimester of pregnancy is studied. Early growth has been a black box for a long time, and it was assumed to occur uniformly, with very small differences between individuals. In fact, this is why gestational age (GA) is often estimated from embryonic growth measurements. The GA is based on the first day of the last menstrual period. However, this day cannot often be determined accurately. Lately, the idea of uniform growth has been largely abandoned, and the thought has taken root that there are important differences in embryonic growth in the first trimester [10–13]. Moreover, these first trimester differences have been related to birth outcomes [14–18]. Therefore, the study of first trimester embryonic growth has become increasingly important. Further understanding of normal embryonic growth and the identification of determinants involved will lead to a better and earlier prediction of adverse pregnancy course and future outcomes. Then women with a family history of adverse birth outcomes may be screened at an early stage, and timely interventions can be made, if necessary. In this respect, it is important to evaluate growth in its totality and not just look at measurements in separation. Indeed, measurements may appear to evolve in a normal manner when considered separately, but when appreciated in a multivariate sense, they may stand out as abnormal. In statistical terms, such observations are called multivariate outliers; see [19]. It is known that such outliers cannot be identified using univariate techniques. The MSITAR model has the ability to spot such outlying growth curves, but remains conceptually simple as the univariate SITAR model.

In Section 2, we discuss the motivating data set in more detail and highlight the research questions that are of interest. In Section 3, we describe the SITAR model exhaustively and show how it can be easily extended to multiple longitudinal series of measurements. In that section, we also indicate how multivariate reference regions can be obtained from the MSITAR model. Marginal and conditional reference contours will be discussed. In Section 4, we apply the MSITAR approach on data of the Rotterdam Predict study. In that section, we will also contrast the SITAR model applied on each of three parameters to the MSITAR model applied to the three measurements jointly. In Section 5, we give some concluding, clinical and statistical remarks. We also outline further extensions of our model.

2. Motivating data set: the Rotterdam Predict study

The Rotterdam Predict study is a periconceptional cohort study carried out at the Department of Obstetrics and Gynaecology at Erasmus MC, Rotterdam, the Netherlands. The overall aim of the study is to examine early human growth and to determine factors associated with complications originating in the periconceptional period and the early stages of pregnancy. The Rotterdam Predict study [12] is the first of its kind in which the early pregnancy is studied meticulously. In addition to regular, 2D, ultrasound images, three-dimensional (3D) ultrasound images are taken.

Every week, between the sixth and the thirteenth week of pregnancy, each participating woman receives a 2D and a 3D ultrasound. From these ultrasound measurements, the crown-rump length (CRL), total arc length (TAL) and the embryonic volume (EV) are determined. The CRL is defined as the shortest distance from the crown (top of the head) to the rump (buttocks). The CRL is one of the most important measurements made in early pregnancy as it is often used for pregnancy dating. The TAL is another measure of the distance between crown and rump introduced by Boogers *et al.* (unpublished manuscript), but now determined on the outside of the embryo along the dorsal side of the back as depicted in Figure 1.

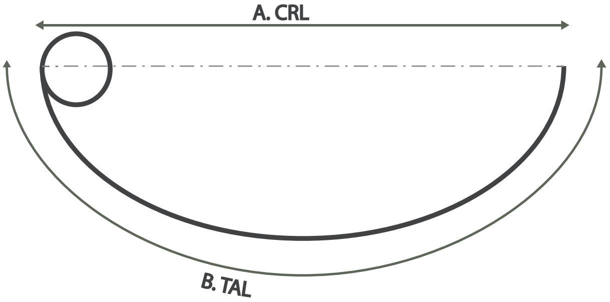


Figure 1. Illustration of crown-rump length (CRL) and total arc length (TAL): the head of the fetus is on the left, and the curved line represents its back. The distance A is the CRL, and the distance B is the TAL.

The relation between CRL and TAL is a function of the GA, which is defined as the number of days since the first day of the last menstrual period.

The embryonic volume is measured based on the 3D ultrasound images. Using the 3D images, holographic projections of the embryo are made in the Barco I-Space (Barco N.V., Kortrijk, Belgium). Using the V-Scope software (version 0.9.0 Ponca City, OK, USA), an interactive projection is made on the floor and walls using eight projectors in a way that allows depth perception using stereoscopic imaging [20]. From this image, the EV of the embryo was determined based on the gray level of the voxels (3D pixels) of the embryo. It has been suggested that EV gives a better indication of the GA than CRL [11]. In addition, EV is believed to be an important parameter to detect growth disorders [21]. Unfortunately, not all measurements could be performed using the 3D images because the holographic projector is an expensive piece of equipment that is also used for other research in Erasmus MC.

For this study, women above the age of 18 years and before the eighth week of pregnancy were recruited. The cohort is not a random sample of the general population. Rather, women were made aware of the study by posters placed in the Obstetrics and Gynaecology outpatient clinic of the Erasmus MC. A part of the women (23%) did not go to the Erasmus MC outpatient clinic prior to inclusion, but learned about the study through word-of-mouth. As a consequence, the women included in the cohort are likely to have a higher risk of pregnancy complications.

We included 259 eligible singleton pregnancies in the Rotterdam Predict study between 2009 and 2010. Forty-four women were excluded because of ectopic pregnancy or miscarriage occurring before the 16th week of gestation. These pregnancies were excluded because the course of the pregnancy was different from regular pregnancies and we are only interested in normal healthy pregnancies. Another 12 pregnancies were excluded because they could not be dated based on the last menstrual period. This leaves 203 pregnancies for the analysis.

Finally, not all of the images could be used for all measurements mostly because of a low image quality or an unfavorable position of the embryo. There are on average approximately five available measurements per pregnancy for the CRL and the TAL measurement and approximately four measurements for the EV. The EV is often more difficult to measure accurately because sometimes the volume of the fetus

Table I. Descriptive statistics of the selected women from the predict study (n = 203).

Characteristic	Mean/*N*	(*SD*/%)
Age	32.2	(4.8)
Missing	11	(5%)
Ethnicity		
Dutch	150	(74%)
Other western	16	(8%)
Other non-western	30	(15%)
Missing	7	(3%)
Education		
Low	18	(9%)
Middle	56	(28%)
High	113	(56%)
Missing	16	(8%)
Body mass index	24.6	(4.1)
Missing	8	(4%)
Parity		
Nulliparous	124	(61%)
Multiparous	74	(36%)
Missing	5	(2%)
Periconceptional alcohol use		
Yes	87	(43%)
No	109	(54%)
Missing	7	(3%)
Periconceptional smoking		
Yes	31	(15%)
No	165	(81%)
Missing	7	(3%)

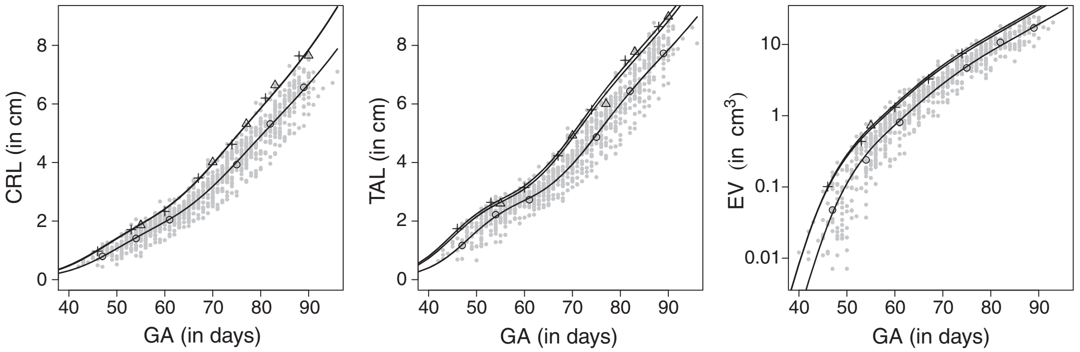


Figure 2. Scatterplots of the ultrasound measurements taken in the first 3 months of pregnancy. Three randomly selected pregnancies are indicated. The estimated profiles for these individuals are obtained from the multivariate Superimposition by Translation and Rotation model (Section 4.2). CRL, crown-rump length; TAL, total arc length; EV, embryonic volume; GA, gestational age.

cannot be clearly separated from its surroundings on the ultrasound. However, in some cases, EV could be determined while CRL or TAL could not be measured.

In Table I, we provide some characteristics of the study sample, which reveals that the participants are predominantly highly educated. Because they are recruited mostly from a tertiary hospital, they are at a higher risk for pregnancy complications. Therefore, one cannot immediately extrapolate the results of this study to the general population. As a comparison, in the regular hospital population in the Netherlands, about 16% of the women giving birth is of non-western origin, about 51% is nulliparous, and the median age is between 30 and 34 years [22]. Notice that we do have some missing values for some variables. The cases with missing data include two cases where the pregnancy was conceived after oocyte donation. For these cases, the relation between maternal characteristics and growth is likely to be different than in regular pregnancies. In addition, there are 10 pregnancies for which one or more of the covariates we will consider are missing. So for analyses that use covariates, we have 191 pregnancies left. We could also have imputed the missing values within the MCMC scheme. However, this increases computational complexity. Because we feel it distracts from the main idea of the paper, we did not follow this path.

Here, we use the Rotterdam Predict study to illustrate how the multivariate SITAR model can be used to model human embryonic growth. In particular, we indicate that our model is better capable to detect abnormal pregnancies.

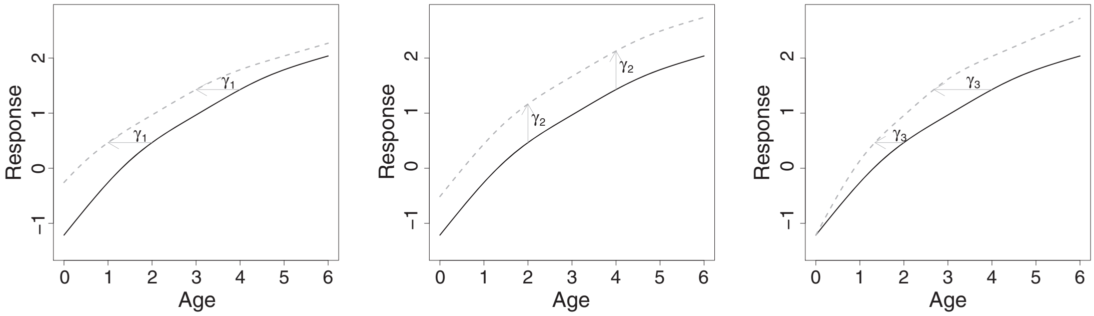
In Figure 2, we show the scatterplots of the three features of embryonic growth versus GA, wherein some randomly selected growth profiles are highlighted. Note that we plotted EV on a logarithmic scale. Clearly, there is a large amount of correlation between the measurements of each pregnancy, both within each separate outcome but also between the different series.

3. The Superimposition by Translation and Rotation model

3.1. The univariate Superimposition by Translation and Rotation model

The SITAR model is based on the ‘shape invariant’ model of infant growth suggested by Beath [5]. The idea behind the model is that there exists some general growth curve or prototype function (f), such that all individual growth profiles can be reduced to this general curve by translating them horizontally and vertically and by stretching them horizontally. So we have $y_{ij} = \gamma_{i2} + f(\gamma_{i3}[t_{ij} + \gamma_{i1}]) + \varepsilon_{ij}$, where y_{ij} is the outcome of individual i measured at time t_{ij}. The γs are the subject-specific effects and express the subject-specific horizontal shift (γ_{i1}), the vertical shift (γ_{i2}) and the stretch (γ_{i3}) with respect to the general growth curve. ε_{ij} is the measurement error. See Figure 3 for a schematic representation of the effect of the random effects. Note that stretching the general curve horizontally is something that is related to accelerated failure time models in which the scale parameter has a similar role as γ_{i3}.

In the SITAR model, the curve of the general pattern is modeled by a natural cubic spline function. A spline of degree d is a smooth function that is piecewise constructed from polynomials of degree d. For a cubic spline, this degree is three. The points where the polynomial functions join are called the inner



(a) Horizontal shift **(b)** Vertical shift **(c)** Horizontal Stretch

Figure 3. Schematic representation of Superimposition by Translation and Rotation model with horizontal shift (γ_{i1}), vertical shift (γ_{i2}) and stretch (γ_{i3}).

knots, while the knots at the boundary are the outer knots. A spline is called 'natural' when the second derivative at the outer knots of the spline is zero. The spline can then be linearly extrapolated beyond the outer knots. Proper extrapolation is needed because the horizontal shifts imply that the domain in which the splines will be evaluated are not exactly known beforehand. With κ inner knots, the natural cubic spline has $\kappa + 2$ independent coefficients.

Now the univariate SITAR model can be expressed as follows:

$$\begin{aligned}
y_{ij} &= \gamma_{i2} + z_{ij}^T \boldsymbol{\beta} + \varepsilon_{ij}, \\
z_{ij} &= \boldsymbol{B}\left(\exp(\gamma_{i3})\left(t_{ij} + \gamma_{i1}\right)\right), \; (i = 1, \dots, N;\; j = 1, \dots, n_i), \\
\boldsymbol{\gamma}_i &\sim \mathrm{N}\left(\mathbf{0}, \Sigma_\gamma\right), \\
\varepsilon_{ij} &\sim \mathrm{N}\left(0, \sigma^2\right).
\end{aligned} \tag{1}$$

We made the assumption here that $\boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2}, \gamma_{i3})^T$ is normally distributed and its components are centered at zero to avoid identifiability problems. $\boldsymbol{B}(t)$ is a function that returns the basis of the natural cubic spline, evaluated at t. Thus, z_{ij} is a vector of length $\kappa + 2$, as is the regression coefficient vector $\boldsymbol{\beta}$. We also assume that ε_{ij} is independently distributed with variance σ^2 and independent of the γs.

The SITAR model has been successfully applied to different data sets, appears to work well in practice and has parameters that are easily interpretable by clinicians [6].

3.2. *The multivariate Superimposition by Translation and Rotation model*

The univariate SITAR model can be easily extended to $K(> 1)$ outcomes. The MSITAR model is given by the following:

$$\begin{aligned}
y_{ijk} &= \gamma_{i2k} + z_{ijk}^T \boldsymbol{\beta}_k + \varepsilon_{ijk}, \\
z_{ijk} \equiv z_{ijk}\left(\boldsymbol{\gamma}_{M_i}\right) &= \boldsymbol{B}\left(\exp\left(\gamma_{i3k}\right)\left(t_{ijk} + \gamma_{i1k}\right)\right), \; (i = 1, \dots, N;\; j = 1, \dots, n_{ik};\; k = 1, \dots, K), \\
\boldsymbol{\gamma}_{M_i} &\sim \mathrm{N}\left(\mathbf{0}, \Sigma_{M_\gamma}\right), \\
\varepsilon_{ijk} &\sim \mathrm{N}\left(0, \sigma_k^2\right),
\end{aligned} \tag{2}$$

where y_{ijk} is the kth response of individual i at time t_{ij}, z_{ijk} is a basis of a natural cubic spline for the kth response of individual i at time t_{ij}, $\boldsymbol{\beta}_k$ is the vector of spline coefficients for the kth response, $\gamma_{ilk}\,(l = 1, 2, 3)$ are the three subject-specific effects for individual i and series k $(k = 1, \dots, K)$, and $\boldsymbol{\gamma}_{ik} = (\gamma_{i1k}, \gamma_{i2k}, \gamma_{i3k})^T$ is the vector of all subject-specific effects for outcome k and individual i. When we combine the subject-specific effects for all outcomes for an individual, we obtain $\boldsymbol{\gamma}_{M_i} = \left(\boldsymbol{\gamma}_{i1}^T, \dots, \boldsymbol{\gamma}_{iK}^T\right)^T$. Σ_{M_γ} is the $3K \times 3K$ joint covariance matrix of all random effects $\boldsymbol{\gamma}_{M_i}$. ε_{ijk} is the measurement error which we assume to be independently distributed and independent of the γs. Finally, σ_k^2 is the measurement error variance for series k. Note that it is assumed that the K outcomes are independent, given the random effects.

The MSITAR has the advantage over the univariate SITAR model that the three series of measurements can be modeled jointly. In general, joint modeling of multiple series of growth curves allows for the following:

- Evaluating relationships between the growth curve series over time.
- Establishing multivariate reference regions and thereby checking whether individuals not only show a normal growth pattern for an individual series but also for, for example, ratios of series.
- Better prediction of a future growth profile of a series when no or few past observations are available by also incorporating information from the past profiles of other series.

We were also interested in the effect of covariates on the parameters of the model. That is, we assumed that the means of the $\boldsymbol{\gamma}$s could depend on covariates. Let $\boldsymbol{x}_i$ be the vector of covariates for subject i of length n_p. We then assume that $\boldsymbol{\gamma}_{M_i} \sim \mathrm{N}(A\boldsymbol{x}_i, \Sigma_{M_\gamma})$. Here, A is a $3K{\times}n_p$ matrix of parameters with elements $\alpha_{k,p}$. We will also use the notation $\boldsymbol{\alpha}$ to denote $\mathrm{vec}(A) = \left(\alpha_{1,1}, \alpha_{2,1}, \dots, \alpha_{3K,n_p}\right)^T$, which is the vectorized form of the matrix A, that is the column matrix that consists of the columns of A that are stacked on top of each other.

3.3. MCMC implementation

We used a Bayesian approach to estimate the model parameters, employing vague priors for all parameters. When possible, we also opted for priors that are conditionally conjugate. Specifically, we assumed for the αs and βs independent normal priors with zero mean and a standard deviation equal to $\sigma_{\alpha,0} = \sigma_{\beta,0} = 1,000$, which is large relative to the plausible values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. For σ^2, an inverse γ prior was chosen with shape parameter (α_σ) and rate parameter (β_σ) equal to 0.001, that is, $\mathrm{IG}(0.001, 0.001)$. Finally, for Σ_{M_γ}, we have chosen an inverted Wishart prior with degrees of freedom δ equal to three times the number of outcomes and a scale matrix Ψ equal to 0.01 times the identity matrix, that is, $\mathrm{IW}(3K, 0.01\mathrm{I}_{3K})$. For a motivation of these choices, see [23, p. 260]. As more (external) information is collected on growth during early pregnancy, we believe that, in due time, some of these vague priors may be replaced by more informative priors.

The posterior distribution is proportional to the product of the likelihood and the priors, and therefore we obtain:

$$\begin{aligned} p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\gamma}_M, \Sigma_{M_\gamma} \mid \boldsymbol{y}) \propto & \prod_{ijk} \mathrm{N}\left(y_{ijk} \mid \gamma_{i2k} + z_{ijk}\left(\boldsymbol{\gamma}_{M_i}\right)^T \boldsymbol{\beta}_k, \sigma_k^2\right) \prod_i \mathrm{N}\left(\boldsymbol{\gamma}_{M_i} \mid \boldsymbol{0}, \Sigma_{M_\gamma}\right) \\ & \times \prod_k \mathrm{N}\left(\boldsymbol{\alpha}_k \mid \boldsymbol{0}, \sigma_{\alpha,0}^2 I_{n_p}\right) \prod_k \mathrm{N}\left(\boldsymbol{\beta}_k \mid \boldsymbol{0}, \sigma_{\beta,0}^2 I_{\kappa+2}\right) \\ & \times \prod_k \mathrm{IG}\left(\sigma_k^2 \mid \alpha_\sigma, \beta_\sigma\right) \ \mathrm{IW}\left(\Sigma_{M_\gamma} \mid \delta, \Psi\right), \end{aligned} \tag{3}$$

with $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_K^T\right)^T, \boldsymbol{\sigma}^2 = \left(\sigma_1^2, \dots, \sigma_K^2\right)^T, \boldsymbol{\gamma}_M = \left(\boldsymbol{\gamma}_{M_1}, \dots, \boldsymbol{\gamma}_{M_N}\right)^T$ and $\boldsymbol{y} = \left(y_{111}, y_{121}, \dots, y_{1n_i1}, y_{211}, \dots, y_{N,n_{NK},K}\right)^T$.

We implemented a Gibbs sampling procedure to estimate the model parameters. The (block) full conditionals for $\boldsymbol{\beta}, \boldsymbol{\alpha}\ \sigma_k^2$ and Σ_{M_γ} are given in the supplementary material. Most full conditionals are of standard form and hence standard samplers can be used. However, this is not the case for the full conditional distribution of the subject-specific effects, given by:

$$\boldsymbol{\gamma}_{M_i} | \boldsymbol{y}, \dots \sim N_{3K}\left(\boldsymbol{\gamma}_{M_i} \mid A\boldsymbol{x}_i, \overline{\Sigma}_\gamma\right) \times \prod_{j,k} \mathrm{N}\left(y_{ijk} \mid \gamma_{i2k} + \gamma_{i3k} B(t_{ij} + \gamma_{i1k}) \boldsymbol{\beta}_k, \sigma_k^2\right), \tag{4}$$

where a Metropolis step is needed.

The above model can be estimated in JAGS [24]. We also implemented the sampler directly into C++ making use of the ‘Eigen’ and ‘Boost’ libraries [25,26]. By embedding the program into the R language, its output can be post-processed with R functions, such as CODA [27]. For the full conditionals of the subject-specific effects, we used, in the C++ program, a Random Walk Metropolis algorithm with a multivariate t proposal distribution with 5 degrees of freedom. The acceptance rate of the Metropolis algorithm was tuned so that it was approximately 30%. The source code of our programs can be obtained upon request from the first author.

3.4. Model selection and evaluation

Model selection was done using the popular Deviance Information Criterion (DIC) [28]. More specifically, we compared the MSITAR model with the ensemble of univariate SITAR models to check the necessity of modeling the three series of growth curve responses jointly.

To check the assumptions of our model, we performed posterior predictive checks. That is, based on the estimated model, we simulated replicated series. We then compared the distribution of the test statistic calculated on the replicates with the test statistic based on the data which we actually observed. Specifically, we looked at the chi-squared goodness-of-fit test recommended by Gelman [29]. That is, we drew posterior samples for each of the K outcomes

$$\chi^2_{\mathrm{GM},k} = \sum_{i,j} r^2_{ijk} \qquad \text{with} \qquad r_{ijk} = \frac{y_{ijk} - E\left(y_{ijk} \mid \tilde{\boldsymbol{\theta}}\right)}{sd\left(y_{ijk} \mid \tilde{\boldsymbol{\theta}}\right)} \tag{5}$$

and compared it with $\chi^2_{\mathrm{GM},k,\mathrm{rep}}$ obtained from replacing y_{ijk} by the corresponding sampled value $\tilde{y}_{ijk}$ from the posterior predictive distribution in (5), $p(\tilde{y}_{ijk}|\boldsymbol{y})$. In this expression, $\tilde{\boldsymbol{\theta}}$ stands for the sampled total parameter vector, and $E(y_{ijk})$ is defined as $\boldsymbol{B}(\tilde{\gamma}_{i3k}(\tilde{\gamma}_{i1k} + t_{ijk}))\tilde{\boldsymbol{\beta}}_k + \tilde{\gamma}_{i2k}$. We then checked the proportion of samples in which $\chi^2_{\mathrm{GM},k,\mathrm{rep}} \geqslant \chi^2_{\mathrm{GM},k}$. In a similar way, we computed the posterior predictive checks evaluating the skewness and kurtosis of the error terms ε_{ijk}. For instance, the statistic $\chi^2_{\mathrm{SKEW},k} = \sum_{i,j} r^3_{ijk}$ checks the skewness of the errors, which we assumed zero.

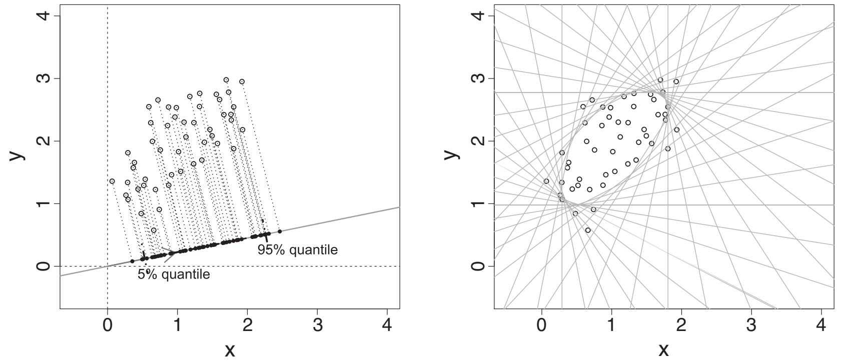
Finally, we extended the model so it could accommodate a residual error with excess kurtosis. Specifically, we have replaced the normal residual error by a scaled t-distribution. In our C++ program, we simulated from a t-distribution by considering it as a scale mixture of normals. We considered the inverse of the number of degrees of freedom to be uniformly distributed on $\left[0, \frac{1}{3}\right]$. With this prior, we give a larger prior probability to smaller numbers of degrees of freedom, that is, to distributions that are further away from the normal.

3.5. Predictive ability of the model

The predictive ability of the models was evaluated by the root mean squared prediction error (RMSE). We want to know, once we have an estimated model, how well it can predict a future observation of a new pregnancy given the available measurements of this pregnancy. This was evaluated using five-fold cross-validation, whereby the total data set was split into five (approximately) equal parts, each containing 20% of the pregnancies. Each time, 80% of the observations served as training set and the remaining 20% part as validation set. In the validation part, we removed the last value for each combination of pregnancy and outcome, which we will denote by $y_{ik,\mathrm{LAST}}$, which is to be predicted. In the supplemental material, we provide the details of the procedure we used to draw samples from the predictive distribution of this last observation. We then took the mean of these samples for each i and k, denoted by $\hat{y}_{ik,\mathrm{LAST}}$ and calculated the RMSE as $\sqrt{\sum_{i=1}^{N_k}(y_{ik,\mathrm{LAST}} - \hat{y}_{ik,\mathrm{LAST}})^2/N_k}$. Here, N_k is the number of pregnancies for which we observed the kth outcome.

3.6. Reference contours

Our model allows for the construction of multivariate reference ranges. With a single outcome, a reference range is a central predictive interval in which 100(1-q)% of the observations are located. Often, equal-tail predictive intervals are taken. This can be generalized to multiple dimensions so we obtain a reference region excluding the most outlying observations and containing the least outlying observations (or for lack of a better word, the 'normal' observations). However, such an extension is not trivial because univariate quantiles are not easily adapted to the multivariate case [30]. Here, we followed the approach of Kong and Mizera [31] and look at contours obtained from a collection of directional quantiles. For a direction indicated by a unit vector $\boldsymbol{s}$, the *qth directional quantile* in the direction of $\boldsymbol{s}$, that is, $Q(q, \boldsymbol{s})$ of a multivariate random variable $\boldsymbol{y}$, is defined as the quantile of $\boldsymbol{s}^T\boldsymbol{y}$, which corresponds to the orthogonal



(a) Directional quantiles are the quantiles of the projected points

(b) The directional quantile lines circumscribe a convex region: thereference contour

Figure 4. The construction of directional quantiles.

projection of $\boldsymbol{y}$ on $\boldsymbol{s}$. To construct the directional quantile contour $Q(q)$, one varies the direction given by the vector $\boldsymbol{s}$ to cover all angles. Each $Q(q, \boldsymbol{s})$ defines a half space where $100(1 - q)\%$ of the observations lie. By taking the intersection of these half spaces, one obtains the contour $Q(q)$; see Figure 4 for a graphical representation of this construction. Note, however, that in this figure, only a limited number of directions $\boldsymbol{s}$ has been taken to enhance the legibility of the figure. In general, $Q(q)$ will exclude more than a fraction q of the observations. A contour that excludes a fraction q of the observations is then obtained by an optimization routine to find a value q' for which $Q(q')$ does have this property.

To generate the directional quantile contours based on our model, we generated samples $\boldsymbol{y}$ for a new individual from the posterior predictive distribution of our model and drew a large number of directional quantiles based on this sample. This results in unconditional or marginal contours. It is also possible to derive conditional contours, that is, when there is interest in the expected range of future 'normal' outcome values for subjects given their history (i.e., past covariates and responses). These values can be generated in much the same way as the unconditional ones. The difference is that we now sample γ_{M_i} from the conditional distribution (4) of those individuals given their history.

Conditional quantile contours can also be constructed based on the distribution of the residuals in a way that avoids sampling. For example, if a multivariate outcome is (conditionally) multivariate normal, the probability contours are elliptical. However, this fails to take the uncertainty in the parameter estimates into account properly. We believe this is important in a clinical setting when limited information is available and parameter uncertainty is relatively large.

3.7. *Effect of time scale*

The γ_{i3k} random effects result in stretching the time axis in order to, together with the other random effects, align the curve of the ith individual with the overall 'mean' smooth curve. We argue that this stretch effect is most effective when the first measurement of each individual is taken just after time zero. When the origin of the time axis is remote from the measurement times, the stretch effect is rather similar to the horizontal shift effect, so sampling becomes more difficult. However, in human embryonic growth studies, measurements cannot be taken at conception. In fact, in the Rotterdam Predict study, the first measurement was taken between days 40 and 50 since no reliable measurements can be made before this time with the current techniques. For this reason, model (2) is based on the GA minus 36 days. In this way, the transformed GA at the first occasion that measurements could be taken is small but positive. In general, when we apply a linear transformation on the time scale, we have: $\tilde{t}_{ijk} = \frac{t_{ijk}+a}{b}$ so the second line of model (2) becomes $z_{ijk} = \tilde{B}(exp(\tilde{\gamma}_{i3k})(\tilde{t}_{ijk} + \tilde{\gamma}_{i1k})$. Here $\tilde{B}$ is a new spline function that is created by applying the same linear transformation on the nodes of B. When we only change the time scale, the effect is similar to increasing the prior variance of the horizontal shifts by a factor $\sqrt{b}$. The effect of a is to change the origin of the time scale. In this case, the effect of the transformation is more profound, and finding an appropriate prior would be much more complicated than changing the time scale. In theory, optimal a and b could be estimated from the data by treating them as extra parameters. However, we have chosen here not to do so to avoid further computational complexity.

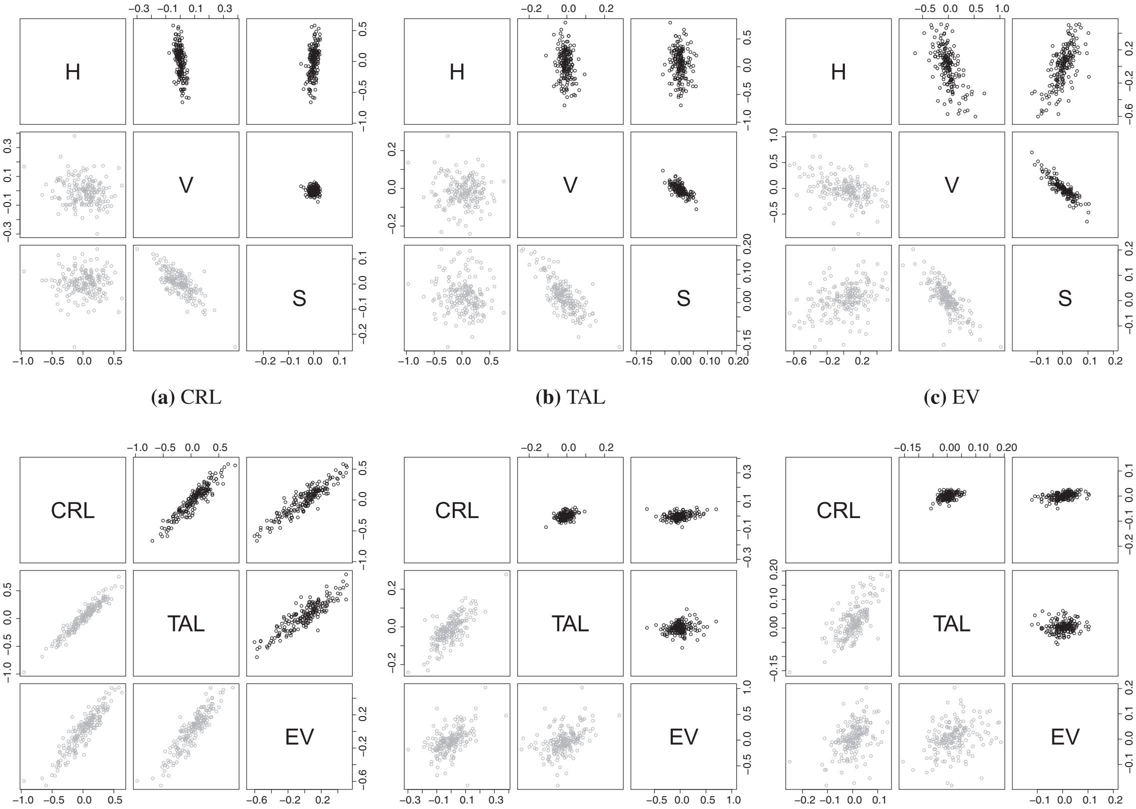
4. Analysis of the motivating data set

In this section, we present the results of our analysis. First, we estimated the SITAR model separately for each of the three series (CRL, TAL and EV) and looked at the correlation between the subject-specific effects obtained in these models. We then estimated the joint model for all the outcomes together. In the next subsection, we focus on the inclusion of covariates. In all analyses, we used $\kappa = 3$ with equidistant nodes, which provided enough flexibility for the prototype growth curve. The univariate models were run for 600,000 iterations (with 300,000 burn-in iterations) and the multivariate model for 1,200,000 iterations (with 600,000 burn-in iterations). For all models, three chains were initiated with different starting values. Convergence was checked visually by examining the trace plots of the Markov chains and more formally by the Gelman and Rubin [32] test. The effective sample size ranged from approximately 1,200 for the $\boldsymbol{\beta}$ values to 11,000 for the σ^2 values. On an Intel i5-24000 processor running at 3.1 GHz with 8 GB RAM and the 64 bit version of Windows, the time needed to run 1,200,000 iterations for the multivariate model was approximately 18 h.

4.1. *Univariate analyses*

We used a logarithmic transformation for all responses. This reduced heteroscedasticity in the responses and also ensures that predictions are always positive. Note that the vertical shift on the log scale corresponds to a stretch effect on the original scale.

We wanted to investigate the relation between the subject-specific effects that belong to the same pregnancy. Therefore, we have made scatterplots of the components of the posterior means of the centered effects. By centered effects, we mean the deviations of these effects from their expected values (i.e., $\boldsymbol{\gamma}_M - XA^T$). In Figure 5, we organized the scatterplots of pairs of variables into six scatterplot matrices. In the upper triangular part of each scatterplot matrix, we show the pairwise scatterplots of



(**a**) CRL (**b**) TAL (**c**) EV

(**d**) Horizontal Effects (**e**) Vertical Effects (**f**) Stretch Effects

Figure 5. Scatterplots of the posterior means of the subject-specific effects $\boldsymbol{\gamma}_M - XA^T$ from the univariate model (upper triangular) and multivariate model (lower triangular). ‘H’ denotes the horizontal shift, ‘V’ the vertical shift, and ‘S’ the stretch effect.

Table II. Posterior predictive checks of the *univariate* SITARs.

	PPC: normal errors			PPC: t-distributed errors		
	CRL	TAL	EV	CRL	TAL	EV
χ^2_{GM}	0.48	0.52	0.49	0.48	0.35	0.45
χ^2_{SKEW}	0.27	0.12	0.52	0.41	0.22	0.78
χ^2_{KURT}	0.00	0.00	0.00	0.69	0.61	0.90

Note: SITAR, Superimposition by Translation and Rotation; PPC, posterior predictive checks; CRL, crown-rump length; TAL, total arc length; EV, embryonic volume.

the posterior mean of the centered random effects within the same series (subplots a–c) and the scatterplot between similar effects in different series (subplots d–f). The correlation between the horizontal and vertical effects is always negative. Other than that, there does not seem to be an obvious pattern in the correlations of each outcome. When we look at the correlation between similar effects of different outcomes, the very strong correlations of 0.9 between the horizontal shifts stand out (subplot d). The other correlations are smaller but still positive (0.1–0.4). The relative size of the horizontal shifts, compared with their sd's, which are not shown here, appears greater than those of the vertical and stretch effects.

In Table II, we show the results of the posterior predictive checks. Here, numbers close to either zero or one indicate that model assumptions might not be met. As is obvious from the leftmost columns of this table, there are some deviations from the model assumptions when we assume a normally distributed measurement error. The residuals appear to be platykurtic distributed. This could be remedied by replacing the normal residual error by a t-distribution (this is illustrated in the rightmost part of the table). The posterior means of the degrees of freedom were 4.9, 3.9 and 3.5 for the CRL, TAL and EV respectively. In the Supplementary Material, we provide Q–Q plots of the subject-specific effects. There appear to be some, relatively minor, deviations from normality. In particular, the stretch effect of the TAL seems to have heavier tails than expected. An obvious way to remedy this is also to change the distribution of the subject-specific effects. Unfortunately, we were unable to fit a model where both the subject-specific effects were non-normal. Because changing the residual errors to a scaled t-distribution also improved the Q–Q plots of the subject-specific effects, we continued with this model.

4.2. Multivariate analysis

We then estimated the joint model for all three series together, that is, the MSITAR model. For this, we used the results of our univariate models as starting values of the sampling procedure. The scatterplots of the subject-specific effects within each series are shown in the lower triangular subplots of Figure 5a–c. The points have slightly fanned out. This is most evident in the vertical and shift effects in the CRL and TAL (Figure 5a and b). The correlation between the horizontal and vertical effects is still negative like in the univariate models. Again, there does not seem to be a pattern in the correlations that involve the stretch effect. Most correlations have the same sign as in the univariate case. Notable exceptions are the relations between the stretch effect and the two other effects in the CRL (compare the rightmost column and bottom row of Figure 5a). In Figure 5d–f, we show the scatterplot of similar subject-specific effects between similar effects in different series. In Figure 5d, we see that, like in the univariate models, the correlation between the horizontal effects is strong (around 0.9). The correlation between the vertical shift and the stretch effects has also become stronger when compared with the univariate models (Figure 5d and f). As for the univariate models, we found some deviations from the model assumptions with the posterior predictive checks (Table III). Again, the observed kurtosis does not match that of the normal distribution. When we replaced the normal distribution with a t-distribution, this problem was solved for CRL and TAL. However, the t-distributions seem to be unable to accurately model the kurtosis of the EV. The posterior means of the degrees of freedom are 4.5, 4.0 and 3.3 for the CRL, TAL and EV respectively. For three pregnancies, we have plotted the estimated profiles based on the MSITAR model with t-distributed errors in Figure 2. The estimated profiles from the model with a normal error look very much like these ones and are not shown.

4.3. Comparison of the univariate and multivariate models

We compared the collection of the three univariate models with the MSITAR model using DIC. We obtained for the multivariate model a DIC of −7222 with 922 effective parameters (p_D). This value

Table III. Posterior predictive checks of the *multivariate* SITAR models.

	PPC: normal errors			PPC: t-distributed errors		
	CRL	TAL	EV	CRL	TAL	EV
χ^2_{GM}	0.50	0.48	0.19	0.50	0.28	0.04
χ^2_{SKEW}	0.37	0.14	0.25	0.22	0.15	0.53
χ^2_{KURT}	0.00	0.00	0.09	0.21	0.70	1.00

Note: SITAR, Superimposition by Translation and Rotation; PPC, posterior predictive checks; CRL, crown-rump length; TAL, total arc length; EV, embryonic volume.

Table IV. Root mean squared prediction error of the ensemble of SITAR models and the MSITAR model (both with normal and t-distributed errors (MSITAR-T)).

	RMSE		
Variable	SITARs	MSITAR	MSITAR-T
log (CRL)	0.051	0.053	0.052
log (TAL)	0.042	0.043	0.044
log (EV)	0.172	0.171	0.172
log (CRL/TAL)	0.042	0.041	0.042

Note: SITAR, Superimposition by Translation and Rotation; MSITAR, Multivariate Superimposition by Translation and Rotation; CRL, crown-rump length; TAL, total arc length; EV, embryonic volume; RMSE, root mean squared prediction error.

is substantially smaller than for the collection of univariate models for which the DIC is −6785 with $p_D = 1030$. Hence, the MSITAR model is preferred.

In Table IV, we compare the RMSE between the univariate and multivariate models. The RMSEs largely agree. So from these results, we must conclude that there does not seem to be any benefit in using the MSITAR model for prediction. However, this general conclusion overlooks some special situations that are highly relevant in practice. Indeed, one advantage of the multivariate model is that it also enables us to make predictions for one of the outcomes when no previous measurements on that kind of outcome are available. For example, we refer to the EV of the pregnancy marked with 'triangles' in Figure 2. The reason is that, in the MSITAR model, the correlation among the series is used to predict the responses of the missing series while, in the univariate models, there is nothing that can provide this information. Even when we have an outcome for which we only have a single previous measurement of the same outcome as the outcome which we want to predict, the multivariate model does slightly better. For example, when we predict the last measured EV in those pregnancies where we had only one or two previous observations, the RMSE of the multivariate models was 0.29 vs 0.33 for the univariate model. For the CRL and the TAL, we have more than two observations for almost all of the pregnancies.

4.4. *Covariates*

Using the MSITAR model, we aimed to look at the effect of periconceptional conditions such as maternal age, nulliparous pregnancies (i.e., first-time pregnancies), smoking and alcohol use of the mother and maternal body mass index on CRL, TAL and EV. All continuous covariates were standardized while for the binary variables, we used effect coding. In Table V, we show the estimated coefficients. Here, we only show the estimated coefficients from the MSITAR model with t-distributed errors; however, the results are similar for the univariate models and models with normal distributed errors. We see that the pattern of the coefficients is similar across the different outcomes. It seems that the horizontal effects are most 'significant'.

While the parameters have an obvious meaning, a predictor can influence the response of an outcome in multiple ways. For example, smoking at the same time causes children to be 'slow' (i.e., behind the curve) and bigger. Therefore, to better appreciate the effect of a particular covariate, we plotted the 'average' profiles for a typical pregnancy. In Figure 6, we have graphically shown the effect of smoking on the three series.

Table V. Regression coefficients (posterior means and standard deviations) of the MSITAR model.

	Horizontal effect		Vertical effect		Stretch effect	
	Coef	SE	Coef	SE	Coef	SE
			CRL			
Maternal age	0.045	0.021	0.017	0.014	−0.0060	0.0074
Nulliparous pregnancy	−0.056	0.021	0.021	0.013	−0.0028	0.0071
Smoking	−0.085	0.025	0.021	0.017	−0.0085	0.0086
Alcohol use	−0.006	0.024	−0.023	0.013	0.0075	0.0068
BMI	0.025	0.022	−0.013	0.013	0.0018	0.0068
			TAL			
Maternal age	0.071	0.022	0.009	0.009	−0.0089	0.0056
Nulliparous pregnancy	−0.053	0.022	0.021	0.009	−0.0069	0.0056
Smoking	−0.091	0.026	0.028	0.011	−0.0169	0.0075
Alcohol use	−0.004	0.022	−0.017	0.008	0.0050	0.0054
BMI	0.008	0.024	0.002	0.009	−0.0048	0.0057
			EV			
Maternal age	0.048	0.019	−0.012	0.072	0.0046	0.0170
Nulliparous pregnancy	−0.033	0.019	0.021	0.063	0.0014	0.0155
Smoking	−0.054	0.023	0.205	0.074	−0.0563	0.0185
Alcohol use	−0.008	0.020	0.087	0.056	−0.0283	0.0136
BMI	0.019	0.019	0.052	0.076	−0.0145	0.0185

Note: MSITAR, Multivariate Superimposition by Translation and Rotation; BMI, body mass index; CRL, crown-rump length; TAL, total arc length; EV, embryonic volume; Coef, coefficient; SE, standard error.

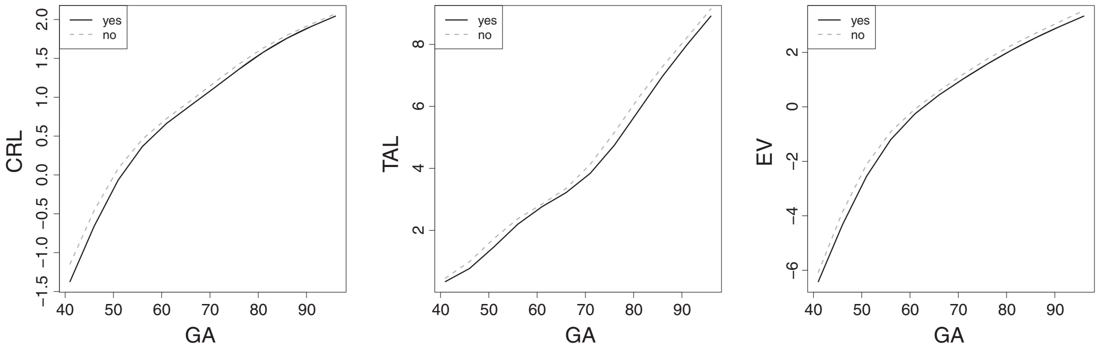
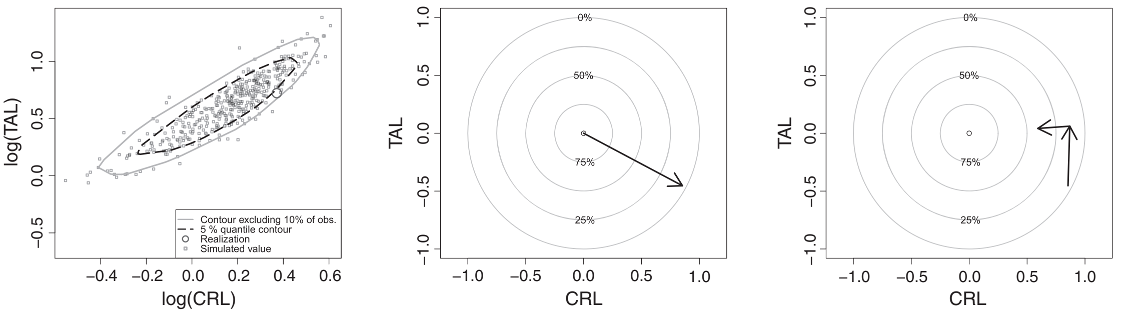


Figure 6. Effect of smoking. CRL, crown-rump length; TAL, total arc length; EV, embryonic volume; GA, gestational age.

4.5. *Reference contours*

Unconditional reference contours allow us to determine the range of ‘normal’ ultrasound images at a particular GA for a new pregnancy. In Figure 7, we demonstrate how these contours can be constructed. First, one generates a sample of $\boldsymbol{y}$ for a new individual from the estimated MSITAR model and draws all directional quantiles based on this sample. This is illustrated in Figure 7a where we have simulated values for CRL, TAL and EV for an embryo of 53 days old from the Rotterdam Predict study. Note that this pregnancy appears quite normal univariately with a CRL of 1.45 mm and a TAL of 2.06 mm but becomes rather unusual when we look at it bivariately. Note also that the fraction of the sample that is outside of the contour of the 5% directional quantiles is actually larger than 10%. As a reference, we have also plotted the quantile contour that does encompass 90% of the observations. Another way to visualize how ‘normal’ a certain observation is to plot the minimum directional quantile and the direction in which this minimum is reached, that is, the pair $(min_s(Q(q,\boldsymbol{s})), argmin_s(Q(q,\boldsymbol{s}))$. This is done in Figure 7b. By connecting the values observed at different time points, as is done in Figure 7c, one can visualize the



(a) Simulated values and reference contours **(b)** Direction and size of minimum depth **(c)** Following an individual over time

Figure 7. Reference contours. TAL, total arc length; CRL, crown-rump length.

trajectory of an individual pregnancy relative to the whole group. In this case, we see that although the pregnancy is quite outlying at the start, it becomes more ordinary later.

For a conditional reference contour, prediction is done assuming that we already have a number of ultrasound images from a certain pregnancy as well as background information such as smoking status, alcohol use and maternal age. Now, we wish to know what measurement values would be considered as normal at a future time point. We again simulate replicate observations from the model, but now, we use the full conditional (4) to simulate the subject-specific effects, taking into account all the information we have for this pregnancy. Plots of the conditional reference contours and the evolution of the multivariate quantile over time can be made in much the same way as for the unconditional contours and quantiles and are therefore not shown here.

4.6. Simplifying the random effect structure

Triggered by the high correlation between the horizontal effects, we investigated if it would be reasonable to assume that these parameters are the same in each of the separate growth curves. Indeed, when we simplified the model in this respect, the DIC improved by 70 points in the model with normally distributed errors and by 15 points in the model with t-distributed errors.

5. Discussion and conclusions

We have extended the SITAR model of Cole *et al.* to include several outcomes, which we refer to as the MSITAR model. The model is quite flexible in modeling different growth patterns while at the same time remaining easy to interpret. In this paper, we have applied this model on embryonic growth data; however, it can also be used to model other kinds of growth data and even in completely different fields of study.

We have shown that the MSITAR model provides a better fit to the data of the Rotterdam Predict study than the collection of univariate SITAR models. Oddly enough, this does not result in an improved predictive ability. However, a benefit of the MSITAR model is that, in case we have missing values in one of the series, we can recover some of that missing information from the other series. This is very relevant in a clinical setting where it is impossible to make frequent measurements of all outcomes. By means of the MSITAR model, we can predict future values for an outcome when we have a single previous measurement for that outcome or even no previous measurements at all as long as we have measurements for other outcomes. Furthermore, the MSITAR model enables one also to look at the relation between the various series which is useful on its own.

Surely, other techniques suitable to model flexibly multivariate longitudinal data are available; see, for example, [9, 33, 34]. However, we argue that the parameters of these models are less interpretable than those of the MSITAR model. A related technique is formed by the class of three-mode models discussed in [35]. This approach requires balanced measurements, which is often not the case in growth studies.

We have successfully used the MSITAR model to assess various potential determinants of embryonic growth. However, because all three types of effect influence the outcome simultaneously, we have to calculate the net effect at different time points as we did in Section 4.4, which diminishes the usefulness of the interpretation of the γ parameters.

The Bayesian procedure lends itself to the construction of reference contours in which all parameter uncertainty is taken into account. These can be used to detect multivariate outliers that would otherwise go unnoticed. Furthermore, covariates and past measurements of a pregnancy can be incorporated so that the reference values can be individualized.

As is generally known, the Bayesian MCMC estimation procedure allows us to easily change some of model components, which is a great advantage over some other computational approaches. Here, we replaced the usual normally distributed error term by t-distributed errors. In addition, we have examined the effect that covariates may have on early embryonic growth. Currently, we are also looking if we can use the subject-specific effects in our model to predict birth outcomes. In the Bayesian methodology, the individual estimates are automatically shrunken, which is advantageous for prediction.

In theory, the MSITAR model can be extended to more than three outcomes. However, from a practical point, we must admit that the model will quickly become too complex to fit. Still, to apply the MSITAR model to more than three dimensions, the current approach might be combined with a dimension reduction technique as in [36]. Furthermore, here we have analyzed the 'normal' pregnancies. Another approach could be to take basically all pregnancies. This strategy might be preferred because in the early period after conception, it is sometimes hard to know whether the pregnancy will be aborted. In fact, one of the reasons for setting up the Rotterdam Predict study is to discover abnormalities. Such a sample will require, however, that we model the random part of the MSITAR model more flexibly. We are currently working on these and other extensions in a Bayesian context.

References

1. Maršál K, Persson PH, Larsen T, Lilja H, Selbing A, Sultan B. Intrauterine growth curves based on ultrasonically estimated foetal weights. *Acta Paediatrica* 1996; **85**(7):843–848.
2. Nieto FJ, Szklo M, Comstock GW. Childhood weight and growth rate as predictors of adult mortality. *American Journal of Epidemiology* 1992; **136**(2):201–213.
3. Cameron N, Demerath EW. Critical periods in human growth and their relationship to diseases of aging. *American Journal of Physical Anthropology* 2002; **119**(S35):159–184.
4. Forni S, Piles M, Blasco A, Varona L, Oliveira HN, Lôbo RB, Albuquerque LG. Comparison of different nonlinear functions to describe Nelore cattle growth. *Journal of Animal Science* 2009; **87**(2):496–506.
5. Beath KJ. Infant growth modelling using a shape invariant model with random effects. *Statistics in Medicine* 2007; **26**(12):2547–2564.
6. Cole TJ, Donaldson MDC, Ben-Shlomo Y. SITAR–a useful instrument for growth curve analysis. *International Journal of Epidemiology* 2010; **39**(6):1558–1566.
7. Goldstein H. Efficient statistical modelling of longitudinal data. *Annals of Human Biology* 1986; **13**(2):129–141.
8. Reinsel Greg. Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. *Journal of the American Statistical Association* 1982; **77**(377):190–195.
9. Macdonald-Wallis C, Lawlor DA, Palmer T, Tilling K. Multivariate multilevel spline models for parallel growth processes: application to weight and mean arterial pressure in pregnancy. *Statistics in Medicine* 2012; **31**(26):3147–3164.
10. Bottomley C, Daemen A, Mukri F, Papageorghiou AT, Kirk E, Pexsters A, De Moor B, Timmerman D, Bourne T. Assessing first trimester growth: the influence of ethnic background and maternal age. *Human Reproduction* 2009; **24**(2):284–290.
11. Rousian M, Koning A, Van Oppenraaij R, Hop W, Verwoerd-Dikkeboom C, Van der Spek P, Exalto N, Steegers E. An innovative virtual reality technique for automated human embryonic volume measurements. *Human Reproduction* 2010; **25**(9):2210–2216.
12. Van Uitert EM, Exalto N, Burton GJ, Willemsen SP, Koning AH, Eilers PH, Laven JS, Steegers EA, Steegers-Theunissen RP. Human embryonic growth trajectories and associations with fetal growth and birthweight. *Human Reproduction* 2013; **28**(7):1753–1761.
13. Van Uitert E, Van Ginkel S, Willemsen S, Lindemans J, Koning A, Eilers P, Exalto N, Laven J, Steegers E, Steegers-Theunissen R. An optimal periconception maternal folate status for embryonic size: the Rotterdam Predict study. *BJOG: An International Journal of Obstetrics & Gynaecology* 2014; **121**(7):821–829.
14. Smith GC, Smith MF, McNay MB, Fleming JE. First-trimester growth and the risk of low birth weight. *New England Journal of Medicine* 1998; **339**(25):1817–1822.
15. Mukri F, Bourne T, Bottomley C, Schoeb C, Kirk E, Papageorghiou AT. Evidence of early first-trimester growth restriction in pregnancies that subsequently end in miscarriage. *BJOG* 2008; **115**(10):1273–1278.
16. Bukowski R, Smith GCS, Malone FD, Ball RH, Nyberg DA, Comstock CH, Hankins GDV, Berkowitz RL, Gross SJ, Dugoff L, Craigo SD, Timor-Tritsch IE, Carr SR, Wolfe HM, D'Alton ME, F. A. S. T. E. R Research Consortium. Fetal growth in early pregnancy and risk of delivering low birth weight infant: prospective cohort study. *BMJ* 2007; **334**(7598):836.
17. Mook-Kanamori DO, Steegers EP, Eilers PH, Raat H, Hofman A, Jaddoe VV. Risk factors and outcomes associated with first-trimester fetal growth restriction. *The Journal of the American Medical Association* 2010; **303**:527–534.
18. Carbone JF, Tuuli MG, Bradshaw R, Liebsch J, Odibo AO. Efficiency of first-trimester growth restriction and low pregnancy-associated plasma protein-A in predicting small for gestational age at delivery. *Prenatal Diagnosis* 2012; **32**(8):724–729.

19. Rousseeuw PJ, Van Zomeren BC. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association* 1990; **85**(411):633–639.
20. Koning AHJ, Rousian M, Verwoerd-Dikkeboom CM, Goedknegt L, Steegers EAP, Van der Spek PJ. V-scope: design and implementation of an immersive and desktop virtual reality volume visualization system. *Studies in Health Technology and Informatics* 2009; **142**:136–138.
21. Rolo LC, Nardozza LMM, Araujo Júnior E, Nowak PM, Bortoletti Filho J, Moron AF. Measurement of embryo volume at 7–10 weeks' gestation by 3D-sonography. *Journal of Obstetrics & Gynaecology* 2009; **29**(3):188–191.
22. De Graaf JP, Ravelli ACJ, Visser GHA, Hukkelhoven C, Tong WH, Bonsel GJ, Steegers EAP. Increased adverse perinatal outcome of hospital delivery at night. *BJOG: An International Journal of Obstetrics & Gynaecology* 2010; **117**(9): 1098–1107.
23. Lesaffre E, Lawson A. *Bayesian Biostatistics*, Statistics in Practice. Wiley: New York, 2012.
24. Plummer M. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, 2003, pp. 20–22.
25. Boost C++ libraries, 2012. http://www.boost.org/ [Accessed on 1 June 2014].
26. Guennebaud G, Jacob B, Niesen J, Heibel H, Nuentsa D, Hertzberg C, Capricelli T. Eigen v3, 2010, p. 2. http://eigen.tuxfamily.org/index.php? [Accessed on 1 June 2014].
27. Plummer M, Best N, Cowles K, Vines K. CODA: convergence diagnosis and output analysis for MCMC. *R News* 2006; **6**(1):7–11. http://CRAN.R-project.org/doc/Rnews/ [Accessed on 1 June 2014].
28. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002; **64**(4):583–639.
29. Gelman A, Meng X-L, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* 1996; **6**:733–807.
30. Serfling R. Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica* 2002; **56**(2):214–232.
31. Kong L, Mizera I. Quantile tomography: using quantiles with multivariate data. *Statistica Sinica* 2008; **22**:1589–1610.
32. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992; **7**(4): 457–472. http://www.jstor.org/stable/2246093 [Accessed on 1 June 2014].
33. Xu S, Styner M, Gilmore J, Gerig G. Multivariate longitudinal statistics for neonatal-pediatric brain tissue development. *Proceedings*, 2008, 69140C–69140C–11.
34. Rosen O, Thompson WK. A Bayesian regression model for multivariate functional data. *Computational Statistics & Data Analysis* 2009-09; **53**(11):3773–3786.
35. Oort FJ. Three-mode models for multivariate longitudinal data. *British Journal of Mathematical and Statistical Psychology* 2001; **54**(1):49–78.
36. Slaughter JC, Herring AH, Thorp JM. A Bayesian latent variable mixture model for longitudinal fetal growth. *Biometrics* 2009; **65**(4):1233–1242.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.