The International Journal of Biostatistics

Volume 7, *Issue* 1 2011 *Article* 32

Modeling Fetal Weight for Gestational Age: A Comparison of a Flexible Multi-level Spline-based Model with Other Approaches

Luc Villandré, *McGill University Health Centre*
Jennifer A. Hutcheon, *University of British Columbia*
Maria Esther Perez Trejo, *McGill University*
Haim Abenhaim, *McGill University*
Geir Jacobsen, *Norwegian University of Science and Technology*
Robert W. Platt, *McGill University*

Recommended Citation:
Villandré, Luc; Hutcheon, Jennifer A.; Perez Trejo, Maria Esther; Abenhaim, Haim; Jacobsen, Geir; and Platt, Robert W. (2011) "Modeling Fetal Weight for Gestational Age: A Comparison of a Flexible Multi-level Spline-based Model with Other Approaches," *The International Journal of Biostatistics*: Vol. 7: Iss. 1, Article 32.
DOI: 10.2202/1557-4679.1305

Modeling Fetal Weight for Gestational Age: A Comparison of a Flexible Multi-level Spline-based Model with Other Approaches

Luc Villandré, Jennifer A. Hutcheon, Maria Esther Perez Trejo, Haim Abenhaim, Geir Jacobsen, and Robert W. Platt

Abstract

We present a model for longitudinal measures of fetal weight as a function of gestational age. We use a linear mixed model, with a Box-Cox transformation of fetal weight values, and restricted cubic splines, in order to flexibly but parsimoniously model median fetal weight. We systematically compare our model to other proposed approaches. All proposed methods are shown to yield similar median estimates, as evidenced by overlapping pointwise confidence bands, except after 40 completed weeks, where our method seems to produce estimates more consistent with observed data. Sex-based stratification affects the estimates of the random effects variance-covariance structure, without significantly changing sex-specific fitted median values. We illustrate the benefits of including sex-gestational age interaction terms in the model over stratification. The comparison leads to the conclusion that the selection of a model for fetal weight for gestational age can be based on the specific goals and configuration of a given study without affecting the precision or value of median estimates for most gestational ages of interest.

KEYWORDS: multi-level models, fetal growth, small for gestational age

Author Notes: The authors would like to thank Michael S. Kramer for his helpful comments and suggestions. J.A.H. is the recipient of a post-doctoral Fellowship Award from the Canadian Institutes of Health Research (CIHR); R.W.P. is a Chercheur-Boursier of the Fonds de la Recherche en santé du Québec. R.W.P., L.V. and M.S.K. are members of the Montreal Children's Hospital Research Institute, which receives operating funds from the Fonds de la recherche en santé du Québec. This work was also supported by grant MOP-84379 from the CIHR.

1 Introduction

Poor fetal growth is strongly associated with adverse perinatal outcomes such as neurological damage, seizures, organ failure and perinatal mortality (Gabbe et al., 2007). Conventionally, a fetus or infant whose weight is in the smallest 10 percent of the population at a given gestational age is classified as small for gestational age (SGA) (Zhang et al., 2010b), and considered to be at increased risk of perinatal morbidity and mortality (Gabbe et al., 2007). Accurate weight-for-gestational-age percentiles are therefore needed to determine which fetuses may require additional investigations and closer monitoring. Conventional weight-for-gestational-age reference charts (Alexander et al., 1996, Kramer et al., 2001, Källén, 1995, Skjaerven et al., 2000), however, have an important shortcoming. Their trajectories of fetal "growth" are derived from the cross-sectional weights of infants born at different gestational ages, not from serial measurements of individual fetuses throughout pregnancy. Since preterm livebirths (infants born prior to 37 weeks of gestation) have been shown to be, on average, smaller than their in-utero peers (Fry et al., 2002, Hediger et al., 1995, Marsál et al., 1996, Ott, 1993, Weiner et al., 1985), models of fetal growth based on cross-sectional birth weight measurements are biased at younger gestational ages.

The introduction of obstetrical ultrasound has made it possible to estimate fetal weights prior to birth, and several models of fetal growth based on ultrasound estimates of fetal weight have been proposed. Hadlock et al. (1991) fitted a series of polynomial models to estimated fetal weight as a function of gestational age. The best model used a quadratic polynomial and log-transformed weight values, and was selected on the basis of the largest coefficient of determination (R^2), of the smallest standard deviation and of the inspection of the residuals for uniformity of variance. However, since their dataset contained only a single estimated weight per pregnancy, it was not designed to be a model of fetal growth per se.

Royston (1995) fitted a multi-level model to longitudinal estimated ultrasound fetal weight data. A Box-Cox transformation and fractional polynomials were used in order to stabilize variance and linearize the relationship between gestational age and transformed weight. A linear mixed model, which allowed for a random slope and a random intercept, was fitted to the transformed data.

Hooper et al. (2002) fitted a quadratic polynomial to the natural logarithm of the weight data. This polynomial regression fit produced residuals, which were then normalized by fitting a linear spline to the normal probability plot of the residuals and fitting a model for the standard deviation of the residuals. Combining those two estimates, they transformed the residuals into z-scores, which were split into their measurement error and latent score components, the latter being defined simply as the raw score stripped of its measurement error component.

Pan and Goldstein (1997) fitted a multi-level model for pediatric growth on weight z-scores computed by applying Cole's Lambda, Mu, Sigma (LMS) method (Cole and Green, 1992). The skewness, mean and variance functions, at the core of the LMS method, were estimated using cubic splines. They were combined to create a transformation function that uses weight values as input and outputs the required z-scores. The model was then used to derive unconditional and conditional norms for growth (conditional on past growth). Conditional norms could be estimated by fitting a linear regression model that has z-scores at a number of previous time points as covariates.

The lack of population-level, longitudinal fetal weight measurements has limited widespread adoption of longitudinal fetal growth references. However, there is now renewed interest in the collection of data to update fetal growth references, including ongoing work by the US National Institute of Health to develop a new National Standard for Normal Fetal Growth (Zhang et al., 2010a). Such work will need to select a longitudinal model for fetal weight for gestational age. Although different approaches to modeling longitudinal fetal weight measurements have been proposed, it is unknown to what extent the choice of model affects the estimates of fetal weight for gestational age, and which model, if any, is superior. A systematic comparison of the models is needed. Existing models may also benefit from improvements. Increasing model flexibility by including a large number of random effects parameters can now be more easily implemented thanks to advances in mixed model statistical software. Mixed models additionally offer convenient ways to deal with heteroscedasticity. Finally, adequately accounting for the influence of fetal sex, an important physiological determinant of fetal weight, may also improve fetal weight-for-gestational-age models.

2 Objectives

In this light, we sought to:

- Present a new, parsimonious, and more flexible modeling strategy for fetal weight for gestational age based on a sex-specific multi-level model,
- Systematically compare medians from the proposed modeling strategy with those of Royston (1995), Hooper et al. (2002) and Pan and Goldstein (1997), a description of which can be found in the appendices.

DOI: 10.2202/1557-4679.1305

3 Data

The dataset consists of longitudinal ultrasound data and birth data for singletons collected in Scandinavia from 1986 to 1988 for the *Successive SGA Birth study* (Bakketeig et al., 1993). Ultrasound biometric measurements were obtained at antenatal study visits at 17, 25, 33 and 37 weeks. Fetal abdominal circumference, femur length and biparietal diameter were combined using Hadlock's formula (Hadlock et al., 1984) to derive estimates of fetal weight. The original study included 1945 pregnancies, which included a 10% sample of the general obstetrical population (n=561) and an over-sampling of high-risk pregnancies (n=1384). In the 10% random sample, there were 454 pregnancies with both ultrasound and birth data; these were retained for the present analysis. Of these 454 pregnancies, 3 pregnancies were excluded because they had no complete birth weight - gestational age or estimated fetal weight - gestational age pairs, leaving 451 pregnancies for analysis. Estimates of gestational age were established using the date of the last normal menstrual period (*LNMP*) confirmed by an estimate of gestational age from early ultrasound through the following algorithm: In the case where the 17 week ultrasound estimate differed from that based on *LNMP* by less than 14 days, the *LNMP* estimate was chosen. If the difference was larger or equal or if the date of *LNMP* was unknown, the ultrasound estimate was chosen. The dataset is the largest of which we are aware to have 5 or more serial ultrasound measurements from an unselected obstetrical population. The representativeness of the study population, frequency of ultrasound measurements, and high study quality make it an ideal input for comparing models.

4 Methods

We fitted a linear mixed model to the fetal weight measurements using restricted cubic splines based on the truncated power basis (Harrell, 2001) to flexibly model the association between gestational age and weight. We included random effects on both the intercept and gestational age terms in order to account for between-fetus variability. Finally, since the variance of weight measurements increases with gestational age (Hadlock et al., 1991) we considered two approaches to manage heteroscedasticity:

1. The weight variable was transformed by means of a variance-stabilizing technique,
2. A variance structure that allows for heteroscedasticity was imposed on the residuals or the random effects.

A Box-Cox transformation was a natural choice in approach 1. The choice of the scaling parameter was driven by the need to make the output data as normally distributed and homoscedastic as possible. In order to ensure this, we adopted a *REML*-based approach devised by Gurka et al. (2006). However, applying the inverse transformation to the transformed model results in biased estimates of the mean, but not the median, on the original scale (Duan, 1983), so we present medians in subsequent analyses.

4.1 Sex-based differences in fetal weight

Evidence suggests that weight differences between male and female fetuses could be important as early as 14 completed weeks of pregnancy (Schwärzler et al., 2004). Therefore, in order to produce sex-specific weight-for-gestational-age curves in such a way that their shapes remain flexible and that the number of variance parameters to be estimated is reduced, we included sex-gestational age interaction terms in the model.

4.2 Model formulation

The model was formally specified as

$$g(\mathbf{W}_i) = \beta_0 + b_{i,0} + (\beta_1 + b_{i,1})\mathbf{T}_{i,1} + ... + (\beta_{k-1} + b_{i,k-1})\mathbf{T}_{i,k-1} + \\ + \beta_k \mathbf{T}_{i,1} S_i + \beta_{k+1}\mathbf{T}_{i,2} S_i + ... + \beta_{2k-2}\mathbf{T}_{i,k-1} S_i + \beta_{2k-1} S_i + \boldsymbol{\varepsilon}_i, \quad (1)$$

where

- $\mathbf{W}_i$ is a vector of weight values for individual i,
- $g(.)$ is the Box-Cox or the identity transformation,
- $\boldsymbol{\beta}_m$, $m = 1, 2, ..., 2k-1$, is the vector of fixed effect coefficients,
- $\mathbf{b}_{i,j}$, $j = 1, 2, ..., k-1$, is the vector of random effects,
- S_i is the code for sex,
- $\mathbf{T}_{ij}$ is gestational age expressed in the truncated power series basis,
- k is the number of knots,
- $\boldsymbol{\varepsilon}_i \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon)$,
- $\mathbf{b}_i \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_b)$.

DOI: 10.2202/1557-4679.1305

4.3 Comparison with other proposed approaches

Since the models proposed by Royston (1995), Hooper et al. (2002) and Pan and Goldstein (1997) ignore sex, they were fitted separately on data for male and female fetuses. Although our own model makes the assumption of a common variance-covariance structure for the weights of males and females, this assumption was relaxed in order to make the comparisons meaningful. This is equivalent to using stratification.We then fitted the model under the initial assumption of shared variance-covariance parameters to establish any potential advantages of not stratifying.

The comparisons were based on the location and the precision of the median estimates, as evidenced by their 95% pointwise bootstrap confidence bands, found after fitting the different models to the Scandinavian dataset. In addition, at term ages, we compared the median estimates to those found in a national birthweight-for-gestational-age chart (Skjaerven et al., 2000) based on births in Norway between 1987 and 1998. Comparisons with this chart were not made at preterm ages, since charts based on the cross-sectional weights of preterm newborns are known to be biased relative to fetal weight charts (Fry et al., 2002, Hediger et al., 1995, Marsál et al., 1996, Ott, 1993, Weiner et al., 1985).

4.4 Software

We carried out the analyses in *R* 2.9.1.

5 Results

5.1 Data characteristics

The observed weights for gestational age in the Scandinavian sample are shown in Figure 1. The strong and continuous increase in the variance of weight over time highlights the need to use modeling methods capable of dealing with heteroscedasticity. In Figures 2 and 3, randomly selected individual weight trajectories are shown. The occasional stabilization or decrease of weight at term (37 to 41 completed weeks of gestation) observed in certain trajectories may be the result of measurement error in estimated fetal weight at the 37 week antenatal visit. Work by Bertino et al. (1996) as well as Hooper et al. (2002) indicates that the observed flattening of the weight-for-gestational-age trajectory may also be due to a progressive decrease in growth velocity after 35 completed weeks, although it seems unlikely that this would explain the weight losses observed. Missing data were rare, with

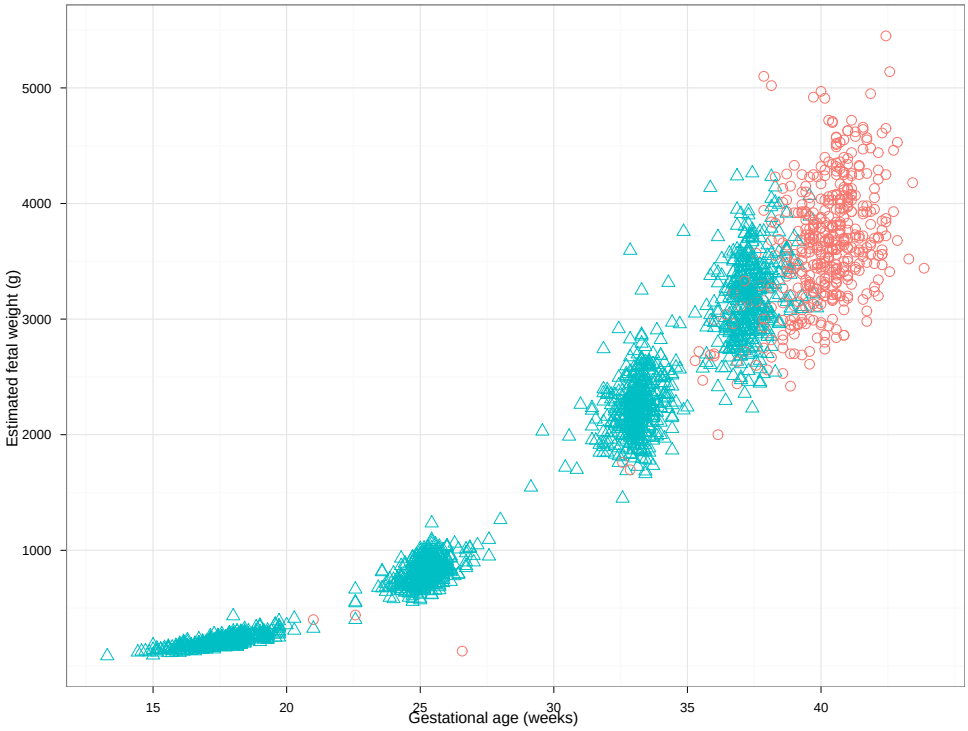

Figure 1: **Pregnancies from the Scandinavian dataset**: Circles represent weight/gestational age pairs at birth and triangles represent estimated weight/gestational age pairs in utero.

98% of births (440/451) having at most one missing weight measurement. Among the 451 pregnancies, 19 resulted in a pre-term birth, i.e. before 37 weeks (259 days), leading to a preterm birth rate of approximately 4%.

5.2 Mixed model characteristics

Basic spline regression models require the prior specification of knot positions. As recommended in the absence of substantive knowledge (Harrell, 2001), we placed knots at the 5th (119 days), 27.5th (175 days), 50th (232 days), 72.5th (262 days) and 95th (287 days) percentiles. Varying the number of knots did not result in noticeable improvements in model fit.

DOI: 10.2202/1557-4679.1305

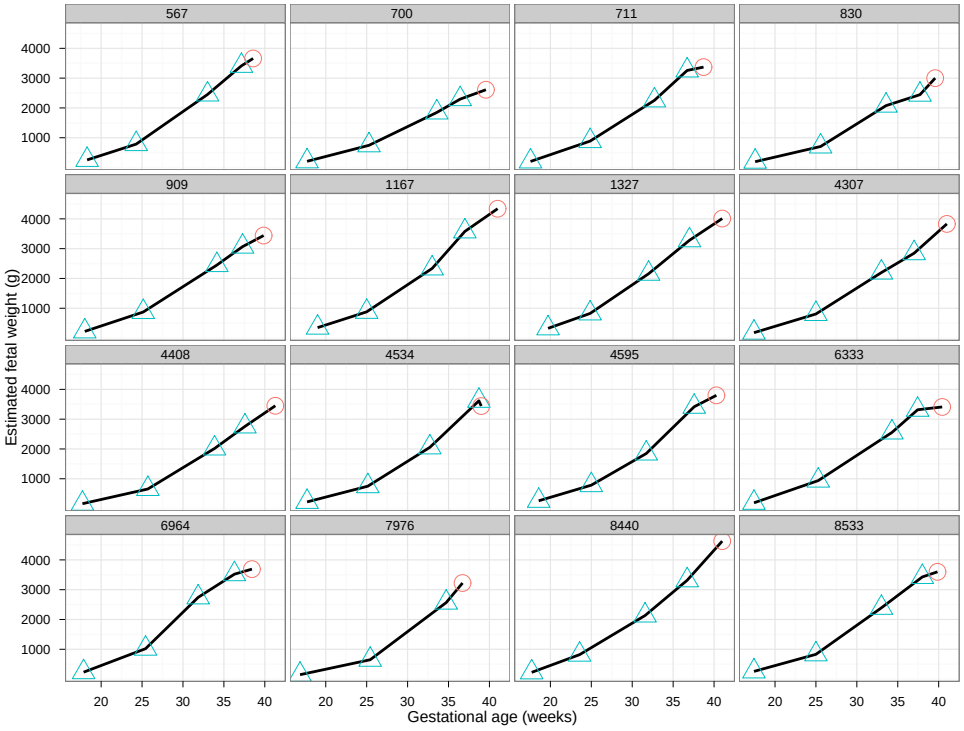

Figure 2: **Individual weight-for-gestational-age paths for males sampled from the Scandinavian dataset**. Circles represent weight/gestational age pairs at birth and triangles represent estimated weight/gestational age pairs in utero.

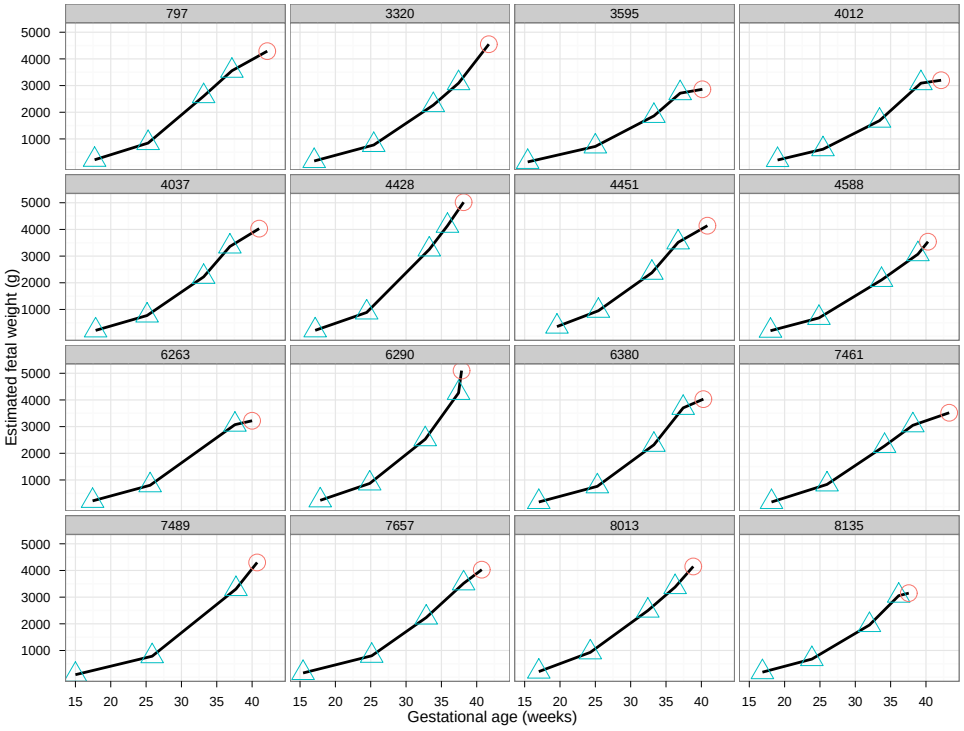

Figure 3: **Individual weight-for-gestational-age paths for females sampled from the Scandinavian dataset**. Circles represent weight/gestational age pairs at birth and triangles represent estimated weight/gestational age pairs in utero.

DOI: 10.2202/1557-4679.1305

Table 1: *Linear mixed model parameter estimates*

Para.	Estimates	SE	SD(Ran. eff.)
Intercept	−1.087	0.101	0.3521
GA_1	0.086	0.000	0.0018
GA_2	−0.012	0.002	0.0034
GA_3	−0.003	0.010	0.0000
GA_4	−0.127	0.033	0.0411
Sex	0.325	0.141	-
Sex*GA_1	−0.002	0.001	-
Sex*GA_2	0.010	0.003	-
Sex*GA_3	−0.037	0.014	-
Sex*GA_4	0.064	0.046	-
Number of groups: 451			
Number of observations: 2142			
Box-Cox parameter value: 0.2			
Log-likelihood: −1090.291			
Random effects covariance structure: Diagonal			
Residual covariance structure: σ^2I			
Residual standard deviation: 0.24			

Note: We used a variance-stabilizing transformation on the weights. The GA_i's represent gestational age reexpressed in the spline basis.

A summary of the fitted models is found in tables 1 and 2. A visual comparison of fitted curves for individual fetuses revealed that imposing non-zero off-diagonal elements in the random effects' variance/covariance matrix did not strongly affect the fitted values, even if a certain amount of covariance did exist between the random effects. Imposing a highly-parameterized structure on these estimators also led to convergence problems. Therefore, we selected a diagonal variance structure. With untransformed weights, a random intercept was unwarranted, because the weights of all fetuses were essentially identical prior to 10 weeks.

On the transformed scale, specifying a residual correlation structure with non-zero off-diagonal elements did not sizably change the parameter estimates or their standard deviations. Therefore, we imposed the assumption of uncorrelated residuals, which is reasonable considering that measurements are taken many weeks apart. On the untransformed scale though, the assumption that residuals have an $AR(1)$ covariance structure led to slightly different median estimates, but with noticeably lower standard deviations than when they were computed under the assumption of no correlation between the residuals.

Overall, transforming the data seemed to be beneficial. Although the median weight estimates from both fits, i.e. the one obtained by using the original weight values and the one obtained by using transformed weight values, were similar, the bootstrapping procedure revealed that those from the model fitted to transformed

Table 2: *Linear mixed model parameter estimates*

Para.	Estimates	SE	SD(Ran. eff.)
Intercept	−745.578	39.904	0.4497
GA_1	7.815	0.294	3.3468
GA_2	27.481	1.187	0.0020
GA_3	−33.919	4.429	95.5489
GA_4	−119.152	16.776	84.4450
Sex	66.511	55.811	-
Sex*GA_1	−0.546	0.412	-
Sex*GA_2	2.108	1.666	-
Sex*GA_3	−8.740	6.214	-
Sex*GA_4	−4.926	23.382	-
Number of groups: 451			
Number of observations: 2142			
Log-likelihood: −14013.518			
Random effects covariance structure: Diagonal			
Residual covariance structure: AR(1)			
Residual standard deviation: 84.445			

Note: The GA_i's represent gestational age reexpressed in the spline basis.

weight values had lower variance starting from approximately 272 days. It also revealed that transforming the data makes convergence of the linear mixed model components estimation procedure more likely. Indeed, the fitting procedure always converged when transformed weight values were used, whereas 28 runs out of 1028 (2.7%) led to a convergence error with untransformed data. For all these reasons, we retained only the model fitted to the dataset containing transformed weights for further comparisons.

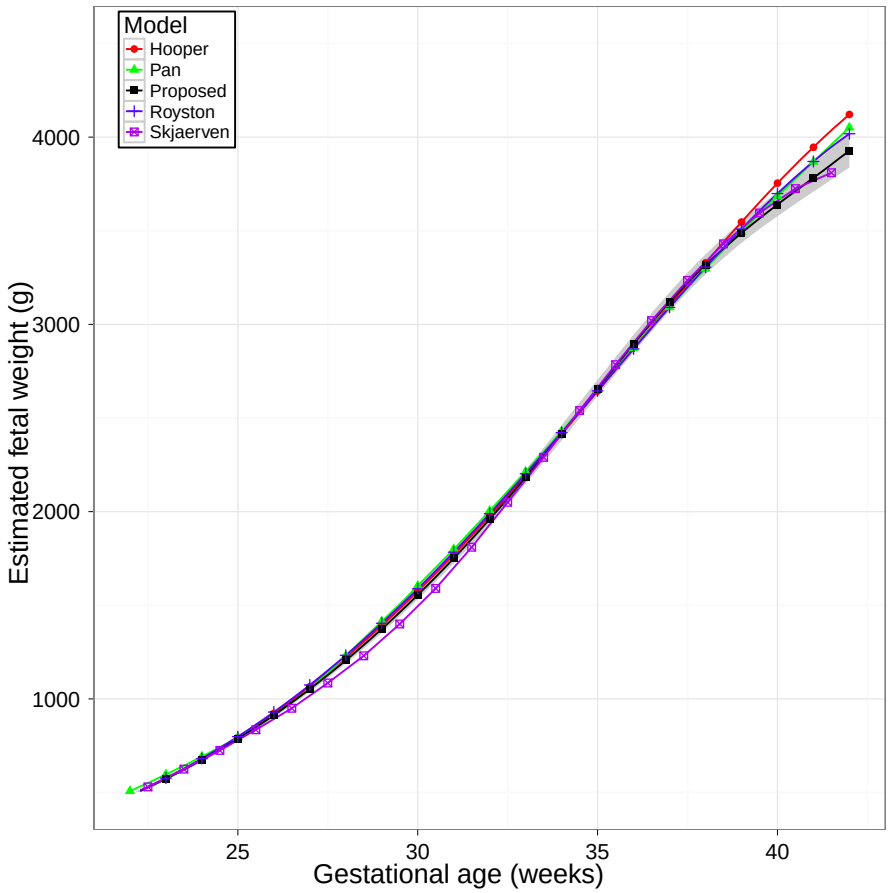

Figure 4: **Weight-for-gestational-age medians**. The shaded region represents the proposed model's 95 % pointwise confidence band.

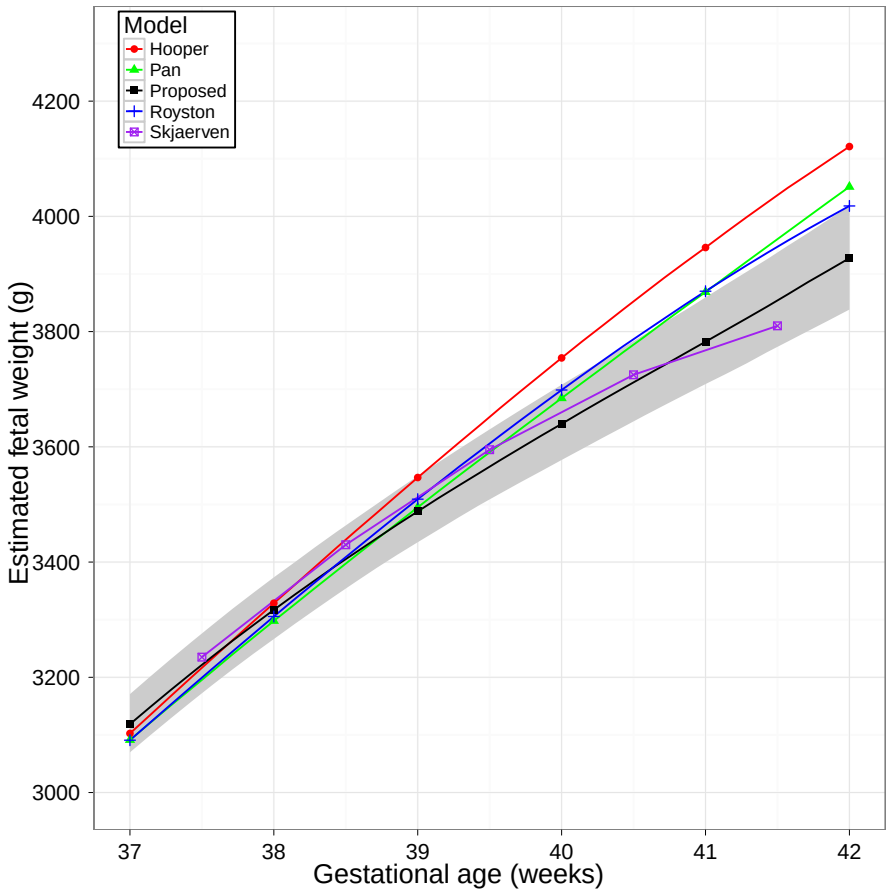

Figure 5: **Weight-for-gestational-age median trajectories**. The shaded region represents the proposed model's 95% pointwise confidence band.

DOI: 10.2202/1557-4679.1305

Table 3: *A comparison between the fitted medians (in grams) and their* 95% *confidence bands, and the means reported by Skjaerven et al. (2000)*

Gest. age	Royston	Hooper *et al.*	Pan and Goldstein	Proposed model	Skj. *et al.*
25.5	862 [846,883]	857 [841,874]	851 [836,867]	849 [836,864]	835
33.5	2312 [2274,2354]∗	2302 [2263,2343]∗	2320 [2286,2357]∗	2298 [2263,2341]∗	2290
36.5	2981 [2937,3028]∗	2988 [2944,3034]∗	2984 [2939,3030]∗	3009 [2962,3060]∗	3020
37.5	3199 [3154,3246]∗	3216 [3167,3269]∗	3196 [3149,3246]∗	3222 [3172,3276]∗	3235
38.5	3409 [3363,3458]∗	3439 [3378,3499]∗	3397 [3348,3453]∗	3405 [3354,3464]∗	3430
39.5	3606 [3557,3658]∗	3652 [3579,3721]∗	3591 [3534,3653]∗	3566 [3509,3630]∗	3595
40.5	3787 [3733,3841]	3852 [3764,3936]	3777 [3709,3849]∗	3712 [3644,3783]∗	3725
41.5	3947 [3885,4008]	4037 [3931,4136]	3960 [3874,4045]	3854 [3774,3937]∗	3810
42.5	4082 [4009,4150]	4203 [4077,4319]	4142 [4031,4247]	4001 [3904,4099]	3850

Note: Intervals followed by an asterisk contain the mean weight-for-gestational-age value reported in Skjaerven et al. (2000). We selected the midpoint of each week to make comparisons with the Skjaerven values more meaningful, as they are reported on a completed-week basis. The confidence intervals around the Skjaerven means are very narrow, and thus need not be reproduced.

5.3 Model comparisons

Results of sex-specific models were essentially similar; the results obtained using male fetuses are presented (results for females available upon request). Figure 4 and table 3 show that median estimates from the different methods remain very close. The Skjaerven et al. (2000) curve follows the anticipated trajectory: since its mean estimates are based on livebirths only, they fall below the other models' median estimates at preterm. Due to the positive skewness of the livebirth weight-for-gestational-age distributions at preterm, differences between medians would be even larger at that time. From conception to late term, confidence bands (based on 1000 bootstrap iterations) for all models overlapped, although differences in median estimates were larger beyond 41 weeks, as illustrated by figure 5. Starting at 40 weeks, medians obtained after fitting our model were the closest to the mean values reported by Skjaerven et al. (2000). Before 40 weeks, other estimates were occasionally closer, but the range between the maximum and minimum median estimates always remained small. Further, all 95% pointwise confidence bands between 32 and 40 completed weeks contained the means reported by Skjaerven et al. (2000).

Figure 5 and table 3 further show that with our method, as well as with that of Pan and Goldstein (1997), progression in weight for gestational age seems to become linear after 40 weeks. This would be at odds with the results of Skjaerven et al. (2000) though, which show a steady decline of the growth rate after 38 completed weeks. This decrease may have been due to errors in the estimation of gestational ages or selective delivery of fetuses at later gestational ages. Alternatively, since a restricted cubic spline imposes linearity after the last knot, it should

Table 4: *Linear mixed model parameter estimates*

Para.	Estimates	SE	SD(Ran. eff.)
Intercept	−1.097	0.100	0.4053
GA_1	0.086	0.000	0.0024
GA_2	−0.012	0.002	0.0033
GA_3	−0.002	0.009	0.0000
GA_4	−0.129	0.032	0.0486

Number of groups: 217
Number of observations: 1033
Box-Cox parameter value: 0.2
Log-likelihood: −560.351
Random effects covariance structure: Diagonal
Residual covariance structure: σ^2I
Residual standard deviation: 0.235

Note: We used a variance-stabilizing transformation on the weights. The GA_i's represent gestational age reexpressed in the spline basis. We only used the data pertaining to male fetuses.

be verified whether the assumption of linearity is in contradiction with the data, although sparseness limits inference. We originally placed the last knot at the 95th quantile, 287 days. Moving it to 294 days did not strongly affect the shape of the curve (data available upon request), indicating that the observed pattern was not solely an artefact due to the use of a restricted cubic spline.

Our model included an interaction term for the effect of sex on fetal weight for gestational age. Since the models against which ours was compared did not include such a term, they were fitted separately to male and female fetuses. Although the other models could be modified so as to take sex into account, we thought it better to select the method, namely stratification, that required the simplest assumptions, even though splitting the data in such a way can be seen as inefficient, since features of the distribution of weights are likely shared between the two groups (Johnsen et al., 2006). In this light, assuming at least a shared variance/covariance structure would not seem unreasonable. This is equivalent to our model's inclusion of sex-gestational age interaction. Indeed, it can be understood from model equation 1 that, on one hand, weight values obtained for individual male and female fetuses are assumed to share the same covariance structure and, on the other hand, the extent of the contribution of the fixed effect component to observed or estimated fetal weight varies differently through time based on fetal sex. Although the inclusion of additional data does not impact strongly the value or the precision of the median estimates from our model, it does affect the random effects variance structure (see tables 1 and 4, and figures 6 and 7 in the appendix). If this structure was of specific interest, then we would benefit from not stratifying.

DOI: 10.2202/1557-4679.1305

6 Discussion

In this article, we presented a new, flexible, sex-specific model for fetal weight for gestational age, and systematically compared the model to other proposed approaches. We established that at most gestational ages, the choice of model does not have a meaningful impact on the median estimates of fetal weight for gestational age, insomuch as the models' 95% pointwise confidence bands have a large degree of overlap. However, the greater accuracy of estimates of weight-for-gestational-age medians at late term of the proposed model is important, because information on fetal growth at post-term ages may inform clinical decision-making on induction of labour. Overestimation of the median is likely to lead to an overestimation of the SGA threshold as well, which would produce an overestimation of the number of pregnancies with abnormal fetal growth. This being said, due to the quickly diminishing number of data points after 40 completed weeks, obtaining low-variance estimates at very late ages becomes difficult. While focusing on the estimation of the 10th percentile would be valuable, larger sample sizes than are currently available would be needed to obtain reasonable accuracy. Estimation of the median weight for gestational age is nevertheless useful, since an understanding of normal fetal growth is needed to be able to define and identify abnormal growth.

The differences between median estimates from the different methods before late term remain small and are therefore unlikely to be of clinical significance. Therefore, the selection process can safely be guided by the model-specific features that make a given model best suited to the underlying characteristics or objectives of a study. For instance, Hooper et al. (2002) proposed a way to isolate the measurement error from the latent weight component of estimated fetal weight, as to allow for the derivation of latent weight percentiles. The proposed model also presents several advantages. Its flexibility, in comparison to polynomial regression models for instance, makes it an obvious candidate for modelling growth. Its parsimony also makes it very appealing. Further it readily offers a concise and straightforward parametrization of variance and covariance. The mixed spline regression model is commonly taught in statistics and epidemiology programs, and scholars are very familiar with its formulation, implications and limitations. It has become a very popular approach to handle non-linear relationships, such as the one between gestational age and fetal weight. For the first time, this standard approach has been systematically compared to methods especially tailored to the problem of fetal growth. Spline regression methods are readily available in most software packages (as compared to some of the specialized approaches used previously). The practical implication of this work for epidemiologists and statisticians interested in modeling fetal

growth is that these models mostly provide similar estimates of median weight-for-gestational-age, so that modeling strategies may be selected based on other criteria such as ease of use.

In general, fitting a model to a dataset in which the response variable has been transformed may produce results that are hard to interpret. For instance, fixed-effect coefficient values will not be expressible on the original scale and the retransformed mean will be biased. (Duan, 1983) However, percentile estimators, which are generally of greater relevance in the study of fetal growth and identification of intrauterine growth restriction, (McIntire et al., 1999) still remain unbiased. In this context, we are more interested in the estimation of median weight for gestational age than in interpreting the coefficients themselves.

Estimates of individual weight-for-gestational-age trajectories are strongly affected by measurement error in estimated fetal weight, which results from the use of a formula to estimate fetal weight from ultrasound biometric measurements, as well as operator-error at the time of ultrasound. Fortunately, this source of error has been shown to be mostly non-systematic (Dudley, 2005), so that while individual-level estimation of fetal weight may be error-prone, population medians should not be greatly affected. Uncertainty in gestational age is a second source of measurement error in longitudinal modeling of fetal weight. However, in our data, the gestational age estimates were validated by the use of both the ultrasound and the LNMP estimates, providing confidence that the measurement error on gestational age remains small on average.

Two different kinds of weight data are being used in the model-fitting process, namely ultrasound-based estimates and precise measurements made at birth. We would expect the coefficient of variation for ultrasound-based estimates to be higher, due to measurement error. Model variance estimates may also be affected. However, the Box-Cox transformation dampens changes in variance, irrespective of their patterns. Therefore, since we fitted models on the transformed scale, it seems unlikely that the median fits were significantly biased by the reduced variance of birth weight data.

Most models proposed to date were only adjusted for the influence of gestational age on fetal weight. It is worth considering if other adjustment terms in addition to sex should be considered. Parity, ethnicity and maternal BMI have all been shown to be significant predictors of fetal weight (Gardosi et al., 1995b,a, Mongelli and Gardosi, 1995). However, the clinical relevance of customization for maternal characteristics remains controversial (Hutcheon et al., 2008b,a). Another way to further personalize a weight-for-gestational-age curve would be to make it conditional on attained weight. Such an approach has been advocated by Royston (1995). A conditional estimate of weight becomes in essence an estimate of growth. However, temporal distance between successive weight measurements

DOI: 10.2202/1557-4679.1305

tends to dampen correlation and conceal the effects conditioning might have on the variance of individual predicted weights on the short term. Since successive measurements in our dataset are often more than four weeks apart, a weight estimate conditioned on a measurement taken less than a month earlier would essentially rely on an arbitrarily imposed covariance structure. Measurement error, too, would negatively impact the predictive power of an individual conditional weight estimate. Results by Hutcheon et al. (2010) indicate that the potential theoretical gains from conditioning are practically negligible when evaluated in the clinical setting.

While the goal of this study was to compare different methods for characterizing normal fetal growth, the methodologies examined in this paper could also be applied to the study of pediatric and adolescent growth. Further, the models could easily be modified to take other predictors into account. %For instance, maternal serum folate concentration in the second trimester and third trimester has been shown to be associated with birth weight (Goldenberg, Tamura, Cliver, Cutter, Hoffman, and Copper, 1992, Scholl, Hediger, Schall, Khoo, and Fischer, 1996) and therefore, measuring this predictor routinely during pregnancy might be warranted.

The dataset used in this study has several advantages, notably its relatively large sample size, the four longitudinal measurements resulting from unselected ultrasound scans and the quality of its weight and gestational age estimates. However, it would be of interest to be able to generate estimates applicable to different obstetrical populations, e.g. a North American population. Future work should determine whether the model can be adapted or rescaled to reflect fetal growth in different obstetrical populations.

Efforts are currently being made to update weight-for-gestational-age reference charts through the creation of a US national ultrasound standard for fetal growth (Zhang et al., 2010a). On the other hand, a new method could also be proposed to improve reference charts by combining data from difference sources that are already available such as data from routine clinical ultrasounds, serial ultrasound research studies, and population birthweight data. As both approaches will require selecting a model for median weight for gestational age, we believe that our work will be especially useful in this context.

Appendix A Royston's method

Royston (1995) proposed a multi-level model to predict weight for gestational age, with either a random intercept alone or with both a random intercept and a random

slope, as well as a flexible fractional polynomial functional form, and a Box-Cox transformation to correct for heteroscedasticity. The model is formulated,

$$W_{ij}^{(\lambda)} = \mu_i + \beta_i g(T_{ij}) + \varepsilon_{ij}, \tag{2}$$

where T_{ij} is the time of the jth measurement on the ith individual, $W_{ij}^{(\lambda)}$ is the weight value after a potential Box-Cox transformation with scale parameter λ, μ_i is the intercept for individual i, β_i is the slope for individual i, $g(.)$ is the fractional polynomial transformation and $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ is the residual term. The random-effect vector (μ_i, β_i) is assumed to follow distribution $MVN((\mu, \beta), \Sigma)$, with Σ being an arbitrary variance-covariance matrix. If the scale parameter is set to 1 then the transformation function reduces to identity.

Royston based model selection on a fixed-effect alternative to the previously specified linear mixed model, that is, one in which random effects are replaced by individual-specific non-random intercept and, if required, slope terms. The choice of the optimal λ is based on its maximum likelihood estimate in model

$$W_{ij}^{(\lambda)} = \mu_i + \beta_1 T_{ij} + \beta_2 T_{ij}^2 + \beta_3 T_{ij}^3 + \varepsilon_{ij}, \tag{3}$$

where μ_i is assumed fixed.

Allowing individual-specific slope terms might render the transformation of the response variable unneeded. However, there may be some lingering residual non-normality or heteroscedasticity. Because of this, four models are compared (all models are fixed-effects models and have at least subject-specific intercepts):

1. W linear in $g(T)$ with a common slope,
2. W linear in $g(T)$ with separate slopes for each subject,
3. $W^{(\hat{\lambda})}$ linear in $g(T)$ with a common slope,
4. $W^{(\hat{\lambda})}$ linear in $g(T)$ with separate slopes for each subject.

Since model 1 is nested in model 2 and model 3 is nested in model 4, they can be readily compared. In order to do this, a F-ratio is computed, as a test of non-parallelism. A significant value for this ratio means that subject-specific curves should not be assumed parallel and thus, that the separate-slope model should be adopted. Non-nested models cannot be readily compared, hence Royston's suggestion to use a pseudo-F-statistic to compare models 2 and 4. This test will verify whether the transformation is successful in reducing residual non-normality and/or heteroscedasticity. Based on the results of the previous steps, an analogous linear mixed model is fitted to replace the fixed-effects model that had previously been derived for selection purposes. In other words, individual-specific intercepts and, if needed, slopes will now be assumed to be correlated normal random variables (within-individual).

DOI: 10.2202/1557-4679.1305

Appendix B Hooper et al.'s method

In the model proposed by Hooper et al. (2002), weight for individual i obtained from the jth ultrasound examination, denoted W_{ij}, is expressed in grams and the corresponding gestational age, T_{ij}, is expressed in weeks. Estimated fetal weight measurements taken between 14 and 42 completed weeks are transformed into z-scores. The transformation is expressed as

$$z = f(w,t) = g(log(w) - q(t))/h(t), \tag{4}$$

$q(t)$ being an estimated quadratic curve obtained by fitting a weighted-least-squares regression model of $log(W_{ij})$ against T_{ij}, with statistical weights inversely proportional to the square root of the number of examinations per subject. The difference between $log(w)$ and $q(t)$ is the residual function. These residuals are transformed to approximate normal scores. This is done by fitting a linear spline with 6 knots to the normal probability plot of the residuals. In other words, theoretical quantiles from a standard normal distribution are regressed against residual quantiles. The spline function is called $g(.)$.

The function $h(t)$ is an estimate of the standard deviation of $g(r)$ at time t, where r corresponds to $log(w) - q(t)$. It is estimated by fitting a quadratic spline to $\{g(r_{ij})^2, t_{ij}\}$ pairs. The standard deviation corresponds to the square root of the fitted spline.

Weight is re-obtained by applying the inverse transformation,

$$w = f^{-1}(z,t) = exp(g^{-1}(h(t)z) + q(t)). \tag{5}$$

With this method, weight-for-gestational-age percentiles can be derived easily by setting z, assumed to follow a standard normal distribution, to the level corresponding to the required quantile.

A latent score is defined as a z-score stripped of its measurement error component. Let W_j denote weight obtained at gestational age t_j on an individual randomly sampled from the population. Let $Z_j = f(W_j, t_j)$ be the corresponding z-score. The joint distribution for $(Z_1, ..., Z_n)$, conditional on $(T_1, ..., T_n)$, is assumed to be multivariate normal with mean $\mathbf{0}$ and variance $\mathbf{1}$. To get latent score estimates, the z-score is reexpressed as

$$Z_j = L_j + U_j, \tag{6}$$

with L_j being the latent score component of Z_j and U_j being measurement error. It follows that

$$var(Z_1 - Z_2) = 2 - 2cov(L_1, L_2). \tag{7}$$

A model called *ALB* is recommended to estimate the covariance structure, now denoted by $c_L(t_1,t_2)$, with parameters in $c_L(t_1,t_2)$ estimated by minimizing the objective function

$$\sum d_{ijk}\{1-0.5(z_{ij}-z_{ik})^2-c_L(t_{ij},t_{ik})\}^2, \tag{8}$$

d_{ijk} being a weight value and $t_{ij} < t_{ik}$. After all parameters contained in $c_L(t_1,t_2)$ have been estimated, variance values can be derived by setting t_1 equal to t_2.

Prediction intervals at gestational age t with $1-\alpha$ coverage probability are bounded by $\pm z_\alpha\sqrt{c_L(t,t)}$, with z_α denoting the αth quantile of the standard normal distribution. A fetus is now considered small for gestational age if its latent score falls under the 10th percentile, i.e. under the lower bound of the 80% confidence interval.

Appendix C Pan and Goldstein's method

Pan and Goldstein (1997) developed a method to quantify growth in children, which involves modeling separately the mean, $M(t)$, the coefficient of variation, $S(t)$, and the Box-Cox power curve, $L(t)$. The method is called *LMS*, for λ ($L(t)$), μ ($M(t)$) and σ ($S(t)$). Maximum penalized likelihood estimation is used to derive the preceding function estimates. These functions belong to the family of cubic splines with knots at each distinct value of t. Percentiles are estimated with the formula

$$C_{100\alpha}(t) = M(t)(1+L(t)S(t)d_\alpha)^{1/L(t)}, \tag{9}$$

with d_α being the $(1-\alpha)$th percentile of the standard normal distribution. An equivalent z-score can thus be obtained,

$$Z(t) = \{(W/M(t))^{L(t)}-1\}/(L(t)S(t)), \tag{10}$$

W being a weight value. $Z(t)$ is referred to as an empirical *LMS* (or *ELMS*) score.

Under the assumption that the *LMS* procedure provides normally distributed scores, a two-level model is constructed:

$$Z_{ij} = \sum_{h=0}^{p}\beta_h T_{ij}^h + \sum_{h=0}^{q} u_{hj}T_{ij}^h + e_{ij}, \tag{11}$$

with

$$E = \{e_{ij}\} \sim N(0,\sigma_e^2 I),$$
$$U = \{u_{hj}\} \sim N(0,D_u),$$
$$i = 1,...,n_j; j = 1,...,m,$$

i being an index for measurements within individual j, n_j being the number of measurements for individual j and m being the number of individuals. This is a typical

DOI: 10.2202/1557-4679.1305

example of polynomial regression with random effects. The polynomial is set to be of degree p and the number of random effects is q. Residuals are assumed to be uncorrelated and homoscedastic. The authors note that this may not be reasonable if measurements are taken close in time.

If serial measurements are available, a norm for a new measurement conditional on the previous two, expressed as

$$Z_{3j} = \beta_1 Z_{1j} + \beta_2 Z_{2j} + \varepsilon_j, \tag{12}$$

with $\varepsilon_j \sim N(0, \sigma_\varepsilon^2)$, can be obtained. Z_{3j} can be standardized by subtracting from it the fixed part of the right-hand side of (12) and dividing the total by the residual variance. In other words, Z^*_{3j}, i.e. Z_{3j} after standardization, can be expressed as

$$Z^*_{3j} = \frac{Z_{3j} - \beta_1 Z_{1j} - \beta_2 Z_{2j}}{\sigma_\varepsilon}, \tag{13}$$

where σ_ε^2 is the residual variance in (12). The standardized estimates are found by substituting the Z_{ij}'s in (13). Corresponding percentiles can then be derived.

Appendix D Extra figures

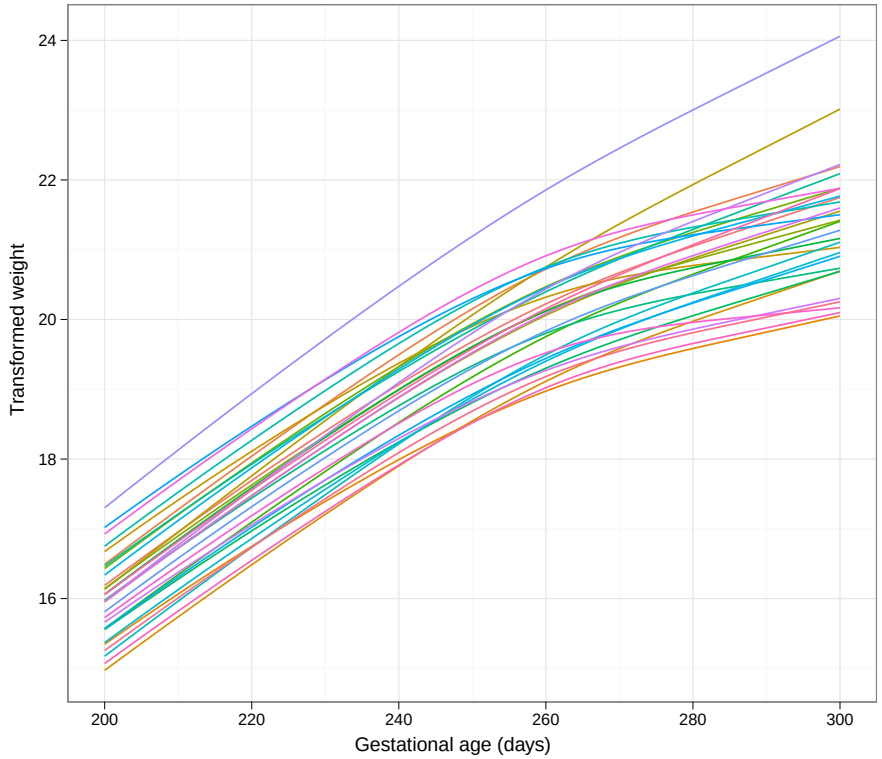

Figure 6: **Simulated individual weight-for-gestational-age trajectories**. The curves have been simulated based on a model stratified on the sex variable.

DOI: 10.2202/1557-4679.1305

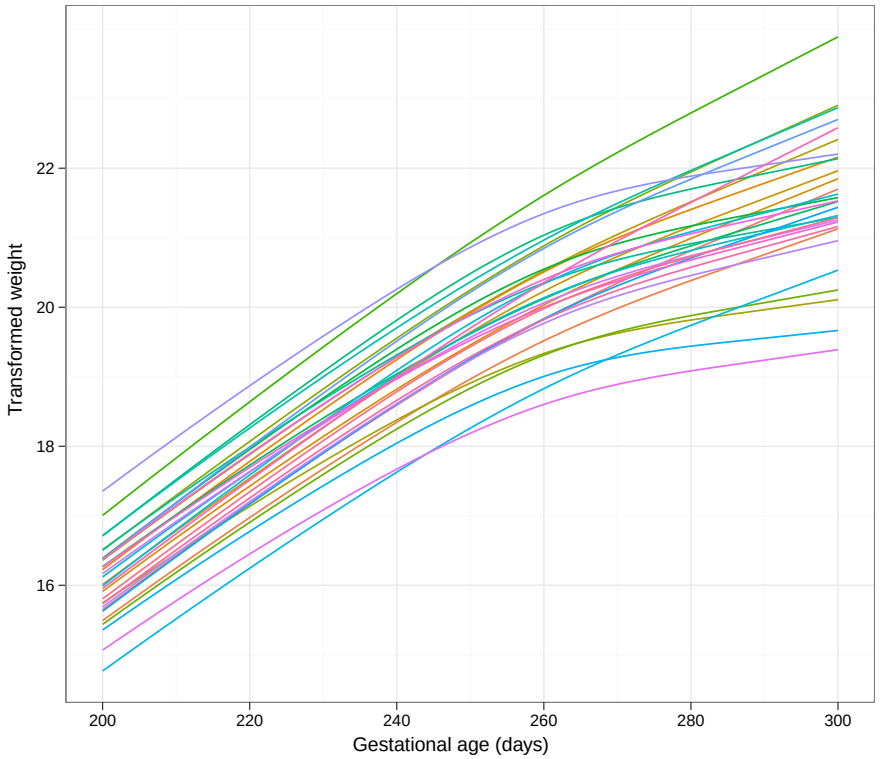

Figure 7: **Simulated individual weight-for-gestational-age trajectories**. The curves have been simulated based on a model with sex-gestational age interaction.

References

Alexander, G. R., J. H. Himes, R. B. Kaufman, J. Mor, and M. Kogan (1996): "A United States national reference for fetal growth." *Obstet Gynecol*, 87, 163–168.

Bakketeig, L. S., G. Jacobsen, H. J. Hoffman, G. Lindmark, P. Bergsjœ, K. Molne, and J. Rœdsten (1993): "Pre-pregnancy risk factors of small-for-gestational age births among parous women in scandinavia." *Acta Obstet Gynecol Scand*, 72, 273–279.

Bertino, E., E. D. Battista, A. Bossi, M. Pagliano, C. Fabris, G. Aicardi, and S. Milani (1996): "Fetal growth velocity: kinetic, clinical, and biological aspects." *Arch Dis Child Fetal Neonatal Ed*, 74, F10–F15.

Cole, T. J. and P. J. Green (1992): "Smoothing reference centile curves: the lms method and penalized likelihood." *Stat Med*, 11, 1305–1319.

Duan, N. (1983): "Smearing estimate: a nonparametric retransformation method," *J. Amer. Statist. Assoc.*, 78, 605–610.

Dudley, N. J. (2005): "A systematic review of the ultrasound estimation of fetal weight." *Ultrasound Obstet Gynecol*, 25, 80–89, URL `http://dx.doi.org/10.1002/uog.1751`.

Fry, A. G., I. M. Bernstein, and G. J. Badger (2002): "Comparison of fetal growth estimates based on birth weight and ultrasound references." *J Matern Fetal Neonatal Med*, 12, 247–252.

Gabbe, S. G., J. R. Niebyl, and J. L. Simpson (2007): *Obstetrics: Normal and Problem Pregnancies, 5th Edition*, volume 1, New York, N.Y.: Churchill Livingstone.

Gardosi, J., M. Mongelli, M. Wilcox, and A. Chang (1995a): "An adjustable fetal weight standard." *Ultrasound Obstet Gynecol*, 6, 168–174, URL `http://dx.doi.org/10.1046/j.1469-0705.1995.06030168.x`.

Gardosi, J. O., J. M. Mongelli, and T. Mul (1995b): "Intrauterine growth retardation." *Baillieres Clin Obstet Gynaecol*, 9, 445–463.

Goldenberg, R. L., T. Tamura, S. P. Cliver, G. R. Cutter, H. J. Hoffman, and R. L. Copper (1992): "Serum folate and fetal growth retardation: a matter of compliance?" *Obstet Gynecol*, 79, 719–722.

Gurka, M. J., L. Edwards, K. E. Muller, and L. L. Kupper (2006): "Extending the Box-Cox transformation to the linear mixed model," *J. Roy. Statist. Soc. Ser. A*, 169, 273–288.

Hadlock, F. P., R. B. Harrist, R. J. Carpenter, R. L. Deter, and S. K. Park (1984): "Sonographic estimation of fetal weight. the value of femur length in addition to head and abdomen measurements." *Radiology*, 150, 535–540.

Hadlock, F. P., R. B. Harrist, and J. Martinez-Poyer (1991): "In utero analysis of fetal growth: a sonographic weight standard." *Radiology*, 181, 129–133.

Harrell, F. (2001): *Regression Modeling Strategies, with Applications to Linear Models, Survival Analysis and Logistic Regression*, New York: Springer.

Hediger, M. L., T. O. Scholl, J. I. Schall, L. W. Miller, and R. L. Fischer (1995): "Fetal growth and the etiology of preterm delivery." *Obstet Gynecol*, 85, 175–182, URL `http://dx.doi.org/10.1016/0029-7844(94)00365-K`.

Hooper, P. M., D. C. Mayes, and N. N. Demianczuk (2002): "A model for foetal growth and diagnosis of intrauterine growth restriction." *Stat Med*, 21, 95–112.

Hutcheon, J. A., G. M. Egeland, L. Morin, S. J. Meltzer, and R. W. Platt (2010): "The predictive ability of conditional fetal growth percentiles," *Paediatric and Perinatal Epidemiology*, 24, 134–139.

Hutcheon, J. A., X. Zhang, S. Cnattingius, M. S. Kramer, and R. W. Platt (2008a): "Customised birthweight percentiles: does adjusting for maternal characteristics matter?" *BJOG*, 115, 1397–1404.

Hutcheon, J. A., X. Zhang, and R. W. Platt (2008b): "The benefits of customizing for maternal factors or the benefits of using an intrauterine standard at preterm ages?" *Am J Obstet Gynecol*, 199, e18–9; author reply e19–20.

Johnsen, S. L., S. Rasmussen, T. Wilsgaard, R. Sollien, and T. Kiserud (2006): "Longitudinal reference ranges for estimated fetal weight." *Acta Obstet Gynecol Scand*, 85, 286–297.

Källén, B. (1995): "A birth weight for gestational age standard based on data in the swedish medical birth registry, 1985-1989." *Eur J Epidemiol*, 11, 601–606.

Kramer, M. S., R. W. Platt, S. W. Wen, K. S. Joseph, A. Allen, M. Abrahamowicz, B. Blondel, and G. Bréart (2001): "A new and improved population-based canadian reference for birth weight for gestational age." *Pediatrics*, 108, E35.

Marsál, K., P. H. Persson, T. Larsen, H. Lilja, A. Selbing, and B. Sultan (1996): "Intrauterine growth curves based on ultrasonically estimated foetal weights." *Acta Paediatr*, 85, 843–848.

McIntire, D. D., S. L. Bloom, B. M. Casey, and K. J. Leveno (1999): "Birth weight in relation to morbidity and mortality among newborn infants." *N Engl J Med*, 340, 1234–1238.

Mongelli, M. and J. Gardosi (1995): "Longitudinal study of fetal growth in subgroups of a low-risk population." *Ultrasound Obstet Gynecol*, 6, 340–344, URL `http://dx.doi.org/10.1046/j.1469-0705.1995.06050340.x`.

Ott, W. J. (1993): "Intrauterine growth retardation and preterm delivery." *Am J Obstet Gynecol*, 168, 1710–5; discussion 1715–7.

Pan, H. and H. Goldstein (1997): "Multi-level models for longitudinal growth norms." *Stat Med*, 16, 2665–2678.

Royston, P. (1995): "Calculation of unconditional and conditional reference intervals for foetal size and growth from longitudinal measurements." *Stat Med*, 14, 1417–1436.

Scholl, T. O., M. L. Hediger, J. I. Schall, C. S. Khoo, and R. L. Fischer (1996): "Dietary and serum folate: their influence on the outcome of pregnancy." *Am J Clin Nutr*, 63, 520–525.

Schwärzler, P., J. M. Bland, D. Holden, S. Campbell, and Y. Ville (2004): "Sex-specific antenatal reference growth charts for uncomplicated singleton pregnancies at 15-40 weeks of gestation." *Ultrasound Obstet Gynecol*, 23, 23–29, URL `http://dx.doi.org/10.1002/uog.966`.

Skjaerven, R., H. K. Gjessing, and L. S. Bakketeig (2000): "Birthweight by gestational age in norway." *Acta Obstet Gynecol Scand*, 79, 440–449.

Weiner, C. P., R. E. Sabbagha, N. Vaisrub, and R. Depp (1985): "A hypothetical model suggesting suboptimal intrauterine growth in infants delivered preterm." *Obstet Gynecol*, 65, 323–326.

Zhang, J., U. Grewal, M. L. Hediger, J. F. Troendle, and C. Zhang (2010a): "The national standard for normal fetal growth," URL `http://www.nichd.nih.gov/about/org/despr/studies/preg/normfetalgrowth.cfm`.

Zhang, J., M. Merialdi, L. D. Platt, and M. S. Kramer (2010b): "Defining normal and abnormal fetal growth: promises and challenges." *Am J Obstet Gynecol*, 202, 522–528, URL `http://dx.doi.org/10.1016/j.ajog.2009.10.889`.

DOI: 10.2202/1557-4679.1305