STATISTICAL METHODS FOR CONSTRUCTING GESTATIONAL AGE-RELATED REFERENCE INTERVALS AND CENTILE CHARTS FOR FETAL SIZE

R. J. Silverwood[a]* and T. J. Cole[b]

[a]Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London

[b]MRC Centre for Epidemiology of Child Health, UCL Institute of Child Health, London

*Correspondence to: R. J. Silverwood, Medical Statistics Unit, Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. Richard.Silverwood@lshtm.ac.uk

INTRODUCTION

Many fetal size variables, for example head measurements, abdominal measurements and femur length, increase over the course of gestation. Reference intervals (RIs) and centile charts provide a means of assessing these measurements, at a given gestational age (GA) or across a range of GAs respectively, and are tools of great importance in clinical medicine.

RIs (sometimes, misleadingly, called 'normal ranges') represent the interval between a pair of symmetrically placed extreme centiles (such as the 5th and 95th for a 90% interval) of a size variable, denoted y, at a given GA. Centile charts plot the values of y corresponding to one or more RIs against the relevant GA over a range of GAs. In the field of fetal size, values which lie outside the RI are regarded as extreme and may indicate the presence of a disorder such as intrauterine growth restriction[1] or macrosomia[2]. More informative, however, than this forced dichotomy is the calculation of a value's centile position, or z-score, relative to the reference population, estimated from knowledge of the distribution of y at a given GA. For a given observation, the proximity of the centile position to 0% or 100% (alternatively the magnitude and sign of the z-score) is then a measure of how extreme the observation is compared to the reference data at that GA. A centile position above 50% (equivalently a positive z-score) signifies a measurement greater than average for that GA, and a centile position below 50% (or a negative z-score) one less than average.

Whilst recent years have seen the publication of a variety of strategies for the construction of RIs, incorrect methods have still been used for fetal measurements of all kinds[1]. The choice of suitable methodology in this field is especially crucial as inaccurate centiles may lead to false conclusions regarding the development of the fetus, resulting in suboptimal clinical care.

In the preceding article Anon *et al*[3] construct centile charts of the axial cerebellar hemisphere circumference (CHC) and area (CHA) through gestation using one such method, based upon regression modelling of both the mean and the standard deviation (SD) across gestational age (GA), as detailed by Altman and Chitty[4] and Royston and Wright[1].

It is the aim of the present article to further examine the statistical approach used by Anon *et al*, whilst taking a more general look at the problem of constructing GA-related RIs and considering alternative approaches to this problem. Techniques for longitudinal data, where each subject contributes repeated observations, as opposed to cross-sectional data, where they contribute only one, require a different approach and are not considered here. Further information on this area can be found in, for example, Royston and Altman[5] and Royston[6].

Whilst many of the techniques explored here could be, and indeed have been, used in the context of anthropometric measurements, the focus here is on applications in the field of fetal size.

THE GENERAL PROBLEM

Prior to the statistical analysis, many RIs and charts for fetal size are already flawed by weaknesses in study design. As with any study, the choice of an appropriate sample is of great importance. Whilst some published studies use routinely collected data, resulting in the inclusion of multiple observations on some fetuses, Altman and Chitty[4] note that these fetuses are likely to be those with clinical indication, introducing bias to the sample. They advocate collecting data specifically for the purpose of developing the RI, with each fetus being included only once.

Within this framework it is important to have as unselected a sample as possible because reference data should relate to 'normal' fetuses. Altman and Chitty[4] suggest it as reasonable to exclude fetuses subsequently found to have a congenital abnormality, though recommend the inclusion of neonatal deaths and fetuses large or small for dates at birth where this is not the case.

Maternal conditions which could affect fetal growth are also deemed reasonable exclusion criteria.

Whilst imprecise estimates of the RI will be obtained when the sample size of the dataset is too small[1] , it is not easy to accurately specify appropriate sample sizes. In particular, when interest is focused on the extreme centiles, as is often the case, several hundred observations may be necessary to obtain estimates at an appropriate level of precision.

There are a variety of available statistical approaches for the calculation of RIs, the most important of which are to be reviewed presently. The method needs to produce reference centiles which change smoothly with GA and provide a good fit to the data. Whilst clearly these requirements are essential, it is also preferable, for the sake of general usability and accessibility, to maintain as simple a statistical model as possible. Accordingly, the choice of approach must strike a balance between these conditions.

A further desirable feature are tools to calculate the relevant centile positions and z-scores for any further measurements, which again should be as user-friendly as possible in their application. Not only is the calculation of z-scores useful on an observation-specific level, they have also been shown to be instrumental in the assessment of chart comparison[7] and quality control[8].

MEAN AND SD MODEL

The statistical approach followed by Anon *et al*[3], here referred to as the 'mean and SD model', is one which has been found to be sufficiently general to cope with a wide range of fetal measurements available from ultrasound scanning[1].

Generally, under the assumption that at each GA the measurement of interest has a Gaussian (or normal) distribution with mean and SD that vary smoothly with GA, the centile curve at a given GA may be calculated by

$$\text{centile}_{GA} = \text{mean}_{GA} + K \text{ x SD}_{GA} \tag{1}$$

where mean_{GA} and SD_{GA} are, respectively, the mean and SD at the required GA, and K is the desired normal equivalent deviate (NED). The NED takes a value corresponding to the proportion of the standard normal distribution (with mean 0 and SD 1) lying to the left of it. For example, the 50th centile (with a proportion of 0.5 of the standard normal distribution to the left of it) has a NED of 0, whilst the determination of a 90% reference range (i.e. the 5th and 95th centile curves) would require $K=\pm1.645$.

The 'mean and SD model' approach aims to find functions that adequately represent how the mean and SD change with GA, allowing any desired centile curve to be readily calculated by appropriate choice of K.

Firstly the mean is modelled by fitting a polynomial curve to the raw data by means of least squares regression analysis. Royston and Wright[1] recommend the initial use of a cubic polynomial ($a + bt + ct^2 + dt^3$, where, for simplicity, GA is represented by t). If the cubic coefficient, d, is not significantly different from zero (approximately if d is less than twice its SD), a quadratic polynomial ($a + bt + ct^2$) should be fitted with the same assessment made of the quadratic coefficient, c. The process should be repeated until no further removal of terms is possible. Whilst quadratic or cubic curves will often give good fit to the data, Altman and Chitty[4] suggest the linear-cubic model ($a + bt + dt^3$) as a good alternative for fetal size data. It is also advocated that the choice of curve be based not only on statistical significance, with the

quality of fit to the data and aesthetic appearance, especially at the extremes of GA, also being of importance.

Anon *et al*[3] found a linear model ($a + bt$) to be sufficient for the CHC curve and a quadratic polynomial to be suitable for CHA.

Once a suitable mean model has been decided upon, attention can turn to the variability in the data. Residuals from the fitted mean model (observed value minus predicted value) should be calculated and plotted against GA to show if and how variability changes with GA[4].

Previously, modelling of the variability was not often considered, even though in the field of fetal size SD almost always changes with GA[9]. Whilst other methods have been proposed[10], the approach most frequently used is that of Altman[9]. It follows from the assumption that the variable under consideration is normally distributed at all GAs that the residuals from the mean model should also be normally distributed. This in turn means that the absolute residuals (residuals with the sign removed) have a half normal distribution. As the mean of a half standard normal distribution is ?(2/?), the mean of the absolute residuals multiplied by ?(?/2) is an estimate of the SD of the residuals. Hence if the SD is not reasonably constant over GA, predicted values from a regression of the absolute residuals on age multiplied by ?(?/2) will give age-specific estimates of the SD of the residuals, and hence of *y*.

An alternative formulation for Altman's approach favoured by Royston and Wright[1], and employed in this instance by Anon *et al*[3], is to produce 'scaled absolute residuals' (SARs) by multiplying the absolute residuals by ?(?/2). The SARs are then regressed on GA, the predicted values from which again estimate the SD of the residuals.

Under either formulation, if the absolute residuals, be they scaled or unscaled, show no trend with GA, the SD is estimated as the SD of the unscaled original residuals (observed value minus predicted value). If there is a trend, polynomial regression is needed to estimate an appropriate curve in the same way as for the mean. Altman[9] suggests it unlikely a curve more complex than quadratic is required for a satisfactory fit to the SD. Superimposing $\pm 1.645 \times SD$ on the residual plot is useful to see how well the SD had been modelled as approximately 90% of the observed residuals should fall within these limits.

Anon *et al*[3] found the CHC SARs to be suitably represented by a linear relationship with GA, whilst those for CHA required a cubic polynomial.

As the regression analysis to estimate the mean should really take into account any increase in SD with GA, at this juncture the mean model can be refit using the reciprocal of the square of the estimated SD as weights. However, Altman and Chitty[4] report that the effect of refitting is almost always rather small.

A useful tool in assessing model fit are z-scores (also known as SD scores), defined as

$$z = (\text{observed } y \text{ value - mean}_{GA}) / (SD_{GA}) \tag{2}$$

where mean_{GA} and SD_{GA} are, respectively, the mean and SD given by the model for the GA at which the observation is made. Hence z-scores represent the observed values expressed on a standard normal scale (with a mean of 0 and SD of 1), with the mean and SD adjusted for GA.

Altman and Chitty[4] recommend three methods of evaluation for the goodness of fit, all of which Anon *et al*[3] appear to have conducted. These methods will be illustrated using data on fetal biparietal diameter (BPD). A subset of 850 of the 19,647 fetuses analysed by Salomon *et al*[11] were

fit with a 'mean and SD model' in the standard manner, as outlined above, resulting in a cubic mean model and a linear SD model.

Firstly, a plot of the z-scores against GA should be checked for existence of any patterns. The z-scores should be randomly scattered about zero at all GAs, with any deviation from this indicating that the mean curve may require modification. This is shown in Fig. 1 for the example dataset, with the BPD z-scores appearing to adhere to this stipulation.

Secondly, a normal plot (essentially a scatterplot of the actual data values plotted against the 'ideal' values from a normal distribution) can be used to check that the z-scores have a close to normal distribution. This is signified by a roughly straight line but can be confirmed more formally using the Shapiro-Wilk W test or Shapiro-Francia W' test. Fig. 2 shows that in the example dataset the BPD z-scores do have a close to normal distribution and this is corroborated by both the Shapiro-Wilk W and Shapiro-Francia W' tests having P-values of 0.998.

Finally, the appropriate proportion of observations should fall between and outside fitted centiles, for example approximately 90% of z-scores should lie between $z = -1.645$ and $z = 1.645$. Deviation from this may imply that a higher order polynomial curve for the SD is needed. For the example dataset, lines corresponding to a BPD z-score of ± 1.645 have been plotted on Fig. 1. A brief examination suggests that approximately 90% of the data lie between the lines, with calculations confirming that 4.9% of the data lie below $z = -1.645$ and 4.2% above $z = 1.645$ (compared to an expected 5% for each). It is unlikely that the values are both *exactly* 5%, so figures such as these indicate an adequate level of fit.

This aspect of the data can be further examined in a plot such as in Fig. 3, a histogram of the z-scores with an overlaid standard normal distribution. If the model fits well then the histogram should match up with the standard normal distribution meaning that the expected and observed centiles lie at the same values. Given the sample size of the dataset, the histogram for the BPD data shows a close to standard normal distribution, indicating an adequate model fit.

Once a satisfactory model has been determined, the centile curves for the desired reference interval may be calculated by substituting the expressions for the mean and SD into equation (1). The z-score for any new individual may be calculated using equation (2) and its centile obtained using the inverse normal distribution. Finally, the calculated centiles should be superimposed on the scatter diagram of observed values against GA to ensure a suitable fit.

Beside the study currently under consideration, this approach to the construction of RIs has been widely used in the field of fetal measurements. Altman illustrated his absolute residual approach by developing reference centiles of fetal foot length[9]. Chitty *et al* constructed new charts for fetal head circumference, biparietal diameter and other head dimensions[12], fetal abdominal circumference and area[13], and fetal femur length[14]. Royston and Wright estimated RIs for fetal head circumference (using the same data as Chitty *et al*[12]), hemoglobin concentration and kidney volume[1]. Salomon *et al* constructed new reference charts and equations for fetal biparietal diameter, head circumference, abdominal circumference and femur length[11].

Extensions to the mean and SD model

Several extensions to the basic 'mean and SD model' approach described above have been posited as ways to improve the performance of the method. The use of logarithmic transformations and fractional polynomials are described below.

Mean and SD model with logarithmic transformation

Many size measurements tend to follow a skewed normal distribution at a given GA, usually a positive skew where the right tail of the distribution is longer than the left. Whilst this clearly conflicts with the assumption that at each GA the data come from a population with a normal distribution, it can often be overcome by the application of a logarithmic transformation. This same solution will also increase the ease with which a model can be fitted if the SD of the original measurements increase rapidly with GA.

Royston[10] suggests initially attempting to fit the mean model to the original measurements. If the residuals from this model show a positive skew then a logarithmic transformation should be performed on the original values, y, and the model refitted on $\log(y)$.

If residuals from the refitted model are once again skewed, it is then recommended to try using a modified logarithmic transformation of the form $\log(y + C)$, where C is positive if the new residuals are negatively skewed, and negative otherwise. A polynomial model of the same degree as the optimal model for $\log(y)$ is then repeatedly fitted, with the value of C varied until the highest (i.e. least significant) P value for the normality test of the residuals is reached. Often a value of C will be found that makes the distribution of residuals satisfactorily normal.

Once acceptable residuals from the mean model are obtained, the rest of the 'mean and SD model' fitting procedure is continued as before. However, it is important to back-transform the curves once the model is finalised using the antilog (exponential if a *natural* logarithmic transformation was used), also remembering to subtract C for a modified logarithmic transformation.

Whilst this simple procedure can easily cope with the problem of skewed data, Altman and Chitty[4] report that very few fetal size measurements require transformation.

The effect of the logarithmic transformation is illustrated here using data on birth weight in 58,940 neonates as analysed by Salomon *et al*[15]. Fig. 4, a scatterplot of birth weight against GA at birth, shows a marked increase in variability with GA and also suggests a slight positive skew to the data at a given GA. In Fig. 5 the birth weights have undergone a natural logarithmic transformation, resulting in a more constant variance over GA and any evidence of skew being removed. The fitting of a 'mean and SD model' on this transformed data should now be relatively more simple.

The modification of the 'mean and SD model' by the addition of a logarithmic transformation is somewhat less common than the unmodified version in the fetal size literature. Royston[10] used a modified logarithmic transformation in an example concerning fetal triglycerides. After fitting an initial quadratic mean model, positive skew was identified in the residuals. A logarithmic transformation was performed on the original values and a quadratic mean model fitted on $\log(y)$. However, this introduced negative skewness, so a modified logarithmic transformation was utilised. Wright and Royston[16], in an example regarding fetal abdominal circumference, also used a logarithmic transformation.

Mean and SD model using fractional polynomials

Fractional polynomials (FPs), formalised by Royston and Altman[17], extend the range of models afforded by conventional polynomials by allowing parameters to also take fractional powers. Whilst a conventional polynomial is of the form

$$a + bt + ct^2 + dt^3 + \ldots,$$

FPs are defined as

$$a + bt^{p1} + ct^{p2} + dt^{p3} + \ldots,$$

where p_1, p_2, p_3,… are chosen from a predetermined set, usually taken to be {-2, -1, -0.5, 0, 0.5, 1, 2, 3}. Here a value of -1 represents the inverse of t and 0.5 the square root of t. By convention the power 0 is defined to be log(t). If one or more power in the model is duplicated then the model will include 'repeated powers', whereby the second term is multiplied by log(t). As an example, a FP of degree 3 with powers (0, 2, 2) (i.e. p_1=0, p_2=2 and p_3=2) is of the form

$$a + b\log(t) + ct^2 + dt^2\log(t).$$

Estimation of the best fitting FP for a given dataset involves both a systematic search for the best power or combination of powers from the permitted set, and estimation of the associated parameter coefficients. This selection process includes fitting a model for each combination of powers in the permitted set. This means, for example, that fitting a fractional polynomial of degree 2 (i.e. of the form $a + bt^{p1} + ct^{p2}$) using the standard set detailed above would involve fitting a different model for each of the 36 permissible combinations of powers. From these models the one with lowest residual standard deviation is chosen to be optimal.

FPs give at least as good a fit to data as a conventional polynomial of corresponding degree and often offer a better fit than conventional polynomials of higher degree. Royston and Wright[1] recommend the use of FPs for modelling the mean or SD curve if a quartic or quintic polynomial is required for an adequate fit to the data.

Over recent years use of FPs in the construction of RIs has become more popular. Kurmanavicius *et al*[18] created ranges for biparietal diameter, occipito-frontal diameter, head circumference and cephalic index using this method, although in each case, bar the cephalic index SD, the best fitting fractional polynomial was found to be a conventional polynomial. Kurmanavicius *et al*[19] also modelled mean abdominal diameter, abdominal circumference and femur length using FPs, with only femur length SD taking a fractional model. Size charts for fetal bones (radius, ulna, humerus, tibia, fibula, femur and foot) were presented by Chitty and Altman[20] after fitting FPs, with all but one mean model, though none of the SD models, taking fractional form.

ALTERNATIVE METHODS

Beside the 'mean and SD model', Wright and Royston[16] report the other most widely applied statistical approaches for estimating GA-specific reference intervals in practice to be those of smoothed crude centiles[21] and LMS[22, 23, 24], detailed below.

Centile curves based on direct centile estimates

For a sufficiently large dataset (several hundred observations at each week of gestation according to Altman and Chitty[4]), one intuitive approach is to calculate empirical estimates for each desired centile at a given GA. Whilst the curves produced by joining these values will be rough, even for large sample sizes, smoother curves can be obtained by considering 'windows' of GAs instead of each GA separately. Here, increasing window size will increase smoothness, though information can easily be lost through oversmoothing[16].

A more formalised version of this approach, with a second stage involving centile smoothing based on the technique of Cleveland[25], is presented by Healy *et al*[21]. This approach makes no assumption about the nature of the distribution of measurements at a given GA but takes advantage of the knowledge that both the centiles themselves and the intervals between centiles at a fixed GA should behave smoothly.

In the first stage, observations are ordered by GA and the first k, where k usually represents 5-10% of the total data, selected. Initial empirical centile estimates at the required values, for example 5%, 10%, 25%, 50%, 75%, 90% and 95%, are calculated from these k measurements by sorting and counting, and then plotted against the median GA of the k observations. This 'window' of k observations is then moved on to encompass measurements 2 to $k+1$, then 3 to $k+2$, etc., with the same estimation procedure repeated on each occasion, until all observations have been included.

The initial centile estimates will be irregular, so the second stage smoothes them to provide more usable centile curves. It is first assumed that each centile curve can be approximated by a polynomial of degree p, so that y_i, the smoothed value of the ith centile, is given by

$$y_i = a_{0i} + a_{1i}t + a_{2i}t^2 + \ldots + a_{pi}t^p,\qquad(3)$$

where t again represents GA. Now consider the proportion corresponding to the ith centile (for example 0.5 for the 50th centile) and define z_i as its normal equivalent deviate (NED), similarly to before.

The coefficients a for a fixed j are then modelled as a polynomial in z_i, so that

$$a_{ji} = b_{j0} + b_{j1}z_i + \ldots + b_{jqj}z_i^{qj},\qquad(4)$$

where the degree q_j of the polynomial may differ from one value of j to another. This restricts the distance between centiles and prevents the resulting curves from crossing. Combining equations (3) and (4) gives a linear model for the centile values which can be fitted by least squares regression. It follows that for any observation a corresponding z-score can be calculated by solving a polynomial equation, though the order of the polynomial may realistically prohibit this. Goodness of fit should be judged by counting the points falling between adjacent centiles.

This method was applied by Wright and Royston[16] to measurements of fetal abdominal circumference and provided an adequate fit.

LMS

The LMS method, introduced by Cole[22, 23] and refined by Cole and Green[24], provides a general method for fitting smooth centile curves to reference data. It utilises the power transformation family of Box and Cox[26] to allow the skewness of the measurement distribution, as well as the median and variability, to vary with age. These three features of the distribution are summarised by the parameters *?*, *?* and *?*, the initials of which (L, M and S) give rise to the name of the method. The original form[22, 23] necessitated age to be split into groups - an arbitrary procedure whereby different groupings would produce different centile curves. This subjective stage was removed by Cole and Green[24] through the addition of a nonparametric aspect. Due to the superiority of the later version, only this is detailed here.

As previously asserted, many size measurements follow a skewed normal distribution. The

use of a suitable power transformation, which stretches one tail of the distribution and shrinks the other, can remove this skewness and 'normalise' the data. One such family of transformations, proposed by Box and Cox[26], is used in the LMS method, with the optimal power at a given age calculated from the data to completely remove skewness in the distribution. As skewness changes with age, the calculated power also changes.

Given a variable of interest y with median μ and a power transformation so that y^λ (or $\log(y)$ if $\lambda = 0$) is normally distributed, we consider the transformed variable

$$
\begin{aligned}
x &= ((y/\mu)^\lambda - 1)/\lambda && \text{if } \lambda \neq 0 \\
&= \log(y/\mu) && \text{if } \lambda = 0
\end{aligned}
\tag{5}
$$

based on the Box-Cox transformation[26]. This transformation maps the median μ of y to $x = 0$ and is continuous at $\lambda = 0$. For $\lambda = 1$ the SD of x is the coefficient of variation (CV) of y, and this remains approximately true for all moderate values of λ[24]. The optimal value of λ now minimises the SD of x.

Denoting the SD of x (and CV of y) by σ, the z-score (or SD score) of x (and hence y) is given by

$$
\begin{aligned}
z &= x / \sigma \\
&= ((y/\mu)^\lambda - 1)/\lambda\sigma && \text{if } \lambda \neq 0 \\
&= (\log(y/\mu))/\sigma && \text{if } \lambda = 0
\end{aligned}
\tag{6}
$$

and is assumed to take a standard normal distribution.

Assume that the distribution of y varies with GA t and that λ, μ and σ at t are read off smooth curves $L(t)$, $M(t)$ and $S(t)$. Then

$$z = ((y / M(t))^{L(t)} - 1) / (L(t)S(t)) \qquad \text{if } L(t) ? 0 \qquad\qquad (7)$$
$$(\log(y / M(t))) / S(t) \qquad\qquad \text{if } L(t) = 0.$$

Rearranging equation (7) shows that centile 100*?* of *y* at *t* is given by

$$C_{100?}(t) = M(t)[1 + L(t)S(t)z_?]^{(1 / L(t))} \qquad \text{if } L(t) ? 0 \qquad\qquad (8)$$
$$M(t)\exp(S(t)z_?) \qquad\qquad \text{if } L(t) = 0,$$

where $z_?$ is the normal equivalent deviate of size *?*. This shows that if *L*, *M* and *S* are smooth, then so are the centile curves.

Cole and Green[24] then introduce a penalised likelihood function, derived from equation (7), with three integrals providing roughness penalties for the curves *L(t)*, *M(t)* and *S(t)*. The extent of these penalties, and hence the smoothness of the curves, are controlled by three smoothing parameters, and these are the only parameters requiring specification in order to fit the model. However, 'equivalent degrees of freedom' (EDF), calculated for each fitted curve as a function of these smoothing parameters, give a more useable measure of the extent of the smoothing.

The illustrative examples of Cole and Green[24], although not from the field of fetal measurements, show values of the *L* curve falling well below zero. This indicates the presence of considerably more skew that a log transformation would remove and the extent of variability of the *L* curves with age reinforces the notion that transformation using a single power for all ages is inappropriate.

Whilst examples of the application of the LMS method for fetal size do not abound, using the same fetal abdominal circumference data as Chitty *et al*[13], Wright and Royston[16] used this approach to fit centile curves to good effect.

DISCUSSION

There are several viable methods available, of varying complexity, for constructing age-related RIs and centile charts. Ideally, methods should be understandable by clinicians, and the results easy to use, even without a statistical computer package. It is desirable that any published method should provide the potential user with the means to calculate the corresponding z-score and centile for a given measurement. The mere provision of a mean model or centile chart, regardless of the quality, is not really adequate. Any approach must also be sufficiently flexible to be applied successfully to many sets of data. Unfortunately, none of the methods currently available fulfil all these criteria, so it is unlikely that any one would be appropriate in all circumstances.

In the simplest setting, if it is plausible that the observed measurements at each GA do indeed come from a population with a normal distribution and, in addition, the variance across the age range is constant, then the use of conventional polynomial regression may be justified. However, the strict adherence to these assumptions is unlikely, meaning that the model may not produce sufficiently reliable reference intervals. Slightly more realistic is the acknowledgement that variance is likely to change over the age range. This feature can be included by fitting the 'mean and SD model' as described previously, though again the assumption of an underlying normal distribution is not always tenable. This issue can often be dealt with by the addition of a (modified) logarithmic transformation prior to the model fitting to correct any skew (distribution

asymmetry). However, this approach still suffers from the well-known limitations of polynomial curve shapes. This last hurdle can be overcome by the relaxation of the restrictions imposed on the powers of the polynomial, allowing the use of fractional polynomials (FPs). As FPs give at least as good a fit to data as a conventional polynomial of corresponding degree, and as the fitting of FPs with most basic statistical software is relatively straightforward, there seems little reason to not adopt them as standard.

All of these variations on the 'mean and SD model' benefit from being relatively conceptually simple and easy to use, with the necessary techniques available in most basic statistical packages. The resulting centile curves and z-scores can be expressed as explicit formulae meaning that the centile position of any individual is easily obtainable. Whilst the method as described here is adequate for most fetal measurements, there are some cases that cannot be handled properly by this approach. It is important to emphasise the strong assumption that at each GA the data come from a population with a normal distribution. Whilst skewed data may sometimes be corrected by a log transformation, this is not always successful, with time-varying skewness especially difficult to accommodate. Even after transformation, kurtosis (a non-normal distribution shape) may remain in the data, again in contravention of the assumption. Variables with a complex curve shape beyond those available from conventional (or even fractional) polynomials may also require alternative techniques.

The method of producing centile curves based on empirical centile estimates as described by Healy *et al*[21] makes no assumption about the nature of the distribution of measurements at a fixed GA, which is an appealing feature. This approach provides a flexible way of constructing centile curves which is capable of handling many patterns of growth due to the lack of a pre-specified functional form. However, there are some drawbacks. Experience is needed to find the best ways of choosing the values of the adjustable parameters involved, and clearly there is some degree of subjectivity here. The estimation of the centile values of further observations is not simple unless a very basic model has been fitted. There is also some vulnerability to outlying values affecting the derived centile values. We concord with the conclusion of Altman and Chitty[4] that this is not a suitable method for the derivation of fetal size charts, except when other methods are unsuccessful.

The LMS method with penalised likelihood[24] is extremely flexible and widely applicable[16]. It is usually easy to produce convincing centile curves, regardless of the complexity of the curve shape, and time-varying skewness is easily dealt with. It also has the appealing byproduct of the L, M and S curves which completely summarise the measurement's distribution over the age range and facilitate further investigation into the underlying structure of the data. Penalised likelihood provides an elegant solution to rid the earlier method of its arbitrary categorisation, with the smoothing of the three curves becoming an integral part of the likelihood maximisation. Now the only arbitrariness in the procedure is the choice of the three smoothing parameters.

There are, however, some general problems with the smoothing approach. Where data is more sparse near the ends of the age range, 'edge effects' (spurious changes in the centiles) may be observed, though this can be avoided by truncating the data at each end. One major drawback of nonparametric estimators is the lack of a succinct formula with which to estimate further centile values. This means that centiles may only be displayed graphically or in tabular form. Finally, the assumption of normality following the Box-Cox transformation may be violated by the presence

of kurtosis, which the transformation does not adjust for.

A more recently proposed generalisation of the LMS approach, the LMSP method of Rigby and Stasinopoulos[27], uses the Box-Cox power exponential (BCPE) distribution to try to overcome the issue of kurtosis. A fourth parameter is introduced in the power transformation in order to account for the observed kurtosis in the distribution and centile estimation proceeds in a manner not dissimilar to that of the conventional LMS method.

Whilst for the first-time user, application of the LMS method may appear a daunting task, the advent of specially designed programs such as Cole and Pan's LMSChartmaker[28], as well as packages for the widely used general statistical programs, mean that with brief instruction this need not be the case.

Wright and Royston[16] advise that a 'simple formula' to allow estimation of centile position for an individual is extremely valuable. If, when considering the statistical approach to follow in light of requirements specific to the data under analysis, this requirement is deemed to be essential, then this would exclude both the LMS method and any approach based on empirical centile estimates. Of the methods examined here, this leaves the only the parametric approach of the 'mean and SD model'. So the choice of approach is really reduced to the trade-off between the simplicity, usability and accessibility of the inferior model provided by the parametric approach, and the superior but less user-friendly model provided by the LMS method.

REFERENCES

1. Royston P, Wright E M. How to construct 'normal ranges' for fetal variables. *Ultrasound Obstet Gynecol* 1998; **11**: 30-8.

2. Sokol RJ, Chik L, Dombrowski MP, Zador IE. Correctly identifying the macrosomic fetus: improving ultrasonography-based prediction. *Am J Obstet Gynecol* 2000; **182**: 1489-95.

3. Anon. Nomograms of the axial fetal cerebellar hemisphere circumference and area throughout gestation. *Ultrasound Obstet Gynecol* (in press).

4. Altman DG, Chitty LS. Charts of fetal size: 1. Methodology. *Br J Obstet Gynaecol* 1994; **101**: 29-34.

5. Royston P, Altman DG. Design and analysis of longitudinal studies of fetal size. *Ultrasound Obstet Gynecol* 1995; **6**: 307-12.

6. Royston P. Calculation of unconditional and conditional reference intervals for foetal size and growth from longitudinal measurements. *Stat Med* 1995: **14**: 1417-36.

7. Salomon LJ, Bernard JP, Duyme M, Buvat I, Ville Y. The impact of choice of reference charts and equations on the assessment of fetal biometry. *Ultrasound Obstet Gynecol* 2005; **25**: 559-65.

8. Salomon LJ, Bernard JP, Ville Y. Analysis of Z-score distribution for the quality control of fetal ultrasound measurements at 20-24 weeks. *Ultrasound Obstet Gynecol* 2005; **26**: 750-4.

9. Altman DG. Construction of age-related reference centiles using absolute residuals. *Stat*

Med 1993; **12**: 917-24.

10. Royston P. Constructing time-specific reference ranges. *Stat Med* 1991; **10**: 675-90.

11. Salomon LJ, Duyme M, Crequat J, Brodaty G, Talmant C, Fries N, Althuser M. French fetal biometry: reference equations and comparison with other charts. *Ultrasound Obstet Gynecol* 2006; **28**: 193-8.

12. Chitty LS, Altman DG, Henderson A, Campbell S. Charts of fetal size: 2. Head measurements. *Br J Obstet Gynaecol* 1994; **101**: 35-43.

13. Chitty LS, Altman DG, Henderson A, Campbell S. Charts of fetal size: 3. Abdominal measurements. *Br J Obstet Gynaecol* 1994; **101**: 125-31.

14. Chitty LS, Altman DG, Henderson A, Campbell S. Charts of fetal size: 4. Femur length. *Br J Obstet Gynaecol* 1994; **101**: 132-5.

15. Salomon LJ, Bernard JP, De Stavola BL, Kenward MG, Ville Y. Birth weight and size in France. Charts and equations. *J Gynecol Obstet Biol Reprod (Paris)* (in press).

16. Wright EM, Royston P. A comparison of statistical methods for age-related reference intervals. *J R Statist Soc A* 1997; **160**: 47-69.

17. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Appl Statist* 1994; **43**: 429-67.

18. Kurmanavicius J, Wright EM, Royston P, Wisser J, Huch R, Huch A, Zimmermann R. Fetal ultrasound biometry: 1. Head reference values. *Br J Obstet Gynaecol* 1999; **106**: 126-35.

19. Kurmanavicius J, Wright EM, Royston P, Zimmermann R, Huch R, Huch A, Wisser J. Fetal ultrasound biometry: 2. Abdomen and femur length reference values. *Br J Obstet Gynaecol* 1999; **106**: 136-43.

20. Chitty LS, Altman DG. Charts of fetal size: limb bones. *BJOG* 2002; **109**: 919-29.

21. Healy MJ, Rasbash J, Yang M. Distribution-free estimation of age-related centiles. *Ann Hum Biol* 1988; **15**: 17-22.

22. Cole TJ. Fitting smoothed centile curves to reference data. *J R Statist Soc A* 1988; **151**: 385-418.

23. Cole TJ. The LMS method for constructing normalized growth standards. *Eur J Clin Nutr* 1990; **44**: 45-60.

24. Cole TJ, Green PJ. Smoothing reference centile curves: the LMS method and penalized likelihood. *Stat Med* 1992; **11**: 1305-19.

25. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc* 1979; **74**: 829-36.

26. Box GEP, Cox DR. An analysis of transformations. *J R Statist Soc B* 1964; **26**: 211-52.

27. Rigby RA, Stasinopoulos DM. Smooth centile curves for skew and kurtotic data modelled using the Box-Cox power exponential distribution. *Stat Med* 2004; **23**: 3053-76.
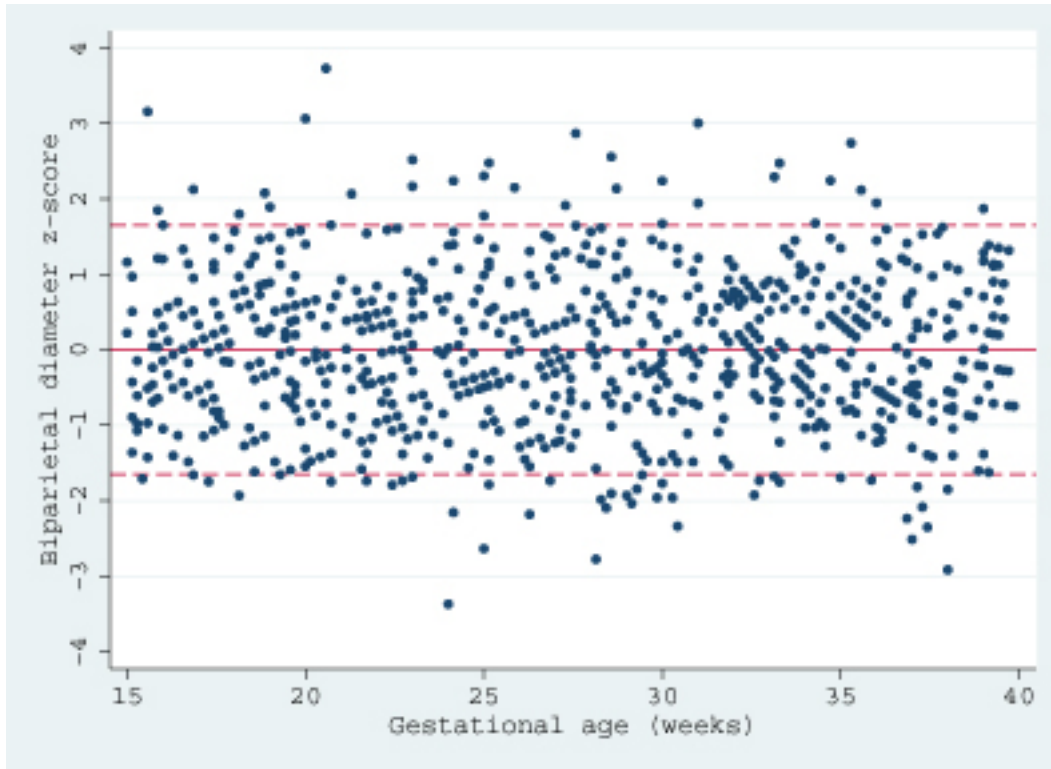
28. Cole TJ, Pan H. LMSChartmaker. 2005. http://www.healthforallchildren.co.uk.

Figure 1. Plot of calculated Z-scores against gestational age in the example dataset.

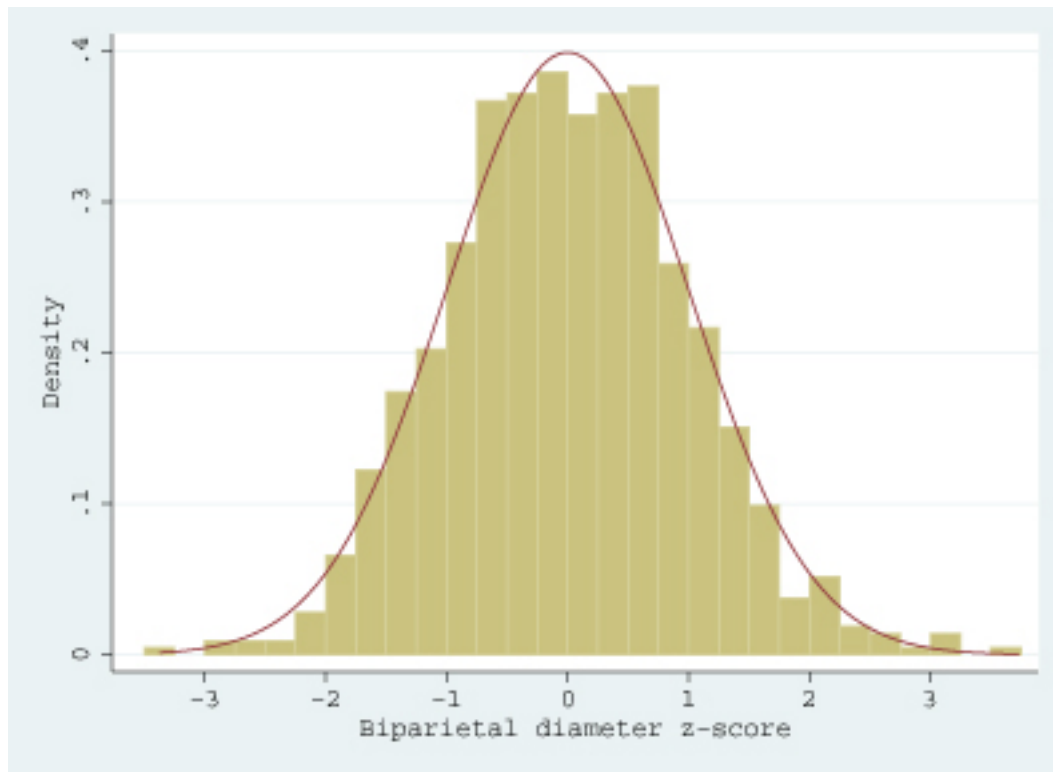**Figure 2. Normal plot of calculated Z-scores in the example dataset.**

Figure 3. Histogram of calculated Z-scores in the example dataset with overlaid standard normal distribution.
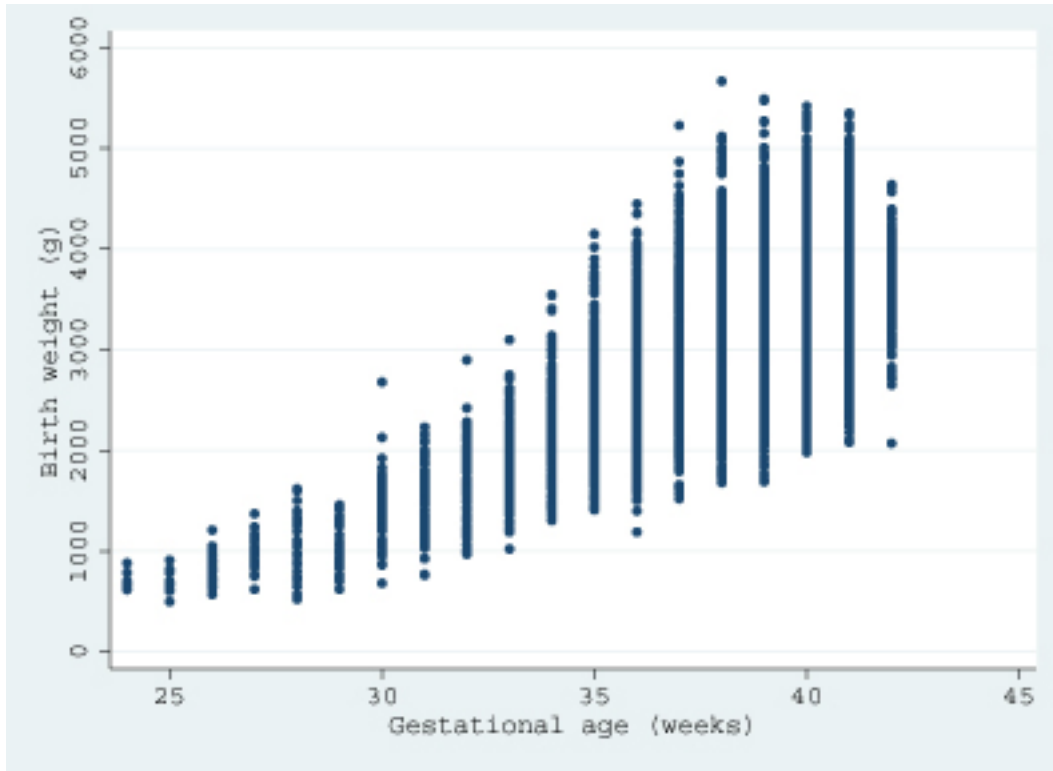
Figure 4. Scatterplot of fetal birth weight against gestational age in the example dataset.
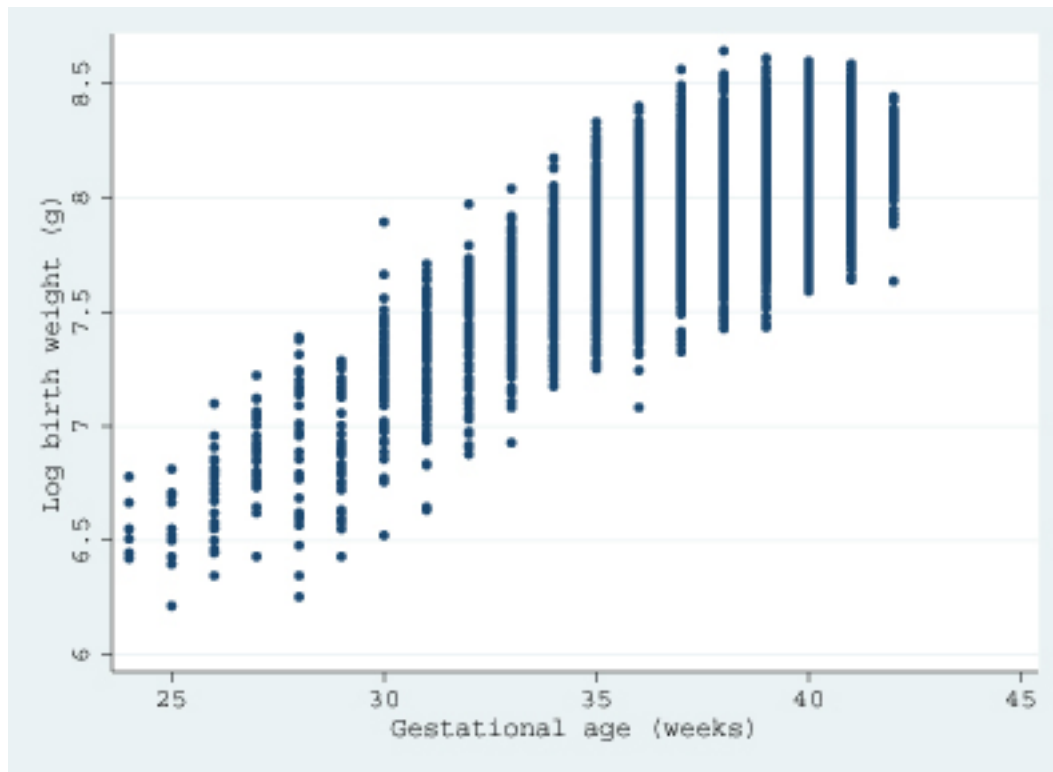
Figure 5. Scatterplot of logarithmically transformed fetal birth weight against gestational age in the example dataset.