DOI: 10.1111/1471-0528.12031
www.bjog.org

Original article

Statistical considerations for the development of prescriptive fetal and newborn growth standards in the INTERGROWTH-21st Project

DG Altman,[a] EO Ohuma,[a,b] for the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st)

[a] Centre for Statistics in Medicine, Wolfson College Annexe, University of Oxford, Oxford, UK [b] Nuffield Department of Obstetrics & Gynaecology and Oxford Maternal & Perinatal Health Institute, Green Templeton College, University of Oxford, Oxford, UK
Correspondence: DG Altman, Centre for Statistics in Medicine, University of Oxford, Wolfson College Annexe, Linton Road, Oxford, UK, OX2 6UD, UK. Email doug.altman@csm.ox.ac.uk

Accepted 9 October 2012.

The INTERGROWTH-21st Project has in its mandate to develop prescriptive standards for fetal, neonatal and preterm post-neonatal growth. The project comprises three components: the Fetal Growth Longitudinal Study (FGLS), the Preterm Postnatal Follow-up Study (PPFS), and the Newborn Cross-Sectional Study (NCSS). We consider here the statistical aspects of the three components as they relate to the construction of these standards, in particular the sample size, and outline the principles that will guide the planned main analyses.

Keywords Sample size, longitudinal study, statistical aspects, fetal growth, prescriptive standards, INTERGROWTH-21st.

Please cite this paper as: Altman DG, Ohuma EO, for the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st). Statistical considerations for the development of prescriptive fetal and newborn growth standards in the INTERGROWTH-21st Project. BJOG 2013; 120 (Suppl. 2): 71–76.

Introduction

The primary objective of the INTERGROWTH-21st project is the production of international, multi-ethnic, standards for fetal, neonatal and preterm postnatal growth.[1] The INTERGROWTH-21st Project includes three major studies: the Fetal Growth Longitudinal Study (FGLS), the Preterm Postnatal Follow-up Study (PPFS), and the Newborn Cross-Sectional Study (NCSS).[2] The methodology of the component studies of INTERGROWTH-21st related to these objectives is broadly the same as for the World Health Organization (WHO) Multicentre Growth Reference Study (MGRS) which developed standards for infant and child growth.[3] In particular, the INTERGROWTH-21st Project has adopted the same prescriptive approach as the MGRS, with selection of healthy populations at low risk of intrauterine growth restriction (IUGR) from several countries across continents. The design and conduct of these are detailed elsewhere.[2,4]

Statistical considerations were influential in many aspects of the design of the INTERGROWTH-21st Project, for example in defining eligibility criteria, dating of pregnancies, the use of replicate ultrasound measurements, the time interval between visits, and quality control procedures, conceptually based on or benefiting from the MGRS experience where appropriate. The justification and logistical implications of these design features are addressed in the specific papers in this series. Here we focus on statistical aspects including sample size and outline the main planned analyses.

Fetal Growth Longitudinal Study (FGLS)

Sample size – initial considerations

The primary output of FGLS will be centile charts for each of seven dimensions of fetal size measured by ultrasound in relation to gestational age. The primary statistical goal was that the sample size should be large enough to yield precise estimates of extreme centiles (e.g. 3rd and 97th). That desire leaves open the question of how we define 'precise', for which there is no standard approach.

Although statistical considerations were important, certain logistical issues were critical too. Thus, for example, a key factor to consider was the number of women who could be

scanned in a centre in a week. This practical issue was important as each centre was provided with one ultrasound machine, specially adapted for the study. Likewise, the total sample size is greatly influenced by the number of centres included in the study although the final number of participating sites (countries) was not defined at the time the protocol was developed. However, the expectation of eight centres proved to be correct.

We also needed to have sufficient power to explore ethnic-specific (i.e. site-specific) growth in FGLS, in the event that ethnic differences did emerge from the data in the main growth indicators. A further consideration was that FGLS should yield an adequate number of newborns for inclusion in PPFS. The minimum target sample size of 4000 for FGLS was determined taking all of these criteria into account. To obtain complete data from 4000 pregnancies and their newborns, 500 mothers on average would have to be enrolled at each of the eight study sites.

The chosen sample size is larger than most previous studies even if each site is considered separately and is thus adequate to produce reliable curves and to explore variability between countries. We estimated that fewer than 5% of women would be lost to follow-up. We also acknowledged that about 10% of women may be excluded from the development of the fetal growth standards because they will have developed complications of pregnancy severe enough to affect fetal growth, as identified in the protocol.[4]

Thus the initial sample size estimation for FGLS was informed by, but not determined only by, statistical considerations. For this reason we have reconsidered the adequacy of the sample size.

Sample size revisited

Due to the complexity of determining the required sample size for growth reference studies a WHO Expert committee, as early as 1995, recommended, as a rule of thumb, a sample of at least 200 individuals in each age group and sex.[5] However, the concept of age groups does not apply to a fetal growth study like FGLS as all the data will be considered in a single analysis,[6] and fetal sex differences will not be explored. In clinical practice fetal sex is not always determined during pregnancy and therefore growth charts have not been gender specific, especially as. this information is never divulged in some cultures. Also, it has been shown that the differences in birth weight between males and females are very small and we expect the differences in fetal growth during pregnancy to be negligible anyway. Therefore, the minimal sample size per site was calculated without taking fetal sex or age group into account.

We considered the sample size for FGLS in relation to the precision and accuracy of a single centile and regression based reference limits, as first proposed by Royston[7] and extended by Bellera and Hanley.[8] Fetal size measurements tend to be close to a normal distribution at each specific gestational age[9]. Data that are normally distributed can be summarised using the mean and standard deviation from which each required centile can be estimated.

Table 1. A summary table relating sample size to precision expressed in SD at selected centiles.

Sample size	Precision achieved at 2.5th or 97.5th centile in SD	Precision achieved at 5th or 95th centile in SD	Precision achieved at 10th or 90th centile in SD
500	0.08	0.07	0.06
1000	0.05	0.05	0.04
1500	0.04	0.04	0.03
2000	0.04	0.03	0.03
2500	0.03	0.03	0.03
3000	0.03	0.03	0.02
3500	0.03	0.03	0.02
4000	0.03	0.02	0.02
4500	0.03	0.02	0.02
5000	0.02	0.02	0.02
5500	0.02	0.02	0.02
6000	0.02	0.02	0.02

The standard error of the P^{th} centile is given by the standard formula for sampling variance of a centile of normal distribution:[10]

$$SE_p = SD\sqrt{[(1 + \frac{1}{2}Z_p^2)/n]}$$

where SE is the standard error, SD is the standard deviation of the measurement (which will increase with gestational age), Z_p is the value of the standard normal distribution corresponding to the P^{th} centile, and n is the sample size. So, for example, for the 2.5th or 97.5th centile $Z_p = 1.96$, giving SE = 0.08SD for a sample size of 500 and 0.03SD for 4000 fetuses (Table 1). Our sample size calculation was based on a cross-sectional design and, bearing in mind that the distance between the 2.5th and 97.5th centiles is about 4SD, it is clear that even at these extreme centiles, fetal size will be estimated with great precision with 500 fetuses per site. These standard errors are, however, overestimated as they ignore the fact that there will be a series of measurements from each fetus.[11] Sample size calculations for growth standards based on longitudinal data are complicated and supported by a rather limited literature.[8,12] Royston defined a design factor, D, as the number of fetuses in a cross-sectional study that would give the same precision as a longitudinal study. Using ultrasound based biparietal diameter the design factor was suggested to be approximately 2.3.[12] Based on that value, the longitudinal component of FGLS with 4000 fetuses would have equivalent precision to a cross-sectional study of over 9000 fetuses.

Analysis strategy of FGLS

The INTERGROWTH-21st analysis strategy and methods for the construction of growth standards will be conceptually similar to those applied in the WHO MGRS,[3] which provides a very useful starting point for the analysis although fetal data are likely to be simpler to model.

Data from all sites will be carefully explored and evaluated to determine if they concur with the assumptions and criteria used by MGRS before adoption of the analysis. It is envisaged that whenever these criteria or assumptions are violated, then further analysis methods or modifications will be applied to the data as deemed appropriate.

It is desirable to be able to use all the data from the eight study sites to provide a single global standard for each measurement and to give the strongest basis for the construction of growth curves for international clinical applications. However, it is important to be satisfied that the data from the different centres are similar enough to be combined. Unfortunately, there is no universally recognised way in which to make a judgment on the acceptable amount of heterogeneity of growth data from several sites. Data from all sites will thus be compared using pre-specified criteria, described below, to determine whether it is reasonable to pool all the data.

We will follow the same biological and statistical strategy applied by MGRS in deciding the combinability of the data from each country. The main growth measures for comparisons between sites, as adopted by MGRS, will be fat-independent measures of linear growth, namely crown rump length (CRL) for early linear fetal size and head circumference (HC) for fetal growth after 14 weeks of gestation. HC is the most suitable parameter for comparisons across ethnic/environmental conditions because the head is the last structure to be affected by external factors influencing fetal growth – the so-called 'brain sparing effect'.[13] Furthermore, it allows continuous evaluation of the same measurement in the post-natal period.

The appropriateness of pooling data from all sites to construct CRL and HC standards will be assessed by comparing site means, standard deviations and the fitted centiles from the analysis of each site to the corresponding values from analyses of data from all sites combined. In particular, a difference of ≥ 0.5 SD between the values for an individual site and the pooled sample (as adopted in MGRS[14]) will be used as a pre-set trigger for considering whether to adjust by site for the purposes of pooling data. That decision would depend on the magnitude and nature of the discrepancies between the data from a site and the total data set. We note that the difference between the 5th and 2.5th centiles is less, 0.32 SD compared to 0.50 SD, and this smaller value might be used as a second, stricter criterion. Furthermore, we will conduct a sensitivity analysis exploring the effect on the pooled mean at different gestational ages and the estimated regression models of removing each of the populations in turn.[14]

Derivation of centile charts

Reference centiles should change smoothly with gestation, and they should provide a good fit to the raw data. It is desirable for the statistical model to be as simple as is compatible with these requirements.[15] In preparation for analysis of MGRS, WHO conducted an extensive literature review of existing methods for the construction of growth curves.[16] A group of statisticians and child growth experts then agreed on the methodology to be used to develop the international infant and child growth charts.

Briefly, WHO adopted a semi-parametric approach by fitting class growth distributions (parametric) for all measurements and applying an appropriate smoothing technique i.e. cubic splines and fractional polynomials (non-parametric) to generate centiles.[17] Five candidate distributions were proposed by the WHO-MGRS experts and were evaluated based on their flexibility and goodness of fit. The Box-Cox power exponential (BCPE) method with four parameters provided the best fit with curve smoothing by cubic splines and was thus selected as the most appropriate method to construct their growth curves.[3,16,18] Villandre et al.[19] have also compared a flexible multi-level spline-based model with other approaches for modelling fetal weight by gestational age. These approaches may be relevant for our analysis and will thus be considered.

The same statistical models would be suitable for analyses of the FGLS data if required. However, we know from many previous studies that fetal size changes smoothly and systematically over gestation, and that the distribution of size is close to normal for any gestational age. We will thus initially apply simpler models, based on fractional polynomial regression functions for the mean and SD of each fetal measurement (e.g. HC) and assuming normality at each gestational age,[9] and only move to the more complex models described above if the fit is seen to be inadequate. Analyses will be performed using STATA software (StataCorp, College Station, Texas, USA).

The distributions of residuals for the fitted centiles for each fetal measurement will be examined both for all sites combined and for each site separately, and plotted against gestational age.

The focus will be on quantification rather than hypothesis testing, including normal Q-Q plots of the fitted z-scores,[20] age-related departures from normality evaluated using worm plots[21] and comparisons of observed percentages that occur above or below estimated centiles against expected.

FGLS is a longitudinal study in which all women should have between two and six sets of ultrasound measurements after the dating scan (most will have four to five measurements) that will be included in the development of

the fetal growth charts. Data from all preterm births, in the absence of one of the severe maternal or fetal exclusion criteria, will contribute to the fetal growth standards until the time of delivery. We will use hierarchical modelling in the statistical software STATA[22] to take proper account of the multi-level structure (between and within fetus variability). We will restrict the analysis to post-dating measurements made between 14^{+0} and 40^{+0} weeks of gestation. This upper limit has been decided because fetal biometry measurements beyond 40 weeks are difficult to make and women attending for ultrasound visits beyond 40 weeks are likely to be isolated cases needing special care. In our study, only a few sites recorded any measurements beyond 40 weeks.[23]

Handling of repeated anthropometric and ultrasound measures

Anthropometric and ultrasound measurements for all newborns and fetuses are taken in duplicate and triplicate respectively. This is important to ensure data quality and to estimate within and between variation among sonographers and anthropometrists. For the analysis, we will compute the average of each set of duplicate or triplicate measurements. In theory this approach tends to underestimate the actual variability for single measurements. In clinical practice, these measurements are only made once because of clinical work load and this single measure is plotted in the standards. We will therefore explore the effect of making a small correction to the observed variability of all repeated measures as previously suggested by Bland and Altman.[24] However, this correction was developed in a different context and so the effect on our data is as yet unknown.

Sensitivity analyses

We will conduct various sensitivity analyses to assess the impact on the results of alternative approaches to the analysis. For example, we will quantify the impact of ignoring the non-independence of multiple measurements of the same fetus over time (i.e. treating all data as independent, as if from a cross-sectional study) and similarly we will evaluate the impact of different ways of treating the three replicate readings per measurement on each ultrasound session.

A further sensitivity analysis will be conducted on the effect of excluding ultrasound images that are found to be of low quality, as part of the quality control process described in another paper in this series.[25] We will also examine the impact on the estimated growth centiles of excluding ultrasound measurements outside the range 14^{+0} and 40^{+0} weeks of gestation.

Newborn Cross-Sectional Study (NCSS)

The INTERGROWTH-21st Project includes the cross-sectional study (NCSS) in which newborn data were collected at birth, as well as pregnancy data from the medical records, in all eight participating hospitals over a fixed period of time (approximately one year). Within NCSS, we first sought to identify an 'FGLS-like' population so as to construct newborn growth standards, by applying the same inclusion criteria used to screen women for FGLS, which were matched to corresponding questions in the Pregnancy and Delivery form completed at birth in the NCSS.

The main objectives of NCSS are (1) to create standards at birth for weight, length and HC according to gestational age and (2) to explore possible risk factors for stillbirth, preterm delivery and impaired fetal growth focusing specifically on their phenotypic sub-types.

To create the newborn standards, pregnancies with severe morbidities as defined for the FGLS cohort will be excluded. These include all multiple births, congenital malformations and fetal deaths. In addition, any mother who took up smoking or used recreational drugs during the index pregnancy or who reported a malaria episode, eclampsia or severe preeclampsia, malignancy, HIV/AIDS or cardiac disease will be excluded, as is the case for the FGLS cohort.

We classified 'FGLS-like' NCSS and 'non-FGLS-like' NCSS cohorts into 5 groups based on estimation of gestational age by ultrasound examination (US) and first day of the last menstrual period (LMP):

1) If an early ultrasound examination included CRL at $<15^{+0}$ weeks and/or HC or BPD at $\leq 24^{+0}$ weeks then the gestational age at delivery based on ultrasound information was used for the analyses.
2) For women who had an ultrasound measurement at $>24^{+0}$weeks, the ultrasound estimation was used if gestational age estimated by both LMP and ultrasound agreed within 7 days.

Both these groups of women are considered to have a reliable gestational age and length of pregnancy. By contrast, three further sub-populations are not considered to have reliable gestational age and so will not be included in the sample to construct the newborn standards or the phenotypic characterisation of preterm delivery and IUGR:

3) Women with an ultrasound at $>24^{+0}$ weeks with a discrepancy >7 days between estimated gestational age by LMP and ultrasound.
4) Women with an ultrasound at $>24^{+0}$ weeks and no LMP.
5) Women with no ultrasound measurements.

We aim to obtain very detailed information from 56 000 pregnancies and their newborns across eight centres to provide a sample of about 20 000 low-risk eligible pregnancies with a similar risk profile at baseline as those enrolled in FGLS. This 'FGLS-like' population, with reliable gestational age estimated by early ultrasound, constitutes the population for developing newborn anthropometric standards.

Standards for weight, length and HC at birth according to gestational age will be derived using broadly similar methods to those for FGLS. We will model newborn size in relation to both gestational age at birth and time since estimated conception.

Fetal growth standards and birth weight for gestational age standards will be related to perinatal outcomes to establish risk levels associated with different growth patterns. The 'ideal' outcome is perinatal mortality but its anticipated infrequent occurrence in this low-risk population (about 10 per 1000) makes it unrealistic to have a sample large enough for the necessary number of events across the gestational age distribution. We therefore decided to use a composite of severe neonatal morbidity and mortality. The Severe Perinatal Morbidity and Mortality Index identifies newborns with at least one of the following conditions: stillbirth, neonatal death occurring up to hospital discharge of the newborn, and newborn stay in neonatal intensive care unit (NICU) for 7 days or more. Our recent studies[26,27] have shown that it requires limited standardisation of clinical diagnoses across hospitals and is well accepted as a marker in large, international, population based studies of severely ill newborns.[28–30] Data from our previous population-based studies in some of the sites indicate that the incidence of this outcome is approximately 5%.

Preterm Postnatal Follow-up Study (PPFS)

A cohort of 'healthy' preterm babies will be recruited from the FGLS population for PPFS. The identification of 'healthy' preterm newborns is a conceptual and clinical challenge that we have discussed in the context of this study and further details of this concept have been published by Villar et al[1]. It has also sample size implications as not all of the selected newborns will be 'late' preterms ($>34^{+0}$ weeks but $<37^{+0}$ weeks). We recognised that the sample size will be influenced by logistic issues and the need to obtain the preterm newborns from the FGLS cohort to be part of PPFS rather than on statistical calculations alone. However, it is a unique cohort as it has documented fetal growth patterns and it is still large by preterm study standards. We shall have very detailed follow-up data, increasing the power of the sample for creating charts. We consider that the possibility of having a full set of fetal and newborn growth patterns from a cohort of preterm newborns is important even if we shall not have the power to explore gestational age sub-groups or early postnatal morbidity.

The analysis of this study has two components: (1) the development of growth standards specific for preterm newborns ($\geq 26^{+0}$ but $<37^{+0}$ weeks), (2) exploratory analysis of the relationship between fetal growth patterns and preterm birth phenotypes. The development of growth standards for this study will follow the same strategy and methodology as those used for FGLS, but we will need to present postnatal growth centiles for infant age accounting for gestational age at birth.

Discussion

The principal, initial statistical element of INTERGROWTH-21st was the question of an adequate sample size. Because the chosen study size also took account of several other factors, especially logistical ones, in this paper we have reconsidered the sample size from a more purely statistical perspective. We have shown that the sample size for FGLS is more than adequate to allow reliable fetal growth centiles to be derived.

We have also outlined the strategy for analysing the data from each of the three studies within INTERGROWTH-21st. Of particular clinical and methodological interest is the question of how to decide whether the data from the eight centres can be combined. While it is clearly desirable for this international study to yield a single set of growth standards, we recognise that some between-country differences may be too large to sustain such a unified approach. There is no methodology to address this issue. Indeed, even the issue of how one should best compare sets of centiles across a range of gestational ages is not easily answered.

We have outlined the steps we will take to compare the data from the different countries based on the criteria used in MGRS. In essence we believe comparisons should be based on the distance between centiles; certainly there is no place here for statistical significance. That said, we think it is neither possible nor desirable to pre-specify fully what amount of heterogeneity is acceptable. We have described our criteria for determining whether the discrepancies between countries might consistently be too large. Should one of these criteria be met a decision will be made based on the exact nature of the variation between centres, using both statistical considerations, biological plausibility and clinical judgement, and bearing in mind the need to balance the advantages of having a single standard against the consequences of pooling heterogeneous data.

Among the statistical methods that we will apply to the various data sets arising from the component studies, the most complex task is deriving growth centiles. It is essential to apply sound statistical methods that give an excellent fit to the actual data. The approach used to derive centiles in MGRS was excellent but perhaps too complex for fetal growth data. We expect that we can simplify the methods for fetal data based on successful analyses of previous studies. We will, however, examine the fit very carefully, and extend the models if necessary. We will also take proper account of key aspects of the study design – specifically we will deal appropriately with the fact that all ultrasound measurements were taken in triplicate and that we will have longitudinal data.

Acknowledgments

A full list of Members of the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st) and its committees appears in the preliminary pages of this supplement.

Disclosure of interests

None.

Contribution to authorship

DG Altman and EO Ohuma wrote the manuscript.

Details of ethics approval

The INTERGROWTH-21st Project was approved by the Oxfordshire Research Ethics Committee 'C' (reference: 08/H0606/139), and the research ethics committees of the individual participating institutions and corresponding health authorities where the Project was implemented.

Funding

This Project was supported by the INTERGROWTH-21st Grant ID# 49038 from the Bill & Melinda Gates Foundation to the University of Oxford, for which we are very grateful. DG Altman is supported by a programme grant from Cancer Research UK (C5529).

References

1 Villar J, Knight HE, de Onis M, Bertino E, Gilli G, Papageorghiou AT, et al. for the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st). Conceptual issues related to the construction of prescriptive standards for the evaluation of postnatal growth of preterm infants. *Arch Dis Child*, 2010;95:1034–8.

2 Villar J, Altman DG, Purwar M, Noble JA, Knight HE, Ruyan P, et al. for the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st). The objectives, design and implementation of the INTERGROWTH-21st Project. *BJOG* 2013; DOI: 10.1111/1471-0528.12047.

3 WHO Multicentre Growth Reference Study. WHO Child Growth Standards based on length/height, weight and age. *Acta Pædiatrica Suppl* 2006;450:76–85.

4 International Fetal and Newborn Growth Consortium. The International Fetal and Newborn Growth Standards for the 21st Century (INTERGROWTH-21st) Study Protocol, 2009. www.intergrowth21.org.uk. Accessed 7 February 2012.

5 World Health Organisation. Physical status: the use and interpretation of anthropometry. Report of a WHO Expert Committee. *World Health Organ Tech Rep Ser* 1995;854:1–452.

6 Cole TJ. The international growth standard for preadolescent and adolescent children: statistical considerations. *Food Nutr Bull* 2006;27(4 Suppl Growth Standard):S237–43.

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

8 Bellera CA, Hanley JA. A method is presented to plan the required sample size when estimating regression-based reference limits. *J Clin Epidemiol* 2007;60:610–5.

9 Silverwood RJ, Cole TJ. Statistical methods for constructing gestational age-related reference intervals and centile charts for fetal size. *Ultrasound Obstet Gynecol* 2007;29:6–13.

10 Healy MJ. Notes on the statistics of growth standards. *Ann Hum Biol* 1974;1:41–6.

11 Localio AR, Berlin JA, Ten Have TR, Kimmel SE. Adjustments for Center in Multicenter Studies: an Overview. *Ann Intern Med* 2001;135:112–23.

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

13 Baschat AA. Fetal responses to placental insufficiency: an update. *BJOG* 2004;111:1031–41.

14 WHO Multicentre Growth Reference Study. Assessment of differences in linear growth among populations in the WHO Multicentre Growth Reference Study. *Acta Pædiatrica Suppl* 2006;450:56–65.

15 Altman DG, Chitty LS. Charts of fetal size: 1. Methodology. *BJOG* 1994;101:29–34.

16 Borghi E, de Onis M, Garza C, Van den Broeck J, Frongillo EA, Grummer-Strawn L, et al. Construction of the World Health Organization child growth standards: selection of methods for attained growth curves. *Stat Med* 2006;25:247–65.

17 Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape. *J R Stat Soc Ser C (Appl Stat)* 2005;54:507–54.

18 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.

19 Villandre L, Hutcheon JA, Trejo ME, Abenhaim H, Jacobsen G, Platt RW. Modeling fetal weight for gestational age: a comparison of a flexible multi-level spline-based model with other approaches. *Int J Biostat* 2011;7(1):32.

20 Royston P, Wright EM. Goodness-of-fit statistics for age-specific reference intervals. *Stat Med* 2000;19:2943–62.

21 van Buuren S, Fredriks M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Stat Med* 2001;20:1259–77.

22 Rabe-Hesketh S, Skrondal A. 2008. *Multilevel and Longitudinal Modeling Using Stata*, 2nd edn. College Station, TX: Stata Press.

23 Altman DG, Chitty LS. New charts for ultrasound dating of pregnancy. *Ultrasound Obstet Gynecol* 1997;10:174–91.

24 Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*, 1999;8:135–60.

25 Sarris I, Ioannou C, Ohuma EO, Altman DG, Hoch L, Cosgrove C, et al. for the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st). Standardisation and quality control of ultrasound measurements taken in the INTERGROWTH-21st Project. *BJOG* 2013; DOI: 10.1111/1471-0528.12315.

26 Villar J, Valladares E, Wojdyla D, Zavaleta N, Carroli G, Velazco A, et al. Caesarean delivery rates and pregnancy outcomes: the 2005 WHO global survey on maternal and perinatal health in Latin America. *Lancet* 2006;367:1819–29.

27 Villar J, Abalos E, Carroli G, Giordano D, Wojdyla D, Piaggio G, et al. Heterogeneity of perinatal outcomes in the preterm delivery syndrome. *Obstet Gynecol* 2004;104:78–87.

28 Wapner RJ, Sorokin Y, Thom EA, Johnson F, Dudley DJ, Spong CY, et al. Single versus weekly courses of antenatal corticosteroids: evaluation of safety and efficacy. *Am J Obstet Gynecol* 2006;195:633–42.

29 Joseph KS, Fahey J, Platt RW, Liston RM, Lee SK, Sauve R, et al. An Outcome-based Approach for the Creation of Fetal Growth Standards: do Singletons and Twins Need Separate Standards? *Am J Epidemiol* 2009;169:616–24.

30 Hannah ME, Hannah WJ, Hewson SA, Hodnett ED, Saigal S, Willan AR. Planned caesarean section versus planned vaginal birth for breech presentation at term: a randomised multicentre trial. *Lancet* 2000;356:1375–83.

