*J. R. Statist. Soc.* A (1997)
**160**, *Part* 1, *pp.* 47–69

# A Comparison of Statistical Methods for Age-related Reference Intervals

By EILEEN M. WRIGHT† and PATRICK ROYSTON

*Royal Postgraduate Medical School, London, UK*

[Received October 1995. Revised May 1996]

SUMMARY

Age-specific reference intervals are commonly used in the routine monitoring of individuals, where interest lies in the detection of extreme values, possibly indicating abnormality. Here, a review is given of the wide range of statistical techniques which have been proposed for the construction of these intervals and issues such as the estimation of confidence bands and goodness of fit are discussed. Three methods, thought to be the most widely applied approaches, are considered in more detail. Comparisons are made on the basis of reference interval estimation for three real data sets.

*Keywords*: CENTILE ESTIMATION; DENSITY ESTIMATION; NONPARAMETRIC QUANTILE REGRESSION; NORMAL DISTRIBUTION; NORMALIZING TRANSFORMATIONS; NORMAL RANGES; REFERENCE INTERVALS

## 1. INTRODUCTION

The reference interval is a tool of some importance in clinical medicine. For a random variable $Y$, it represents the interval between two predetermined centiles (such as the fifth and 95th) of the distribution of $Y$. The population from which $Y$ is drawn is known as the reference population and in medicine is usually (though not always) selected or presumed to comprise 'normal' or 'healthy' people — hence the commonly used, but potentially misleading, phrase 'normal range'. An individual who is being screened for some disorder according to their observed value of $Y$ is judged in relation to the reference population in different ways. In simple terms, abnormality may be suspected if their $Y$ lies below the lower reference limit or above the upper limit. More informatively, their centile position relative to the reference population is estimated from knowledge of the distribution of $Y$. The proximity of the centile position to 0% or 100% is a measure of how extreme the individual's observation is. There is a major statistical difference between the approaches, for in the former case only certain quantiles must be estimated, whereas in the latter the whole distribution function of $Y$ must be approximated, which is more challenging.

It is common for the distribution of $Y$ to be affected by characteristics of the reference population such as age, sex, smoking habit, weight, height and genetic factors. We might wish to construct reference intervals for all such implied subpopulations, but this is clearly impractical. For many clinical and anthropometric variables the major influence is the age (and to a lesser extent the sex) of the individual, and the importance of allowing for age has long been recognized. Historically the construction of 'growth charts' was probably the first example of

†*Address for correspondence*: Department of Medical Statistics and Evaluation, Royal Postgraduate Medical School, Hammersmith Hospital, Du Cane Road, London, W12 0NN, UK.
E-mail: ewright@rpms.ac.uk

 0964–1998/97/160047

age-related reference intervals; see Cole's (1993) review for further details. More recently, following the pioneering work of Campbell and Newman (1971), the ultrasonographic assessment of fetal growth has become clinically routine, and many researchers have proposed gestational-age-specific centile charts and tables for a variety of relevant measurements (e.g. Chitty *et al.* (1994a, b, c)). Reference intervals are also commonplace in clinical chemistry and in clinical epidemiology, e.g. in population screening for heart disease and stroke according to risk factors such as serum cholesterol concentration and blood pressure. Harris and Boyd (1995) gives a useful overview of methods applied to data sets from these areas of research and clinical practice, with both the homogeneous and the age-specific cases being discussed. Fig. 1(a) shows a scatterplot of serum total cholesterol against age in a random sample of 502 normal men aged between 26 and 64 years in a study investigating cardiovascular risk factors (Mann *et al.*, 1988). Using suitable methods, the data could be analysed to provide age-specific reference intervals.

The aim of this paper is to discuss statistical methods for estimating age-specific reference intervals. Three methods are compared (Royston, 1991; Cole, 1988; Healy

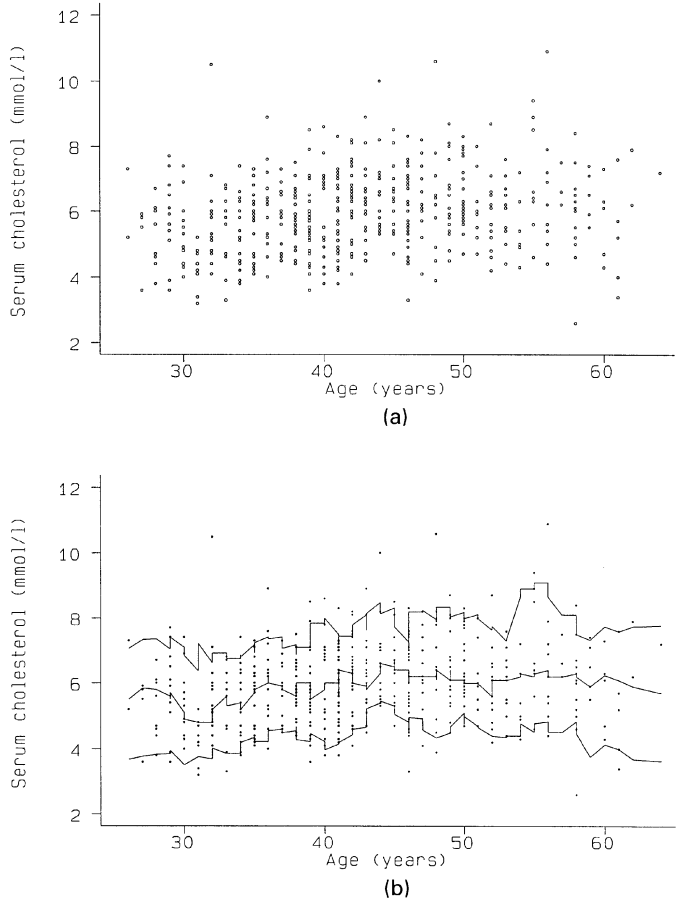


Fig. 1. (a) Scatterplot of serum cholesterol *versus* age; (b) 10th, 50th and 90th empirical centile curves for serum cholesterol

*et al.*, 1988) — these are probably the most widely applied approaches in practice and of most interest to those analysing reference data. The methods are assessed on their fit of estimated age-specific reference intervals to three data sets. These data sets were chosen to cover a range of situations, where the relationship between the measurement of interest and age is weak, strong (but simple in shape) and complex. Bonellie and Raab (1996) compared extreme low centiles for the methods described in Cole (1988), Healy *et al.* (1988) and a polynomial regression approach (similar to Royston (1991)) using a loss function. The aim was to assess the accuracy of possible models for the determination of intra-uterine growth retardation. Here the application is more general and standardized scores are used to assess the model fit across the complete range of centiles. The paper also attempts to give a broad overview of the published literature on this topic, as well as pointing out aspects which would benefit from further research. The methods described here are suitable for cross-sectional studies where there is only one observation on each individual. Techniques for longitudinal data require a different approach and are not considered. The variable $Y$ of interest is assumed to be continuous, though methods exist for binary or ordered categorical variables (Wade *et al.*, 1995).

Section 2 defines terms and reviews ideas and methods used to estimate homogeneous (i.e. no age effect) reference intervals, extended to the age-specific case in Section 3. Some details of the most widely applied methods are illustrated by example. Calculating confidence bands and assessing goodness of fit are discussed in Sections 4 and 5 respectively. Section 6 compares three methods as applied to each of three data sets. Conclusions are presented in Section 7 and Appendix A describes available software.

## 2. ESTIMATING REFERENCE INTERVALS

To set the scene for the more complex case, methods for estimating reference intervals for samples with no age effect are reviewed first.

### 2.1. *Definition*

An $\alpha\%$ reference interval is defined as a range of values for a variable of interest, symmetric with respect to the median on a probability scale, which encompasses $\alpha\%$ of the data. An $\alpha\%$ age-related reference interval is the range bounded by two centile curves which encompasses $\alpha\%$ of the data at each age. For example, observations between the fifth and 95th centile curves lie within a 90% age-related reference interval.

This definition does not distinguish a 'descriptive' reference interval derived directly from sample data (empirical quantiles) from a putative 'true' reference interval (population quantiles, with a sample-based estimator). In practice, both paradigms are used. For a discussion of the rationale of reference intervals, see Albert and Harris (1987). The issue of the coverage of a reference interval based on the normal distribution is investigated by Royston and Matthews (1991).

### 2.2. *Notation*

Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed continuous random variables with cumulative distribution function $F$ and order statistics $Y_{(1)} < Y_{(2)} < \ldots < Y_{(n)}$. Let the location and scale parameters of $\{Y_i\}$ be $\mu$ and $\sigma$

respectively. Shape parameter(s) will be denoted $\lambda$, $\tau$ or $\eta$, according to the context. Let $g(\cdot)$ denote a monotonic data transformation. Let $p$ $(0 \leqslant p \leqslant 1)$ be a position on the distribution function of $Y$ and $C$ the corresponding quantile, such that $F(C) = p$. We shall usually refer to $C$ as 'the 100$p$th centile of $Y$'.

Define the $Z$-score (or SD-score) for $Y_i$ as $Z_i = \Phi^{-1}\{F(Y_i)\}$ where $\Phi^{-1}$ is the inverse normal distribution function. When $F \equiv \Phi$, $Y_i$ is normally distributed and the $i$th $Z$-score is simply $(Y_i - \mu)/\sigma$. The $Z$-score is a standardized measure of the distance of an observation from the centre of the distribution. Properties of $Z$-scores may be used to assess the goodness of fit of the distributional model to the data.

### 2.3. *Empirical Quantiles*

Finding an $\alpha$% reference interval involves estimating the $(100 - \alpha)/2$ and $(100 + \alpha)/2$ centiles of $Y$, which, since $C = F^{-1}(p)$, requires inversion of the distribution function. The simplest estimator of $F$ is the empirical distribution function, defined for an ordered sample of observations $Y_{(1)} < \ldots < Y_{(n)}$ by $\hat{F}(Y_{(i)}) = i/n$, $i = 1, \ldots, n$. The 100$p$th centile may then be estimated by $\hat{C} = \hat{F}^{-1}(p) = Y_{([np])}$ where $[\cdot]$ denotes the nearest integer. The estimator is biased for small sample sizes. When $np$ is not an integer, interpolation between the nearest two order statistics may be used. Decomposing $(n + 1)p$ into its integer part $r$ and its fractional part $s$, we have

$$\hat{C} = Y_{[r]} + s(Y_{[r+1]} - Y_{[r]}),$$

a weighted average of $Y_{[r]}$ and $Y_{[r+1]}$.

### 2.4. *Modelling the Distribution*

An alternative approach to centile estimation is to approximate $F$ by a member of a family of suitable distributions. The normal distribution (or some simple transformation towards normality) is almost exclusively used for this in the literature, presumably owing to its familiarity, mathematical convenience and relative ease of fitting.

Distributions which are related to the normal distribution may be classified according to the number of parameters to be estimated. In a two-parameter model, only the mean and variance are estimated, whereas three-parameter models have a shape parameter (related to skewness) and four-parameter models have two shape parameters.

#### 2.4.1. *Two-parameter distributions*

The assumption that $Y$ is normally distributed, common for example in the medical literature, is likely to be unrealistic, not least because most relevant variables (such as physical measurements) take only non-negative values, whereas the normal distribution is doubly unbounded. Positive skewness is common, and logarithmic transformation is usually effective in considerably reducing or even removing it. When $\ln Y$ is normal, $Y$ has a two-parameter log-normal distribution.

#### 2.4.2. *Three-parameter distributions*

The Box–Cox or power transformation (Box and Cox, 1964) $g(Y) = (Y^{\lambda} - 1)/\lambda$

for $\lambda \neq 0$, $g(Y) = \ln Y$ for $\lambda = 0$, is almost certainly the most popular transformation towards normality. The parameter $\lambda$ has the property that the distribution of $Y$ is negatively skewed for $\lambda > 1$, normal for $\lambda = 1$ and positively skewed for $\lambda < 1$. Estimates of $\lambda$ may be obtained by maximum likelihood (Atkinson, 1985) or by other methods, such as the use of order statistics (Hinkley, 1975), or by restricting $\lambda$ to a small set of values, typically 1 (identity), 0.5 (square root), 0 (logarithmic) or $-1$ (reciprocal), and choosing the best fit by some suitable criterion. Although many other normalizing transformations are possible, only the shifted log-function $\ln(Y - \tau)$ (Royston, 1992; Cheng and Iles, 1990) and the exponential transformation $\exp(\eta Y)$ (Manly, 1976) appear to be used in practice. In the former case $Y$ has a three-parameter log-normal distribution, and in the latter an exponential–normal distribution. Since $Y - \tau > 0$ in the three-parameter log-normal distribution, the parameter $\tau$ may be interpreted as a threshold or origin for $Y$. Estimation of $\tau$ by maximum likelihood is a 'non-regular problem' (Cheng and Iles, 1987) and may present computational difficulties due to an unbounded likelihood function as $Y$ approaches $\tau$. A limiting normal distribution for $Y$ is attained only as $\tau \to -\infty$, a disadvantage compared with the Box–Cox approach for which $\lambda = 1$ (the normal distribution) is an interior point of the parameter space. The parameter $\eta$ of the three-parameter exponential–normal distribution is monotonically inversely related to the skewness (standardized third moment) of $Y$ (Manly, 1976).

2.4.3. *Four-parameter distributions*

Four-parameter distributional models are generally more difficult to fit than models of lower dimension. However, the additional complexity may be justified with data sets which present difficulty in accurate estimation of the tails of the distribution. Each of the three-parameter distributions described above is embedded in what may be termed the shifted power or four-parameter Box–Cox distribution, for which the normalizing transformation is

$$g(Y) = \{(Y - \tau)^{\lambda} - 1\}/\lambda.$$

Among possible alternatives is the Johnson system (Johnson, 1949), which contains two subfamilies of four-parameter distributions known as $S_U$ and $S_B$. The corresponding normalizing transformations are

$$g(Y) = \sinh^{-1}\{(Y - \tau)/\lambda\}$$

and

$$g(Y) = \ln\{(Y - \tau)/(\eta + \tau - Y)\}$$

respectively. The $S_B$-distribution contains the three-parameter log-normal distribution as a limiting case. Slifker and Shapiro (1980) described methods which aid selection of the appropriate density. Parameter estimation for the Johnson system was originally by the method of moments, but maximum likelihood estimation is straightforward and preferable.

2.4.4. *Centile calculation*

Once an approximately normalizing transformation $g(Y) \sim N(\mu, \sigma^2)$ has been

selected, centiles in the transformed scale are estimated from the fitted parameters as

$$\hat{C}_p = \hat{\mu} + \Phi^{-1}(p)\hat{\sigma}.$$

It is then straightforward to transform back to the original scale, using $g^{-1}(\hat{C}_p)$, to obtain a reference interval in the appropriate units. Centile estimators with variants of the $\Phi^{-1}(p)\hat{\sigma}$-term were discussed by Royston and Matthews (1991) but they are numerically almost indistinguishable for practical sample sizes ($n > 100$).

#### 2.4.5. *Other approaches*

The International Federation of Clinical Chemistry (IFCC) Panel on Theory of Reference Values (1987) recommended a two-stage transformation approach, firstly applying Manly's exponential function, followed by a modulus transformation (originally proposed by John and Draper (1980)), which is related to the Box–Cox transformation. The aim here is to remove skewness and kurtosis in turn.

Lawrence and Trewin (1991) discussed the case where several populations are thought to be represented in a sample. The data were modelled by a mixture of normal distributions and parameters were estimated by maximum likelihood. Normal ranges for several biochemical measurements were obtained thus.

Merkouriou and Dix (1988) assumed a linear relationship between ordered observations and centiles. Reference intervals were defined according to the value of the Pearson correlation coefficient between the ordered values and corresponding sample fraction. A related method was proposed by Millward and Dix (1992). Neither method has a convincing statistical rationale and therefore neither can be recommended.

Tsay *et al.* (1979) proposed a graphical method for estimating the mean and standard deviation of log-transformed data on the assumption of a two-parameter log-normal distribution. Reference limits were obtained from a normal $Q$–$Q$-plot by interpolation or, in the case of small samples, by extrapolation. Such limits will be approximately correct if the two-parameter log-normal assumption is valid, but the subjective element of the method renders it unacceptable in comparison with standard estimation procedures.

### 2.5. *Nonparametric Estimators*

Section 2.3 discussed estimating the $p$th quantile as the $Y$-value where the empirical distribution function is closest to $p$. The main drawback to sample quantiles is their inefficiency, caused by variability of individual order statistics. Averaging over the order statistics and weighting according to their proximity to the sample quantile improves the estimate. A common technique for smoothing is kernel estimation (Silverman, 1986) and standard methods of nonparametric regression can be applied to $p$ *versus* $Y$ and estimated at a given point (Parzen, 1979). A major drawback of nonparametric estimators is that a simple closed formula with which to estimate the centile value of an arbitrary individual is unavailable. Centiles may only be displayed graphically or in tabular form.

Other nonparametric estimators (such as those by Stigler (1974), Reis (1980) and Sheather and Marron (1990)) are not discussed here since they appear not to have been applied to the calculation of reference intervals.

## 3. MODELLING AGE EFFECT

Methods for incorporating variation with age into reference intervals are now discussed. Methods of estimation should have a precise goal, e.g. to obtain graphs, tables or formulae representing the required centiles. Ideally, both centile curves and a model describing the formulation should be obtained. (This is an attractive feature of most parametric and some semiparametric approaches.) If this is not possible, then the limitations of the results should be made clear and other information can be documented, such as tables of centile values.

The possible occurrence of negative centile estimates for variables which are positively bound, e.g. measurements of body size, should be avoided. A careful choice of transformation in parametric methods, e.g. using distributions with positive support, would prevent such an eventuality, but resolution is more difficult when using semiparametric and nonparametric methods. Constraining the distance between successive centile estimates (known as 'commonality') can be applied to prevent the curves from touching or crossing. The issue is again more easily addressed in a parametric framework.

All the notation defined in Section 2.2 applies here, except that the data may now be written as pairs of variables $\{(Y_i, T_i)\}_{i=1}^{n}$, where $T$ represents age. For illustration, some of the techniques reviewed will be used to analyse the data in Fig. 1(a).

### 3.1. *Curves Based on Empirical Estimates*

If a sufficiently large data set is available, the most obvious way to create centile curves may be to calculate the empirical estimates for each centile value at each age point. Such curves will be rough, even for large sample sizes. Instead of considering each age value separately, 'windows' of values centred about points on the age axis may be considered, although curves with too much local variation may be produced. Windows are usually chosen to contain 5–10% of the data (Healy *et al.*, 1988). Such curves have been derived for the 10th, 50th and 90th centiles for the cholesterol data (Fig. 1(b)) with a window size containing 5% of the data points and estimation at every fifth observation ordered by age. Increasing the window size, or bandwidth, would make the curves less rough but information may be lost by oversmoothing.

Himes and Hoaglin (1989) proposed a 'resistant smoothing' procedure. The empirical centile curves for a given set of centile values are used as initial estimates. The intercentile differences are calculated between successive centile curves, e.g. 25th–10th empirical centiles. This process is known as delineation. The median and the intercentile differences are smoothed by finding their respective medians within windows. These smoothed curves are then recombined to estimate the centile curves. This may be repeated with different sizes of window to produce smoother centile curves.

### 3.2. *Logarithmic Transformation*

Royston (1991) proposed transforming $Y$ using a shifted logarithmic function, $\ln(Y-\tau)$, and finding the best polynomial fit to the transformed values plotted against $T$. Models are fitted using least squares regression and Royston suggested a backwards stepwise approach to obtain an appropriate degree of polynomial. It is recommended that a cubic polynomial be used as an initial fit, then reduce the degree

if the coefficient for the highest order term is not statistically significant. Clearly, for complex curve shapes, it may be more sensible to start with a higher order fit. The parameter $\tau$ is chosen to minimize the non-normality of the residuals according to a test statistic such as the Shapiro–Francia $W'$ (Shapiro and Francia, 1972) or the Shapiro–Wilk $W$ (Shapiro and Wilk, 1965). (See Royston (1993) for references to algorithms to calculate these statistics.) Royston's (1991) proposal for modelling the age-specific standard deviation, which involved dividing the residuals into three groups according to age, is crude and awkward to implement. A better method is to estimate the standard deviation as the fit from the regression of the absolute residuals (multiplied by $\sqrt{(\pi/2)}$) on $T$ as proposed by Altman (1993). Centiles may then be estimated under the assumption of normality and transformed back to the original scale.

Analysis of the cholesterol data of Fig. 1 reveals an approximately normalizing transformation to be $\ln(Y+3)$ (i.e. $\tau$ estimated as $-3$). The cubic term of a fitted polynomial was found to be non-significant ($p$-value 0.095) and a second-degree polynomial was required to model the mean of the transformed data (see Fig. 2(a)).

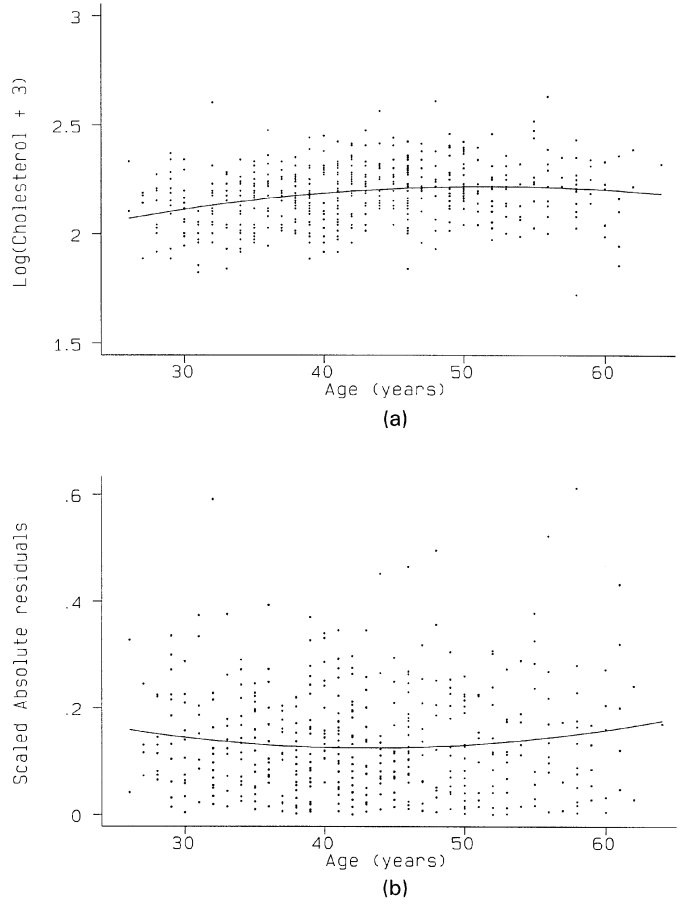


Fig. 2. (a) Fitted median curve for log(cholesterol + 3); (b) plot of scaled absolute residuals *versus* age with fitted standard deviation curve

Slight curvature in the absolute residuals (see Fig. 2(b)) suggests a quadratic also for the standard deviation, although arguably, since these terms had coefficients with $p$-values of 0.037 and 0.033, a constant may be adequate. (The simple linear model had a non-significant slope coefficient.)

Other papers discussing the log-transformation tend to be data specific as opposed to general techniques. Bland *et al.* (1990) assumed that birth weight measurements were approximately normal but used the log-transformation to stabilize the variance across age. Chinn (1992) made the same assumptions when analysing height for age, although she used a shifted log-transformation. In fetal medicine, it appears to be common (see for example Simon *et al.* (1990) and Todros *et al.* (1987)) to log-transform the response variable (e.g. birth weight) and to fit a model of the form $\beta_0 + \beta_1 T + \beta_2 T^3$ where $T$ is menstrual or gestational age.

### 3.3. *LMS Method*

In the original form of the *LMS* method (Cole, 1988), the parameters of a Box–Cox distribution for $Y$ are first estimated within each of several contiguous age groups. Cole (1988) proposed a useful parameterization in which the three parameters $L$, $M$ and $S$ (standing for $\lambda$, $\mu$ and $\sigma$) are essentially the skewness, median and coefficient of variation of $Y$ respectively. (In fact, the skewness depends on both $\lambda$ and $\sigma$.) The parameters are estimated separately within each age group by maximum likelihood and then smoothed across age. Any regression smoothing procedure may be applied here, e.g. polynomial regression or kernel estimation. Centiles are calculated for $Y^{\lambda}$ and back transformed to the original scale. Cole (1988) also mentioned the four-parameter Box–Cox distribution as a possible model but regarded the idea as impractical.

Categorizing the age variable is a subjective procedure and different groupings produce different centile curves. Cole (1990) recommended group sizes of at least 100 when constructing reference standards for human growth. In data sets where the regression effect is less strong, smaller group sizes may suffice. The cholesterol data were divided into 10 age groups of about 50 points each and maximum likelihood estimates found for each of $L$, $M$ and $S$. The estimates for $M$, shown as a curve, are superimposed on a scatterplot of the data in Fig. 3(a). These clearly need to be smoothed before fitting centile curves but subjectively seem to be a satisfactory initial fit. The values to be used for constructing the $L$- and $S$-curves are shown in Figs 3(b) and 3(c). Confidence intervals for $L$ and $S$ are very wide, suggesting a large amount of 'noise' and very little structure in the parameters across age. This would be addressed more formally within the smoothing stage.

Cole and Green (1992) added a nonparametric aspect to the original *LMS* method by using maximum penalized likelihood to estimate the age-related curves for each of the parameters by natural cubic splines. The advantages of the approach are that the subjective grouping step is removed and the curve fitting across age is controlled directly by the values of three smoothing parameters ('equivalent degrees of freedom' (EDF)). The disadvantages are that, since statistical inference for use with penalized likelihood is not well understood, the choice of EDFs is somewhat subjective, and the 'best fitting' curves may not be uniquely determinable. Cole and Green (1992) suggested, as a very rough guideline, comparing the difference in deviance ($-2\log$(penalized likelihood)) between two models where the total number of

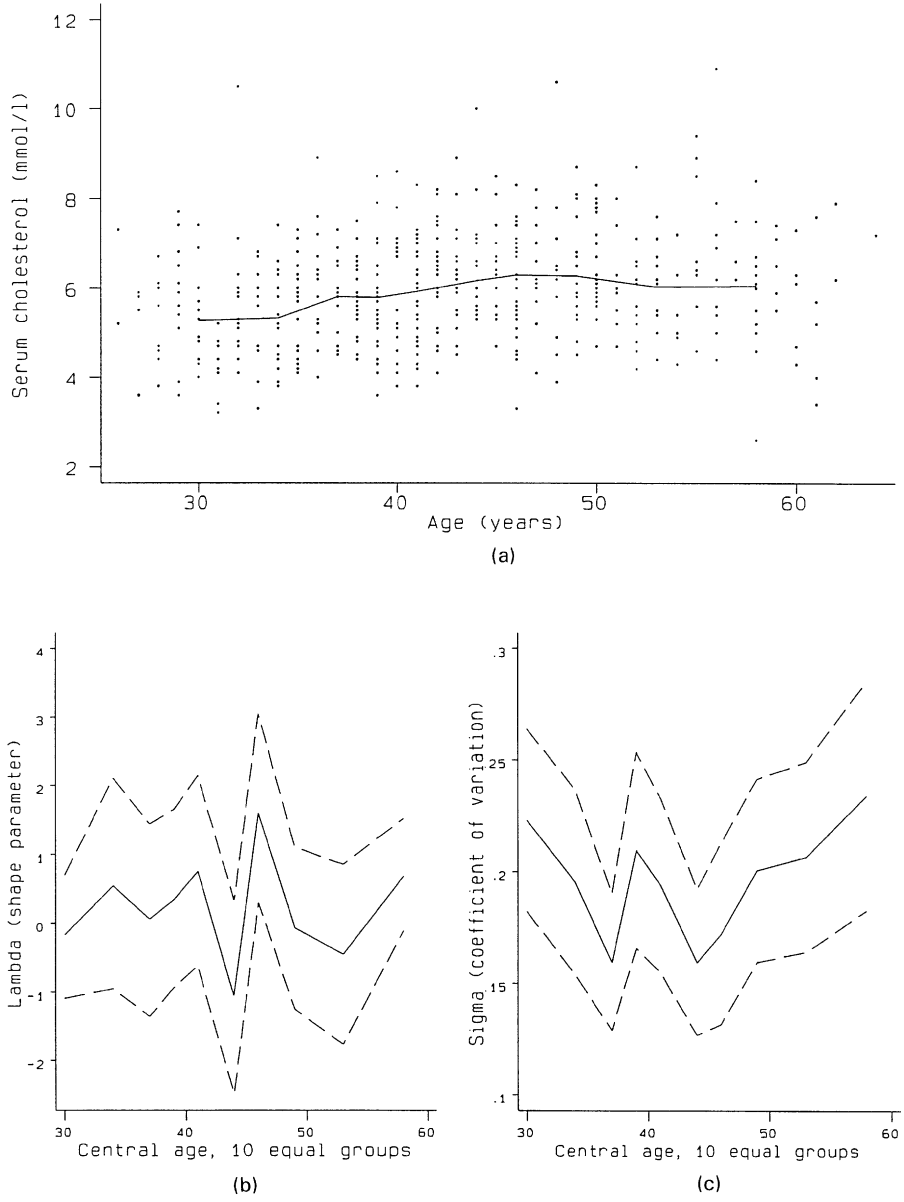


Fig. 3. (a) *M*-curve based on 10 age groups; (b) *L*-curve based on 10 age groups with $\pm 2$ standard error bands (— — —); (c) *S*-curve based on 10 age groups with $\pm 2$ standard error bands (— — —)

EDFs differed by e to a $\chi^2_e$-distribution. This 'test' has no formal statistical justification, although it has an analogy with likelihood ratio tests. The fitted *L*-, *M*- and *S*-values need to be tabulated using a suitable age grouping and interpolated for non-tabulated ages. Despite the drawbacks, the convenience and flexibility of the new method are vastly superior to those of the original.

Wade and Ades (1994) chose a somewhat complex modelling approach when applying the Box–Cox transformation of CD4 lymphocyte counts measured in the blood of children. The *L*-, *M*- and *S*-curves were approximated by non-linear models which involved negative exponential powers and likelihood ratios were used to compare the fits of different possible models. Although the estimated centile curves appear to fit their particular data well, a range of apparently appropriate models was given, with few guidelines on how to choose between them.

### 3.4. *Use of Johnson Distribution*

Thompson and Theron (1990) calculated centiles based on maximum likelihood estimates of the four parameters of the $S_U$-distribution from the Johnson family of densities mentioned in Section 2.4.3 and used low order polynomials to smooth the parameters with respect to age.

### 3.5. *HRY Method*

Cleveland (1979) proposed fitting a line to a scatterplot by using locally weighted regression. Healy *et al.* (1988) extended the idea to estimating centile curves in a two-stage smoothing process with what we shall call the HRY method. A set of $m$ centile positions is chosen (typically 3, 10, 25, 50, 75, 90 and 97 with $m = 7$), with corresponding normal equivalent deviates $z_1, \ldots, z_m$. Initial centile estimates are found by moving a window along the age axis and calculating the $m$ empirical centiles from the data in each window (as described in Section 3.1). Each of the resulting $m$ empirical centile curves is smoothed using a polynomial of degree $d$. For $j = 0, 1, \ldots, d$, the $m$ values $\{a_{j,k}\}_{k=1,\ldots,m}$ of the $j$th polynomial coefficient are themselves smoothed by fitting polynomials of degree $q_j$ in the $\{z_k\}$. This restricts the distance between centiles and prevents the resulting curves from crossing. The model for the $k$th empirical centile curve $c_k(T)$ may be written

$$c_k(T) = a_{0,k} + a_{1,k}T + \ldots + a_{j,k}T^j + \ldots + a_{d,k}T^d + \epsilon_k(T)$$

where

$$a_{j,k} = b_{j,0} + b_{j,1}z_k + \ldots + b_{j,q_j}z_k^{q_j}$$

and $\epsilon_k(T)$ represents residual error.

A model of the form $d = 2$, $q_0 = 2$, $q_1 = 1$, $q_2 = 0$ was fitted to the cholesterol data. To clarify the fitting of polynomials to $Z$-scores, the initial empirical centiles and their fitted values from the quadratic model have been plotted against $z_k$ for three different age values in Fig. 4. (The open circles, crosses and full circles represent points at ages 35, 45 and 55 years respectively.) The model appears to fit fairly well, though a closer approximation to the extreme centiles would require higher order polynomials to be fitted to the $\{a_{j,k}\}$.

To allow for the more complex curve shapes that certain variables observed over a wide age range may display, Pan *et al.* (1990) suggested dividing the data into contiguous age groups, fitting polynomials within each group and smoothing points where these meet using an extra polynomial term. As an alternative, Goldstein and Pan (1992) proposed that these age groups be defined initially (as opposed to after the estimation of the raw centiles) and that the fitted response be made smooth across the join points by equating derivatives of the curves on either side.

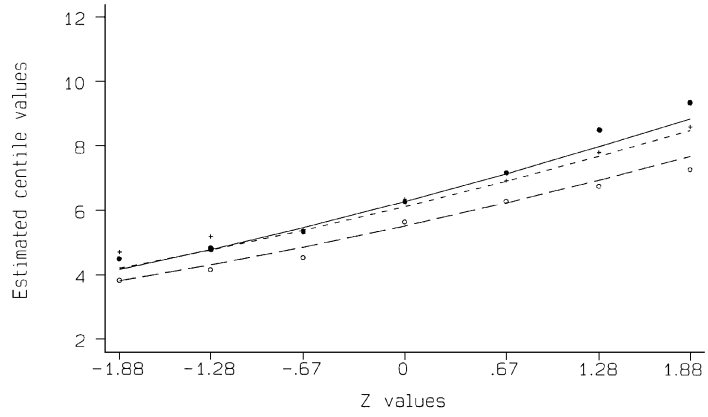


Fig. 4. Estimated empirical centile values (○, +, ●) and fitted values (— — —, - - - - -, ———) at ages 35, 45 and 55 years respectively from the HRY method

### 3.6. *Other Methods*

A technique employed by Hamill *et al.* (1977) to provide height and weight growth standard curves for children in the USA involved calculating the empirical centiles for small intervals of data across age and smoothing over these values by using splines. Jones, in the discussion of Cole (1988), described how to weight a natural cubic spline to estimate quantile curves directly.

Rossiter (1991) used the nonparametric smoothing approach to density estimation, using multidimensional kernels to condition on one or more covariates. The centiles were calculated by solving the integral of the estimates obtained for the density for $p$ ($0 \leqslant p \leqslant 1$) by using a Newton–Raphson procedure.

Jones and Hall (1990) fitted each of the centile curves of interest with kernel estimates of the same degree by using a weighted empirical distribution function (Stone, 1977).

Efron (1991) estimated 'regression percentiles' by minimizing the asymmetric squared loss, a variant of least squares estimation which weights residuals according to their positive or negative sign. The method proposed by Koenker and Bassett (1978) is similar but is based instead on the asymmetric absolute loss function.

## 4. CONFIDENCE BANDS

One way to assess the precision of an estimated age-specific reference interval is to consider a pointwise confidence band for the individual centile curves. When precision at a certain age is low, confidence bands of centiles may overlap, implying indeterminacy between the curves. For centile estimates which are approximately normally distributed, confidence bands may be calculated from $\pm 2$ standard errors. Alternatively, for models whose parameters have been estimated by maximum likelihood, age-specific confidence bands may be constructed by multivariate optimization of the likelihood function at selected ages, followed by graphical interpolation (Wade and Ades, 1994). For semiparametric or nonparametric methods, other approaches must be considered. Bootstrapping (Efron, 1979) is an asymptotically consistent resampling technique which is particularly useful when the theoretical aspects of a

problem become intractable or impractical to resolve, as here. The most basic form is to resample the $n$ pairs of $(Y_i, T_i)$ with replacement and to calculate the centile curves for each of a large number of such bootstrap samples. Another approach is to consider centiles from data simulated under the estimated curves. The 'envelopes' formed by the curves from either of these approaches give approximate confidence intervals for the original centile curve.

## 5. GOODNESS OF FIT

Assessing the fit of centile curves to data appears to have received little attention in the literature. A subjective appraisal of the fitted curves by eye may rule out models which are clearly unsatisfactory, for example, according to criteria derived from subject-matter knowledge. For the semiparametric *LMS* method (Cole and Green, 1992), comparisons of penalized likelihoods for models with different EDFs may give an idea of the complexity of the required model. Goodness-of-fit methods which have been proposed, and some procedures which are simple extensions of routine tests used in other contexts, are discussed below.

If the age effect is ignored, a Pearson $\chi^2$-statistic with $k$ degrees of freedom may be used to assess whether the frequencies of observations found within the $k + 1$ groups defined by $k$ centile curves differ significantly from the expected number. An extension is to categorize the age variable into $l$ contiguous groups, say of roughly equal size, creating a contingency table with $k + 1$ rows and $l$ columns (Healy *et al.*, 1988). The nominal degrees of freedom for a $\chi^2$-test are now $kl$, since the observed and expected numbers of observations are constrained to be equal in each of the $l$ columns. However, the estimation of parameters in the model from which the centiles were derived will further constrain the $\chi^2$-statistic and reduce the effective degrees of freedom, resulting in a conservative test if no adjustment is made. If the unadjusted test is nevertheless significant a poor fit is indicated, whereas a non-significant result may require further model appraisal.

Goodness of fit may be assessed graphically by using the initial step of the HRY (Healy *et al.*, 1988) method, which involves estimating the empirical centiles by local averaging. The same technique may be applied to a plot of $Z$-scores *versus* age to compare the observed and expected centiles (see for example Cole and Green (1992)). However, such diagrams may be difficult to interpret owing to the variability of the observed lines and the serial correlation between neighbouring values. A plot of the $Z$-scores resulting from the second-order polynomial fit described in Section 3.2 is shown in Fig. 5(a). The empirical centile plot for these data (Fig. 5(b)) is very rough but the lines are centred on the expected $Z$-scores ($\pm 1.28$).

A normal $Q$–$Q$-plot of the residuals is commonly used to assess the normality assumption of the data. This technique is also appropriate for some of the methods used to fit centile curves, as is the Shapiro–Wilk $W$-test (Shapiro and Wilk, 1965; Cole, 1988). The test may be conservative since it is not adjusted to account for estimated parameters other than a single mean and standard deviation.

## 6. EXAMPLES

Table 1 contains details of the models fitted to each of three data sets by three different methods. Each model was fitted according to the advice given in their

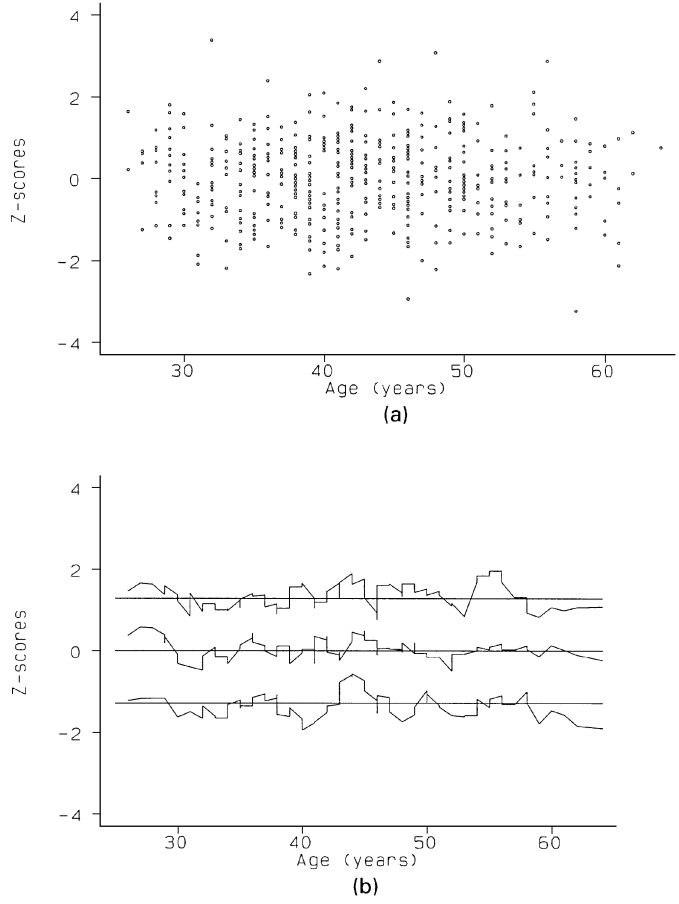


Fig. 5. (a) $Z$-scores for the shifted log-transformed model; (b) empirical quantile plots for $Z$-scores at the 10th, 50th and 90th centiles with horizontal lines denoting $\Phi^{-1}(p)$ where $p = 0.1, 0.5, 0.9$

TABLE 1
*Summary of models fitted to example data sets*

| *Data set* | *Method* | *Model details* |
|---|---|---|
| Cholesterol | LOG | $\log(Y + 3)$; quadratic for mean, quadratic for standard deviation |
| | *LMS* | EDFs: $L = 1$, $M = 4$, $S = 1$ |
| | HRY | $d = 2$, $q_0 = 2$, $q_1 = 1$, $q_2 = 0$ |
| Abdominal circumference | LOG | $\log Y$; quartic for mean, quadratic for standard deviation |
| | *LMS* | EDFs: $L = 1$, $M = 4$, $S = 3$ |
| | HRY | $d = 2$, $q_0 = 2$, $q_1 = 1$, $q_2 = 1$ |
| Triceps skinfold thickness | LOG | $\log Y$; 7th-order polynomial for mean, quadratic for standard deviation |
| | *LMS* | EDFs: $L = 6$, $M = 9$, $S = 6$ |
| | HRY | Cut point at age 8 years: $d_1 = 3$, $q_{10} = 6$, $q_{11} = 6$, $q_{12} = 1$, $q_{13} = 0$ and $d_2 = 4$, $q_{20} = 2$, $q_{21} = 2$, $q_{22} = 1$, $q_{23} = 1$, $q_{24} = 1$ |

original published source. Diagrams representing analyses of each data set (see below) are given in Figs 6–8. Each diagram comprises a $3 \times 3$ array of plots. The rows correspond to three methods of analysis (the shifted log-transformation (LOG), the *LMS* method with penalized likelihood and the HRY method) and the columns to three types of display. Parts (a), (d) and (g) show the raw data with estimated 10th,

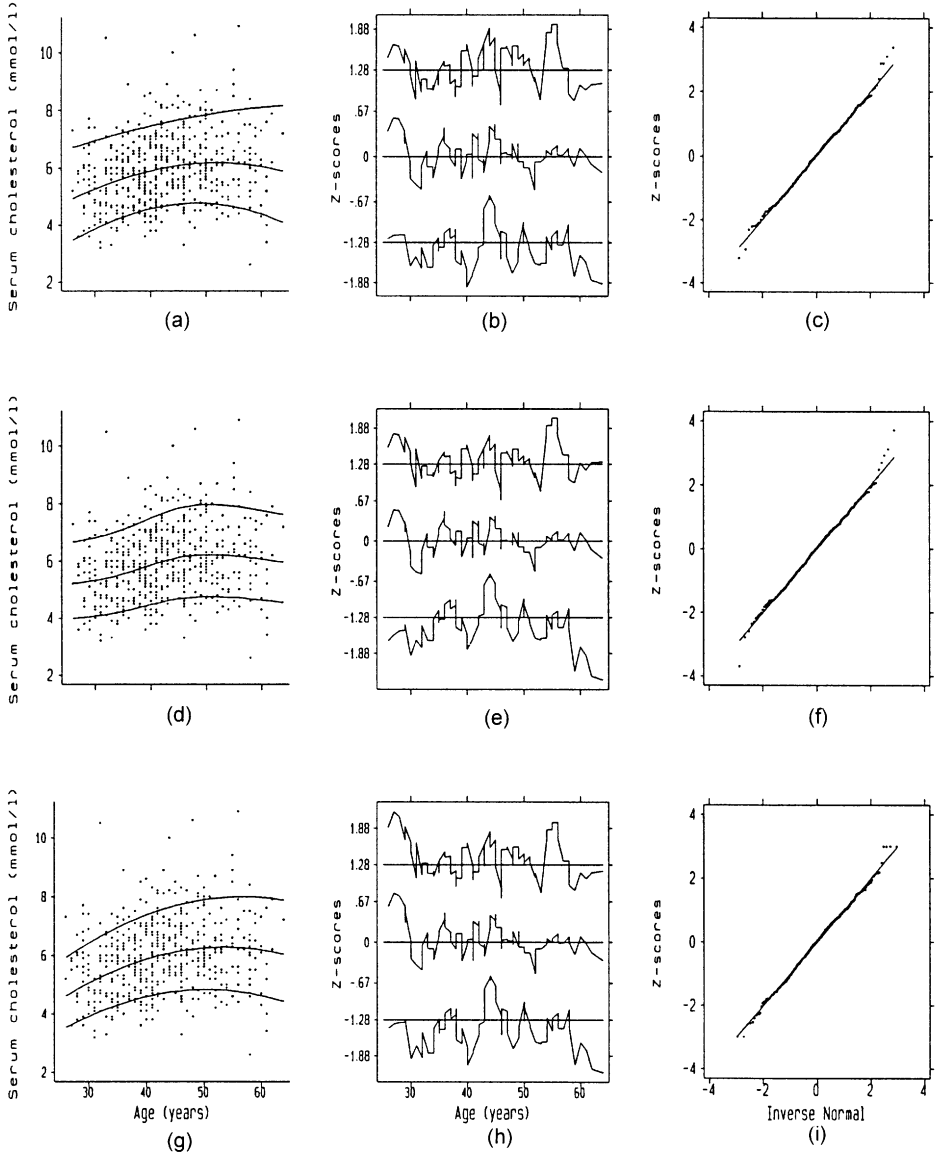


Fig. 6. Serum cholesterol data: (a) 10th, 50th and 90th centiles using the LOG method; (b) empirical centile plots of $Z$-scores from the LOG method at the 10th, 50th and 90th centiles; (c) normal $Q$–$Q$-plots of $Z$-scores from the LOG method; (d), (e), (f) equivalents of (a), (b) and (c) respectively for the *LMS* method; (g), (h), (i) equivalents of (a), (b) and (c) respectively for the HRY method

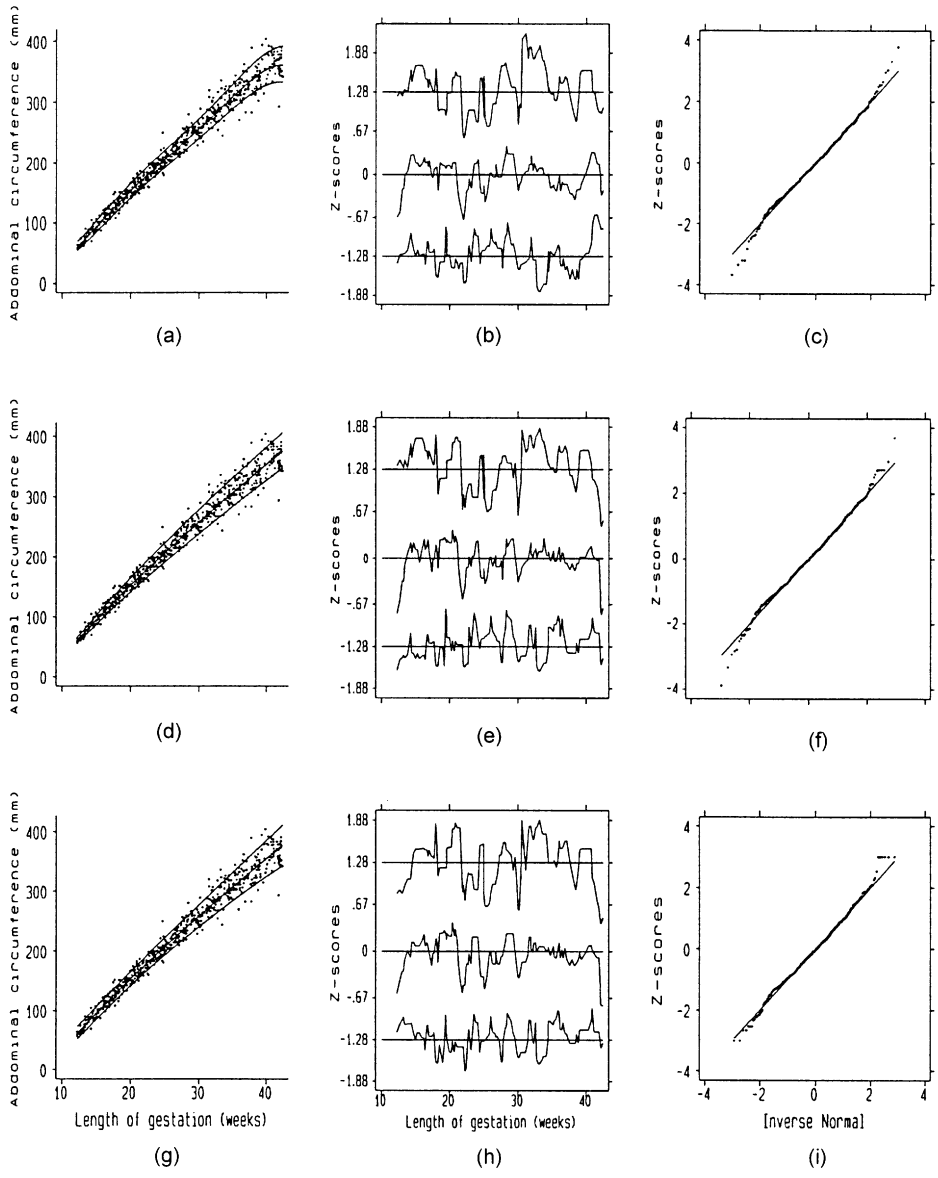


Fig. 7. Abdominal circumference data: format as in Fig. 6

50th and 90th centile curves inscribed. Parts (b), (e) and (h) show empirical quantiles of $Z$-scores. If the model is correct, the three quantile lines should be approximately equal to $-1.28$, 0 and 1.28 respectively, i.e. to the 10th, 50th and 90th centiles of $N(0, 1)$. The quantiles were computed at every fifth observation across age using 5% of neighbouring values at each point. Parts (c), (f) and (i) show normal $Q$–$Q$-plots of the $Z$-scores. For formal testing of the goodness of fit, a Pearson $\chi^2$-statistic (3 degrees of freedom) and $p$-value may be used to compare the numbers of observations lying in

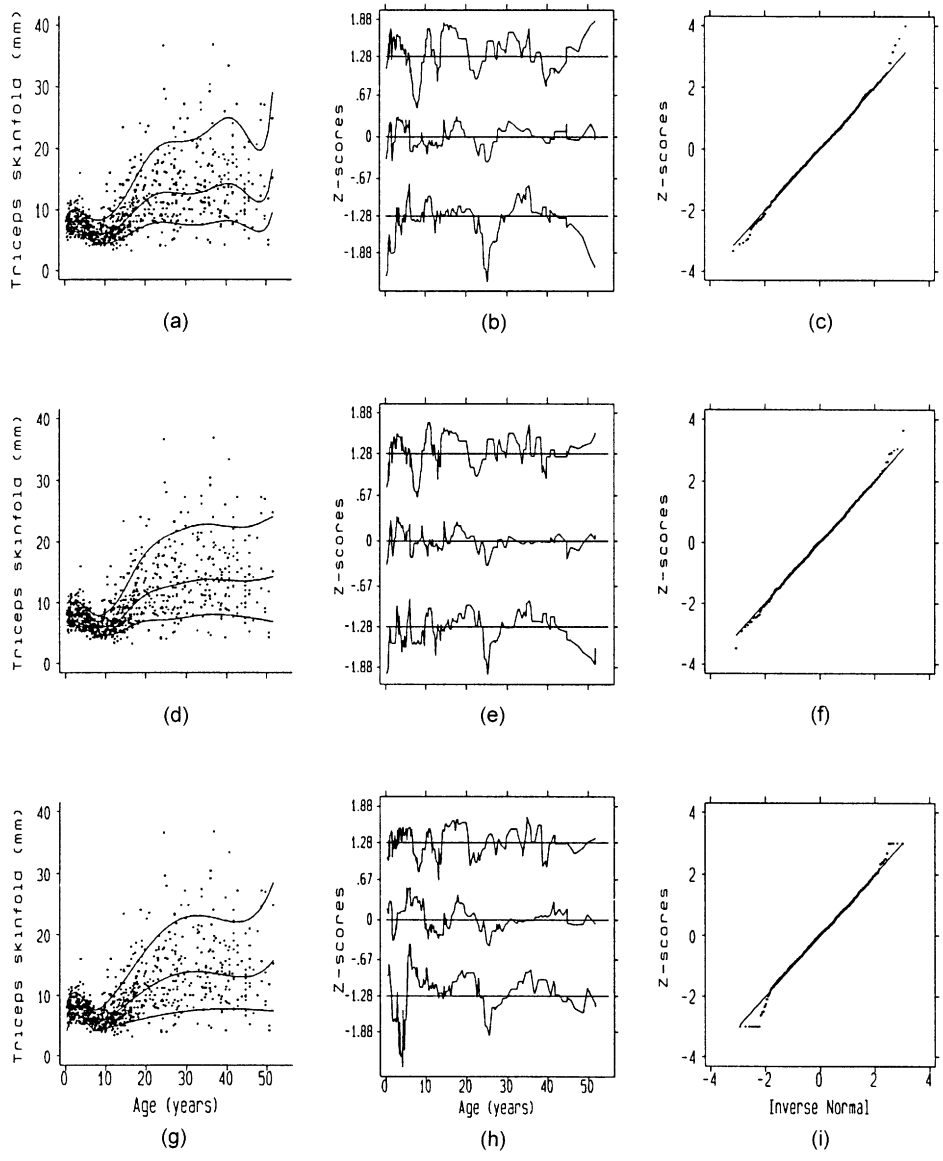


Fig. 8. Triceps skinfold thickness data: format as in Fig. 6

the four groups defined by the estimated 10th, 50th and 90th centiles with expectation, irrespective of age. To assess the age-specific fit, the data were divided into five contiguous age groups of equal size and a Pearson $\chi^2$-test statistic (15 degrees of freedom) was calculated to compare the observed and expected frequencies in the resulting contingency table. Shapiro–Wilk $W$-tests for overall normality of $Z$-scores were also computed. The results for each combination of method and data set are given in Table 2.

TABLE 2
*Results of goodness-of-fit tests*

| *Data set* | *Method* | $\chi^2_3$ | *p-value* | $\chi^2_{15}$ | *p-value* | *W* | *p-value* |
|---|---|---|---|---|---|---|---|
| Cholesterol | LOG | 1.16 | 0.76 | 7.95 | 0.93 | 0.9972 | 0.54 |
| | *LMS* | 0.57 | 0.90 | 5.50 | 0.99 | 0.9968 | 0.43 |
| | HRY | 3.22 | 0.36 | 8.55 | 0.90 | 0.9974 | 0.62 |
| Abdominal circumference | LOG | 2.59 | 0.46 | 12.81 | 0.62 | 0.9950 | 0.04 |
| | *LMS* | 3.62 | 0.31 | 13.81 | 0.54 | 0.9955 | 0.07 |
| | HRY | 4.63 | 0.20 | 10.53 | 0.79 | 0.9950 | 0.04 |
| Triceps skinfold thickness | LOG | 0.11 | 0.99 | 27.43 | 0.03 | 0.9969 | 0.08 |
| | *LMS* | 0.40 | 0.94 | 20.24 | 0.16 | 0.9987 | 0.75 |
| | HRY | 8.79 | 0.03 | 52.87 | $<0.001$ | 0.9940 | 0.001 |

### 6.1. *Cholesterol in Men*

The cholesterol data set has been described earlier. The centile curves for each of the methods appear similar around the centre of the data (i.e. between 30 and 55 years) (Figs 6(a), 6(d) and 6(g)). Differences are visible in the extremes, although all indicate some reduction in median cholesterol concentration after 55 years of age. The pattern is less clear for younger men. Little can be concluded from the empirical quantile plots (Figs 6(b), 6(e) and 6(h)) and normal plots (Figs 6(c), 6(f) and 6(i)) other than that the fits seem satisfactory.

The overall and age-specific $\chi^2$-tests indicate that the proportions of values falling between centile curves are as expected (Table 2). The Shapiro–Wilk tests provide no evidence of departure from normality in the $Z$-scores. Thus the various tests and plots are unable to discern much difference between the methods for this particular data set. Each method seems to give a good fit.

### 6.2. *Fetal Abdominal Circumference*

Measurements of abdominal circumference were taken on 663 fetuses during ultrasound scans at Kings College Hospital, London, at gestational ages ranging between 12 and 42 weeks. Reference intervals derived from these data are reported in Chitty *et al.* (1994b). The strong, but clear, relationship illustrated here is typical of fetal size measurements.

The three sets of fitted centile curves (Figs 7(a), 7(d) and 7(g)) look similar, although those from the HRY approach are slightly further apart at early gestational ages than for the other two methods. As far as may be judged, the empirical quantile plots (Figs 7(b), 7(e) and 7(h)) suggest adequate fits. The normal plots (Figs 7(c), 7(f) and 7(i)) indicate that the distribution of $Z$-scores has somewhat longer tails than the normal distribution. (The short sequence of $Z$-scores equal to 3 in Fig. 7(i) seems to be an artefact of the GROSTAT program, though no mention is made in the instruction manual.)

Table 2 shows that the estimated proportions of observations falling between centile curves differ little from expectation. However, the results of the $W$-tests suggest that the $Z$-scores may not be normally distributed.

Although the curves for the 10th, 50th and 90th centiles appear to fit well, the non-normality in the lower tails of the $Z$-scores indicates that estimates of extreme

centiles would be biased and therefore that more attention should be paid to the higher moments of the data.

### 6.3. *Triceps Skinfold Thickness*

Anthropometric measurements were collected on a sample of 892 Gambian females aged between 3 months and 52 years (see Cole and Green (1992)). Triceps skinfold thickness is correlated with the amount of body fat and is therefore a convenient (if crude) measure of nutritional status. The pronounced dip in skinfold thickness around the age of 8–10 years is an unusual feature (it may correspond to the prepubescent growth spurt) requiring careful modelling. The *LMS* model was the easiest with which to produce a good fit since it only required an increase in EDF, whereas the other methods necessitated considerable thought and trial and error.

The methods produce curves which are similar except at ages over 30 years (Figs 8(a), 8(d) and 8(g)). The HRY and LOG curves are clearly overfitted at the higher ages but a choice of lower order polynomials results in a poorer fit for low ages. The empirical quantile plots (Figs 8(b), 8(e) and 8(h)) are very difficult to interpret. Each of the plots has a dip in the 90th centile around the age of 8–10 years, suggesting a poor fit. The *LMS* method gives the most satisfactory fit overall. (As in the previous example, the short sequences of $Z$-scores at $\pm 3$ in Fig. 8(i) are artefacts of the GROSTAT program.)

The goodness-of-fit tests reinforce the subjective impressions from the empirical quantile plots. The LOG and HRY approaches do not perform well. The Shapiro–Wilk tests detect departures from normality in the $Z$-scores for the LOG and HRY methods. The tests for the *LMS* method appear to be satisfactory.

## 7. DISCUSSION

For cases where the sample size is small and normality and homoscedasticity assumptions are plausible, the use of conventional polynomial regression may be justified but may not produce reference intervals that are sufficiently reliable for, say, routine clinical use. Overall impressions of the general methods for estimating age-related centiles may be summarized as follows.

### 7.1. *LOG Method*

The shifted log-transformation together with polynomial regression is conceptually simple and easy to use. The rules for model selection (based on the statistical significance of the regression coefficients) are clear and implementation is straightforward. The resulting centile curves and $Z$-scores can be expressed as explicit formulae. However, the method suffers from the well-known limitations of polynomial curve shapes. Also, although the $\tau$-parameter and log-transformation adjust for skewness, time varying skewness cannot be easily accommodated. Further, non-normal kurtosis may remain in the data after transformation.

### 7.2. *LMS Method*

The *LMS* method with penalized likelihood is extremely flexible and widely applicable. It is usually easy to produce convincing centile curves even when the data

appear to have a complex shape. The complexity of the shape of the parameter curves is reflected in the number of EDFs. Time varying skewness is easily dealt with, though as with the LOG method some non-normal kurtosis may remain. As with most nonparametric fitting procedures, formal inference for model comparison is unavailable and succinct formulae for the centile curves or $Z$-scores are unobtainable. Implementation of the method requires specially written software.

### 7.3. *HRY Method*

The HRY method is flexible, with the suggestions of Pan *et al.* (1990) and Goldstein and Pan (1992) making it even more so. Time varying skewness and kurtosis can be assessed and accounted for in the choice of polynomial across the $Z$-scores. The choice of degrees of polynomial requires considerable expertise and trial and error, and it is not always clear how to improve the fit. Improper density estimation may occur since the relationship between the empirical cumulative distribution function and the $Z$-scores is not constrained to be monotonic. Although formulae are available for the centile curves, inversion to obtain the $Z$-scores requires an iterative search procedure unless a very simple model has been fitted. The method also requires specially written software.

### 7.4. *Goodness of Fit*

The assessment of goodness of fit is an area requiring further research. In Section 6, a combination of graphical and test-based procedures was of some help in deciding which methods performed best, but no single method enabled firm conclusions about the adequacy of the fitted curves to be drawn. It may be helpful to distinguish methods which examine the fit of specific estimated centile curves from those which look at the distribution of the residuals ($Z$-scores). In each case, overall tests and those based on grouping by age (or perhaps by treating age as a continuous covariate) should be considered. The extent to which the null distributions of relevant test statistics (such as Pearson $\chi^2$ for tables of observed and expected frequencies of the Shapiro–Wilk $W$ for residuals) are perturbed by estimation of model parameters should be assessed. Questions of the power of different procedures and the sensitivity to the number of age groups and/or centile curves chosen should be addressed.

### 7.5. *Conclusions*

The ideal for constructing age-specific reference intervals would be a relatively simple, flexible method which could be applied successfully to many sets of data — perhaps a combination of ideas from existing methods. If there is interest in the extremes of the reference population (i.e. the tails of the reference distribution), particular attention must be paid to an accurate estimation of the low and high centiles. In any case a simple formula which allows estimation of an individual's centile position is extremely valuable. This requirement excludes the $LMS$ method with penalized likelihood, the HRY approach and all methods based on direct quantile estimation. Parametric methods based around maximum likelihood estimation would satisfy this proposal and allow formal inference for comparing models by using likelihood ratio tests. The requirements from the data under analysis should always be kept in mind.

## ACKNOWLEDGEMENTS

This research received financial support from project grant 039911/Z/93/Z from the Wellcome Trust. The authors wish to thank Dr Tim Cole for his helpful comments on an earlier draft of the manuscript and for allowing the use of the triceps skinfold thickness data set.

## APPENDIX A: IMPLEMENTATION

The following is a brief account of available packages and programs which implement the main methods described earlier.

The simplicity of the shifted log-transformation approach (Royston, 1991; Altman, 1993) enables it to be used with most 'standard' statistical packages.

The package GROSTAT, developed by M. J. R. Healy and co-workers for the World Health Organization, is available from Dr H. Q. Pan, Institute of Education, 20 Bedford Way, London, WC1H 0AL, UK. Its manual describes the statistical basis of the methods (Healy *et al*., 1988; Pan *et al*., 1990; Goldstein and Pan, 1992) in detail. GROSTAT is an extension of Healy's general purpose statistical package NANOSTAT. The user interface is crude and only medium resolution graphics are available.

The penalized likelihood version of the *LMS* method (Cole and Green, 1992) has been implemented as a stand-alone Fortran program by Dr T. J. Cole. The user needs access to another package for display and further analysis of the results, e.g. to produce graphs of centile curves.

Wade and Ades (1994) also implemented their method in Fortran, including routines from the Numerical Algorithms Group *Fortran Subroutine Library*.

Quantile regression (Koenker and Bassett, 1978) is available in the statistical package STATA (StataCorp, 1995) as the `qreg` command.

## REFERENCES

Albert, A. and Harris, E. K. (1987) *Multivariate Interpretation of Clinical Laboratory Data*. New York: Dekker.

Altman, D. G. (1993) Construction of age-related reference centiles with absolute residuals. *Statist. Med*., **12**, 917–924.

Atkinson, A. C. (1985) *Plots, Transformations, and Regression*. Oxford: Oxford Science.

Bland, J. M., Peacock, J. L., Anderson, H. R., Brooke, O. G. and De Curtis, M. (1990) The adjustment of birthweight for very early gestational ages: two related problems in statistical analysis. *Appl. Statist*., **39**, 229–239.

Bonellie, S. R. and Raab, G. M. (1996) Fitting centile curves to birthweight data. *Statist. Med*., to be published.

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). *J. R. Statist. Soc*. B, **26**, 211–252.

Campbell, S. and Newman, G. B. (1971) Growth of the fetal biparietal diameter during normal pregnancy. *J. Obstetr. Gyn. Br. Commwlth*., **78**, 513–519.

Cheng, R. C. H. and Iles, T. C. (1987) Corrected maximum likelihood in non-regular problems. *J. R. Statist. Soc*. B, **49**, 95–101.

——— (1990) Embedded models in three-parameter distributions and their estimation. *J. R. Statist. Soc*. B, **52**, 135–149.

Chinn, S. (1992) A new method for calculation of height centiles for preadolescent children. *Ann. Hum. Biol*., **19**, 221–232.

Chitty, L. S., Altman, D. G., Henderson, A. and Campbell, S. (1994a) Charts of fetal size: 2, head measurements. *Br. J. Obstetr*., **101**, 35–43.

——— (1994b) Charts of fetal size: 3, abdominal measurements. *Br. J. Obstetr*., **101**, 125–131.

———(1994c) Charts of fetal size: 4, femur length. *Br. J. Obstetr.*, **101**, 132–135.
Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Am. Statist. Ass.*, **74**, 829–836.
Cole, T. J. (1988) Fitting smoothed centile curves to reference data (with discussion). *J. R. Statist. Soc.* A, **151**, 385–418.
———(1990) The LMS method for constructing normalized growth standards. *Eur. J. Clin. Nutrn*, **44**, 45–60.
———(1993) The use and construction of anthropometric growth reference standards. *Nutrn Res. Rev.*, **6**, 19–50.
Cole, T. J. and Green, P. J. (1992) Smoothing reference centile curves: the LMS method and penalised likelihood. *Statist. Med.*, **11**, 1305–1319.
Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Statist.*, **7**, 1–26.
———(1991) Regression percentiles using asymmetric squared error loss. *Statist. Sin.*, **1**, 93–125.
Goldstein, H. and Pan, H. (1992) Percentile smoothing using piecewise polynomials, with covariates. *Biometrics*, **48**, 1057–1068.
Hamill, P. V. V., Drizd, T. A., Johnson, C. L., Reed, R. B. and Roche, A. F. (1977) NCHS growth curves for children birth–18 years. *Vital and Health Series 11*. National Center for Health Statistics, Washington DC.
Harris, E. K. and Boyd, J. C. (1995) *Statistical Bases of Reference Values in Laboratory Medicine*. New York: Dekker.
Healy, M. J. R., Rasbash, J. and Yang, M. (1988) Distribution-free estimation of age-related centiles. *Ann. Hum. Biol.*, **15**, 17–22.
Himes, J. H. and Hoaglin, D. C. (1989) Resistant cross-age smoothing of age-specific percentiles for growth reference data. *Am. J. Hum. Biol.*, **1**, 165–173.
Hinkley, D. V. (1975) On power transformations to symmetry. *Biometrika*, **62**, 101–111.
International Federation of Clinical Chemistry (IFCC) Panel on Theory of Reference Values (1987) The theory of reference values: Part 5, Statistical treatment of collected reference values; determination of reference limits. *J. Clin. Chem. Clin. Biochem.*, **25**, 645–656.
John, J. A. and Draper, N. R. (1980) An alternative family of transformations. *Appl. Statist.*, **29**, 190–197.
Johnson, N. L. (1949) Systems of frequency curves generated by methods of translation. *Biometrika*, **36**, 149–176.
Jones, M. C. and Hall, P. (1990) Mean squared error properties of kernel estimates of regression quantiles. *Statist. Probab. Lett.*, **10**, 283–289.
Koenker, R. and Bassett, G. (1978) Regression quantiles. *Econometrica*, **46**, 33–50.
Lawrence, C. J. and Trewin, V. F. (1991) The construction of biochemical reference ranges and the identification of possible adverse drug reactions in the elderly. *Statist. Med.*, **10**, 831–837.
Manly, B. F. J. (1976) Exponential data transformations. *Statistician*, **25**, 37–42.
Mann, J. I., Lewis, B., Shepherd, J., Winder, A. F., Fenster, S., Rose, L. and Morgan, B. (1988) Blood lipid concentrations and other cardiovascular risk factors: distribution, prevalence and detection in Britain. *Br. Med. J.*, **296**, 1702–1706.
Merkouriou, S. and Dix, D. (1988) Estimating reference ranges in clinical pathology: an objective approach. *Statist. Med.*, **7**, 377–385.
Millward, M. and Dix, D. (1992) Determining reference ranges by linear analysis. *Lab. Med.*, **23**, 815–818.
Pan, H. Q., Goldstein, H. and Yang, Q. (1990) Nonparametric estimation of age-related centiles over wide age ranges. *Ann. Hum. Biol.*, **17**, 475–481.
Parzen, E. (1979) Nonparametric statistical data modeling. *J. Am. Statist. Ass.*, **74**, 105–131.
Reiss, R. D. (1980) Estimation of quantiles in certain nonparametric models. *Ann. Statist.*, **8**, 87–105.
Rossiter, J. E. (1991) Calculating centile curves using kernel density estimation methods with application to infant kidney lengths. *Statist. Med.*, **10**, 1693–1701.
Royston, P. (1991) Constructing time-specific reference ranges. *Statist. Med.*, **10**, 675–690.
———(1992) Estimation, reference ranges and goodness of fit for the three-parameter lognormal distribution. *Statist. Med.*, **11**, 897–912.
———(1993) A toolkit for testing for non-normality in complete and censored samples. *Statistician*, **42**, 37–43.

Royston, P. and Matthews, J. N. S. (1991) Estimation of reference ranges from normal samples. *Statist. Med.*, **10**, 691–695.

Shapiro, S. S. and Francia, R. S. (1972) An approximate analysis of variance test for Normality. *J. Am. Statist. Ass.*, **67**, 215–216.

Shapiro, S. S. and Wilk, M. B. (1965) An analysis of variance test for Normality (complete samples). *Biometrika*, **52**, 591–611.

Sheather, S. J. and Marron, J. S. (1990) Kernel quantile estimators. *J. Am. Statist. Ass.*, **85**, 410–416.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Simon, N. V., O'Connor, T. J. and Shearer, D. M. (1990) Detection of intrauterine fetal growth retardation with abdominal circumference and estimated fetal weight using cross-sectional growth curves. *J. Clin. Ultrasound*, **18**, 685–690.

Slifker, J. F. and Shapiro, S. S. (1980) The Johnson system: selection and parameter estimation. *Technometrics*, **22**, 239–246.

StataCorp (1995) *Stata Reference Manual, Version 4.0*. College Station: Stata.

Stigler, S. M. (1974) Linear functions of order statistics with smooth weight functions. *Ann. Statist.*, **2**, 676–693.

Stone, C. J. (1977) Consistent nonparametric regression (with discussion). *Ann. Statist.*, **5**, 595–645.

Thompson, M. L. and Theron, G. B. (1990) Maximum likelihood estimation of reference centiles. *Statist. Med.*, **9**, 539–548.

Todros, T., Ferrazzi, E., Groli, C., Nicolini, U., Parodi, L., Pavoni, M., Zorzoli, A. and Zucca, S. (1987) Fitting growth curves to head and abdomen measurements of the fetus: a multicentric study. *J. Clin. Ultrasound*, **15**, 95–105.

Tsay, J. Y., Chen, I.-W., Maxon, H. R. and Heminger, L. (1979) A statistical method for determining normal ranges from laboratory data including values below the minimum detectable value. *Clin. Chem.*, **25**, 2011–2014.

Wade, A. M. and Ades, A. E. (1994) Age-related reference ranges: significance tests for models and confidence intervals for centiles. *Statist. Med.*, **13**, 2359–2367.

Wade, A. M., Ades, A. E., Salt, A. T., Jayatunga, R. and Sonksen, P. M. (1995) Age-related standards for ordinal data: modelling the changes in visual acuity from 2 to 9 years of age. *Statist. Med.*, **14**, 257–266.