Creating unbiased cross-sectional covariate-related reference ranges from serial correlated measurements

ANGELA WADE*
Centre for Paediatric Epidemiology and Biostatistics, Institute of Child Health,
30 Guilford Street, London WC1N 1EH, England
awade@ich.ucl.ac.uk

JUOZAS KURMANAVICIUS
Perinatal Research Unit, Department of Obstetrics, Zurich University Hospital,
Frauenklinikstrasse 10, 8091 Zurich, Switzerland

SUMMARY
Cross-sectional covariate-related reference ranges are widely used in clinical medicine to put individual observations in the context of population values. Usually, such reference ranges are created from data sets of independent observations. If multiple measurements per individual are available, then ignoring the within-person correlation between repeats will lead to overestimation of centile precision. Furthermore, if abnormal measurements have triggered more frequent assessment, the data set will be biased thus producing biased centiles. Where multiple measures per individual exist, the methods commonly used are either randomly or systematically to select one observation per individual or to model individual trajectories and combine these. The first of these approaches may result in discarding a large proportion of the available data and may itself cause bias and the latter requires the form of the changes within individuals to be characterized. We have developed an approach to the modeling of the median, spread, and skew across individuals using maximum likelihood, which can incorporate correlations between dependent observations. Heavily biased data sets are simulated to illustrate how the methodology can eliminate the biases inherent in the data collection process and produce valid centiles plus estimates of the within-person correlations. The “select one per individual” approach is shown to be liable to bias and to produce less precise centiles. We recommend that the maximum likelihood method incorporating correlations be used with existing data sets. Furthermore, this is a potentially more efficient approach to be considered when planning the future collection of data solely for the purposes of creating cross-sectional covariate-related reference ranges.

Keywords: Age-related reference ranges; Correlated measurements; Dependence; Serial measures; Unbiased; z-scores.

*To whom correspondence should be addressed.
1. Introduction

Covariate-adjusted reference ranges may be used to assess individuals at a single point in time (cross-sectional) or to monitor changes within individuals over time (velocity or conditional). Most commonly, the covariate used is age. If an individual presents for diagnosis/assessment and repeat measurements are available, then it will generally be advisable to utilize all of these. However, there are many occasions on which only a single measurement is available and this needs to be evaluated against population values using a covariate-adjusted cross-sectional reference range.

The methodologies for constructing population-based covariate-adjusted cross-sectional reference centiles are now well established and were recently reviewed by the World Health Organization (Borghi and others, 2006, for the World Health Organization [WHO] Multicentre Growth Reference Study Group). These methodologies commonly assume that the measurements used for construction are independent. If the data set contains serial measurements from individuals, then these will be correlated within person and hence the independence assumption is not satisfied. One approach has been to circumvent the problem by systematically or randomly selecting one observation per individual to create a data set of independent measurements (e.g. Kurmanavicius, Wright, Royston, Wisser and others, 1999; Kurmanavicius, Wright, Royston, Zimmermann and others, 1999; Wade and Ades, 1994). However, this is wasteful of the data and may lead to bias.

This paper is concerned specifically with the construction of cross-sectional reference ranges using serial measurements from individuals. The need for any marginal analyses to include assumptions about the form of the correlation has been well documented within other applications (Diggle and others, 2002). The Laird–Ware model (1982) gives a general framework for modeling which allows for variable spacing of observations and varying structures between individuals. The estimation of parameters for this model form has received much coverage (Davidian and Giltinan, 1995; Hand and Crowder, 1996; Vonesh and Chinchilli, 1997; Lindsey, 1999; Diggle and others, 2002). However, very little of the available literature applies specifically to the reference range problem. Within the majority of texts, characterization of the average pattern is of primary importance followed by estimation of the covariance/correlation structure between repeats within individuals where these exist. When reference ranges are to be constructed, the estimation of any skewness and the spread of values at each covariate are at least as important as quantification of the median. This shift of emphasis is necessary as it is usually the extreme centiles that are of most clinical use. By contrast, estimation of the covariance/correlation structure is generally of little or no direct interest in this scenario.

The WHO review recommended the use of methodologies that model the covariate-related changes in distributional features and then combine these to obtain centiles. Commonly, the underlying distribution is assumed to be some transformation of the normal distribution and the kurtosis, skew, spread, and median are modeled. The form of the models used for the distributional features and the mode of identifying the best fit parameters vary according to the specific method chosen (Cole, 1988, Cole and Green, 1992; Wright and Royston, 1997). Previously, we used this methodology within a maximum likelihood framework with exponential models to create age-related centiles for CD4 counts (Wade and Ades, 1994), randomly selecting one measurement per child. We subsequently extended the approach by incorporating suitable correlation structure into the likelihood and thus additionally modeled the correlation between repeats from the same individual as smooth functions of age and time (Wade and Ades, 1998), hence utilizing the entire data set. This extension may be viewed as a generalization of models developed to identify trends in longitudinal data with explicit modeling of the serial correlation (Diggle, 1988). Despite a strong correlation structure, incorporation did not substantially alter the fitted centiles. This finding was not unexpected because measurements were made at ages defined within a strict protocol, and hence, frequency of measuring was independent of previous measurements. A similar approach that has been used is systematically, as opposed to randomly, to select one measurement per individual. For example, Kurmanavicius, Wright, Royston, Wisser and others (1999) and Kurmanavicius, Wright,
Royston, Zimmermann and others (1999) used only the first of serial measurements made during pregnancy to create cross-sectional centiles for fetal biometry.

While it is generally appreciated that incorporation of dependent observations without adjustment for correlations will lead to overestimation of centile precision, the propensity for bias invalidating the centiles has received little discussion. When the number and/or timing of observations are related to outcome, for example, when an abnormal measurement is likely to trigger more frequent assessment for clinical purposes, then the incorporation of correlations may have a large impact on the centiles by reducing or removing the bias inherent in the collection process. The problem is one of informative observation times, whereby future measurement frequency is related to the values of existing measurements for that individual (Lin and others, 2004).

The WHO review (Borghi and others, 2006) identified the following 2 approaches that incorporated correlated measurements into the construction of cross-sectional reference ranges. Laird and Ware (1982) proposed 2-stage random-effects models, while Goldstein (1986) proposed a more general framework of multilevel models which could be parameterized to allow for complex covariance structures and multiple explanatory variables. Marginal distributions obtained from these models would identify cross-sectional patterns of change (Pan and Goldstein, 1997). While these models are flexible and present a solution to the specific problem posed here, they require explicit characterization of a common underlying form for expected trajectories within individuals. Goldstein and others (1994) recognized that the methodology for conditional (longitudinal) references can theoretically yield cross-sectional references. The second approach cited by the WHO review was our previously described maximum likelihood method (Wade and Ades, 1998) requiring characterization only of population changes irrespective of how individual trajectories vary.

In this paper, we illustrate how biases may be removed and precision increased via appropriate modeling even for heavily biased simulated data sets. We compare the precision with which centiles are estimated when correlations are incorporated versus the alternative systematic or random “select one” approach. The methodology is illustrated by application to serial fetal ultrasound measurements collected at the University Hospital in Zurich (UHZ) and previously modeled using only a subset of the data (Kurmanavicius, Wright, Royston, Wisser and others, 1999; Kurmanavicius, Wright, Royston, Zimmermann and others, 1999).

2. Methods

2.1 Statistical methods

In previous papers, we have demonstrated the use of splines, fractional polynomials, and exponentials within the maximum likelihood methodology. Any data collection protocol can be accommodated, as can any amount of variation between the number and timing of measurements per individual. Formal significance tests are easily performed between nested models and confidence intervals constructed for the model parameters and/or the centiles (Wade and Ades, 1994, 1998). Thompson and Fatti (1997) extended the methodology to create multivariate centile charts. In the analyses presented in this paper, we assume that a transformation of the normal distribution is appropriate at each covariate value and we model the changes in the skewness, spread, and median. We maximize the likelihood incorporating a correlation structure between repeated measurements within the same individual.

We used Fortran programs incorporating numerical algorithms group subroutines, which are available within the supplementary material, available at Biostatistics online. An alternative would be to use generalized additive models for location, scale, and shape (GAMLSS; Rigby and Stasinopoulos, 2005). The gamlss command in R can be used to fit centiles with incorporation of random effects for individuals to account for serial correlation.
2.2 Simulations

Full details and results from the simulations are given in the supplementary material, available at *Biostatistics* online. The features and findings were as follows.

Underlying models were assumed so that median and spread were both increasing with gestational age, as this would be typical in most applications. Simulations were based around a scenario often encountered during pregnancy where measurements are made between 15 and 40 weeks and abnormally low values trigger additional repeated measurements. Repeated measurements within an individual were generated with an exponentially decaying correlation function.

The extent of bias in the data sets was dependent on how the frequency of repeat measurements was determined. Fitted centiles based on an assumption of independence were heavily biased. The extent of the correlations between repeats was typically underestimated within the correlation models, although the centiles obtained were not biased. Both precision and accuracy were improved for the correlation model compared to that assuming independence.

With “select one”, the precision of the centile estimates was reduced and the centile estimates were biased. This latter finding shows that selecting a subset of independent measurements does not necessarily yield unbiased centiles. At later gestations, there were more measurements from those fetuses previously with abnormally low values and hence there was more chance of selecting biased assessments in this gestational age range.

3. Application to ultrasound data set

3.1 The data set

Ultrasound measurements were taken from clinic records of pregnant women examined at the UHZ, where routine examinations were performed at 11–13, 18–21, and 28–32 weeks of gestation. High-risk pregnancies were examined at shorter intervals, every 2 or 3 weeks until delivery. Referrals at later gestations from other ultrasound centers were also included. A relatively common reason for such referral would be suspected intrauterine growth retardation (IUGR) due to placental insufficiency which manifests after 25–28 weeks of gestation with reduced growth of the fetal abdomen. Such women then undergo repeat tests until a definitive diagnosis is made. Small values of abdominal circumference (AC) indicate potential IUGR, whereas biparietal diameter (BPD), a measure of skull size, is not expected to be affected by IUGR. The only measurements excluded from the data set were for fetuses found to have a congenital abnormality.

The original analysis used the first fetal measurements made between 12 and 42 weeks from 6557 pregnant women (6557 BPDs and 5807 ACs). Fractional polynomials were used to model age-related changes in the mean and standard deviation, and Shapiro–Francia \( W \) test was used to check the normality of the \( z \)-scores. For these 2 fetal measurements, a linear cubic in age for the mean and linear model for the standard deviation were found to be suitable (Kurmanavicius, Wright, Royston, Wisser and others, 1999; Kurmanavicius, Wright, Royston, Zimmermann and others, 1999).

The current data set, which has expanded since its use in 1999, consists of information from 12 480 women measured between 1 and 28 times. A total of 48 005 BPDs and 45 352 ACs are included. Hence, any select one approach would utilize only about 25% of the available measurements. Since the purpose was to illustrate the effects of this modeling, we used the same model forms as Kurmanavicius and others had previously (linear-cubic model for the mean, linear model for the standard deviation, and no skew) and estimated only their parameters. We allowed the correlation between repeats from the same individual to fall as the time between those repeats increased and characterized this as a 2-parameter exponential model \( (\rho_1 e^{-\rho_2 \text{diff}}) \). Hence, we estimated 5 parameters for the independence and select one models (3 for the mean and 2 for the standard deviation) and an additional 2 (for the correlation structure) for the correlation incorporated models. We compare the fitted centiles with those previously presented by
Kurmanavicius, Wright, Royston, Wisser and others (1999) and Kurmanavicius, Wright, Royston, Zimmermann and others (1999).

3.2 Fitted models

Figures 1(a) and (b) show the 5th, 50th, and 95th centiles from the independence models, the exponentially correlated models, and those previously presented by Kurmanavicius, Wright, Royston, Wisser and others (1999) and Kurmanavicius, Wright, Royston, Zimmermann and others (1999). For both AC and BPD, taking into account the correlations reduces the centile range at earlier gestations. This pattern is compatible with a greater frequency of measurement of fetuses with extreme values. The centiles fitted by Kurmanavicius, Wright, Royston, Wisser and others (1999) and Kurmanavicius, Wright, Royston, Zimmermann and others (1999) lie between the correlation and independence models at early gestations but become increasingly like the independence centiles at later ages.

The effect of incorporating correlations on the centiles at later gestations differs between AC and BPD. The increased AC 5th centile beyond 30 weeks of gestation may be explained by the inclusion of late referrals to UHZ for suspected IUGR. By contrast, BPD is a skull measurement, large values of which may be of greater concern near to term (40 weeks of gestation). The pattern of differences shown in Figure 1(b) suggests that the fetuses with larger BPD were more likely to be measured more frequently in the last 5 weeks of pregnancy. Incorporation of correlations reduces the effect of these larger measurements, and the centiles based on the correlation model are lower.

Figures 1(a) and (b) show how incorporation of correlations can remove the selection bias inherent in clinical data sets. The patterns observed were not anticipated but with hindsight have clinically valid

![Fig. 1](http://biostatistics.oxfordjournals.org/)
explanations. The results show that the adjustments for correlation will not be uniform in either direction or quantity even for seemingly highly related measurements, that is, different assessments of growth from the same ultrasounds in the same group of women. It is perhaps surprising that the centiles based on the first measurement from each fetus (Kurmanavicius, Wright, Royston, Wisser and others, 1999; Kurmanavicius, Wright, Royston, Zimmermann and others, 1999) were more akin to those obtained from the data set of all measurements. However, this is compatible with the finding of the simulation study. Selecting an independent subset does not necessarily remove bias as late referrals to UHZ are probably atypical.

4. Discussion

Our simulations and application demonstrate that incorporation of a correlation structure within the fitting algorithm is to be preferred to the select one approach. Although select one is computationally simpler, our analysis shows that the precision and accuracy of the centiles may be severely affected. The simulations were developed to represent typical clinical scenarios. The irremovable bias obtained was not expected, although easily explained with hindsight. It is important to note that such biases may occur in any data set and will yield biased centiles if select one is used.

Simulations are necessarily limited by choice of the parameters and assumptions incorporated. In particular, we assumed very similar correlation structures within the data generation and fitting phases. We do know that if the correlation structure is severely misspecified, then this will lead to invalid estimation. The comparison with the independence model of zero correlation between repeats within individuals is an extreme case and clearly illustrates this point. There must therefore be some degree of misspecification that can lead to invalidation of the centiles. For all cases where we modeled the correlation structure, we assumed exponential decline with increasing time, a reasonable assumption for clinical applications. The extent of any decline was estimated, and estimates were often biased. However, the model did allow for any extent of decline (including zero), and the simulations illustrate that biased parameter estimates do not necessarily lead to biased centile estimates. When constructing cross-sectional reference ranges from correlated data, estimation of the correlation structure is not of direct interest, but rather is a means toward the important end of obtaining unbiased estimates of centiles.

The data set we describe consists of longitudinal measurements with informative observation times. Techniques that specifically model the timing mechanism could be employed and would give additional information. However, this would necessitate the specification of the conditional distribution of an observation given the history of the process and the centile estimates may not be robust to misspecification (Lin and others, 2004). For the purposes of creating cross-sectional covariate-related reference ranges, the biases in the measurement process that we wish to eliminate are a function of the clinical scenario. Our simulations show that the method we present is capable of producing unbiased centiles from messy data sets of the form likely to be found in clinical practice. The extent and direction of centile adjustment when correlations are incorporated may yield information about the nature of biases inherent in the data set.

Despite the numerous advantages of incorporating some form of correlation, the technique has not been widely used, perhaps because of a mistaken belief that the added complexity of modeling is not warranted. In this paper, we have shown that simpler methods applied to data with informative observation times lead to invalid centile estimation. The process of selecting the first observation per individual, as previously used by Kurmanavicius, Wright, Royston, Wisser and others (1999) and Kurmanavicius, Wright, Royston, Zimmermann and others (1999), may lead to a biased solution and necessitates discarding a large proportion of the data. The resulting reduction in precision will be a function of the percentage of data discarded and the extent of the within-individual correlation. The reduction may not be uniform across the age range since selecting the first measurement from each pregnancy will lead to greater loss of precision at later gestations where there will be fewer women presenting. The UHZ received referrals...
Creating unbiased cross-sectional covariate-related reference ranges

for suspected problems, based on abnormal ultrasound measurements, from other hospitals. While fetuses subsequently found to have a congenital abnormality were excluded, biases are still likely to remain. The simulation results show that incorporating correlation structure into the modeling has the capacity to reduce biases such as these in estimating centiles. The extent of the bias reduction will depend on the degree to which the correlation structure has been adequately modeled.

If data are to be collected specifically for the purposes of creating cross-sectional reference ranges, then the most precise centiles for a given total number of observations will be obtained when these observations are independent. However, there may be a trade-off between the recruitment cost per individual and the cost of following individuals serially (Goldstein, 1979). If subsequent measurements for a recruited individual are easier and/or cheaper to obtain than measurements from new recruits, then some consideration should be given at the design stage to the most efficient way to proceed. It may be ethically easier to justify serial collection from a smaller pool of prospective subjects, or the pool may be limited (e.g. children born to HIV-1-infected mothers). Often, the remit is to produce both cross-sectional and conditional or velocity references (Borghi and others, 2006). In this case, the optimal approach will be to incorporate serial observations into the cross-sectional references using appropriate adjustment for correlation within individuals. The precision with which centiles are estimated under differing correlation structures, model forms, and/or sample sizes can be readily compared using simulations to identify the most appropriate recruitment method to use.

ACKNOWLEDGMENTS

We are grateful to Tim Cole, Harvey Goldstein, and Huiqi Pan for helpful comments and advice. Conflict of Interest: None declared.

FUNDING

Research at the Institute of Child Health and Great Ormond Street Hospital for Children National Health Service (NHS) Trust benefits from R&D funding received from the NHS executive.

REFERENCES


[Received January 12, 2007; first revision November 7, 2007; second revision April 25, 2008; third revision May 20, 2008; fourth revision June 5, 2008; accepted for publication June 6, 2008]