Ultrasound Obstet Gynecol 2012; **39**: 266–273
Published online in Wiley Online Library (wileyonlinelibrary.com). **DOI:** 10.1002/uog.10082

Intra- and interobserver variability in fetal ultrasound measurements

I. SARRIS*, C. IOANNOU*, P. CHAMBERLAIN*, E. OHUMA*†, F. ROSEMAN*, L. HOCH*, D. G. ALTMAN† and A. T. PAPAGEORGHIOU*; for the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st)

**Oxford Maternal and Perinatal Health Institute (OMPHI), Green Templeton College and Nuffield Department of Obstetrics and Gynaecology, University of Oxford, John Radcliffe Hospital, Oxford, UK; †Centre for Statistics in Medicine, University of Oxford, Wolfson College Annexe, Oxford, UK*

KEYWORDS: fetal biometry; reproducibility; ultrasound measurement; variability

ABSTRACT

Objective *To assess intra- and interobserver variability of fetal biometry measurements throughout pregnancy.*

Methods *A total of 175 scans (of 140 fetuses) were prospectively performed at 14–41 weeks of gestation ensuring an even distribution throughout gestation. From among three experienced sonographers, a pair of observers independently acquired a duplicate set of seven standard measurements for each fetus. Differences between and within observers were expressed in measurement units (mm), as a percentage of fetal dimensions and as gestational age-specific Z-scores. For all comparisons, Bland–Altman plots were used to quantify limits of agreement.*

Results *When using measurement units (mm) to express differences, both intra- and interobserver variability increased with gestational age. However, when measurement of variability took into account the increasing fetal size and was expressed as a percentage or Z-score, it remained constant throughout gestation. When expressed as a percentage or Z-score, the 95% limits of agreement for intraobserver difference for head circumference (HC) were ± 3.0% or 0.67; they were ± 5.3% or 0.90 and ± 6.6% or 0.94 for abdominal circumference (AC) and femur length (FL), respectively. The corresponding values for interobserver differences were ± 4.9% or 0.99 for HC, ± 8.8% or 1.35 for AC and ± 11.1% or 1.43 for FL.*

Conclusions *Although intra- and interobserver variability increases with advancing gestation when expressed in millimeters, both are constant as a percentage of the fetal dimensions or when reported as a Z-score. Thus, measurement variability should be considered when interpreting fetal growth rates. Copyright © 2012 ISUOG. Published by John Wiley & Sons, Ltd.*

INTRODUCTION

In addition to estimation of gestational age[1,2] and screening for anomalies[3], fetal ultrasound measurements are commonly used for monitoring fetal growth[4]. In a mixed-risk obstetric unit it is not uncommon for 20% of women to have third-trimester scans for growth[5], and in practice these usually involve different observers. Reproducibility of third-trimester results is important, as this is the period when growth assessment is most likely to influence clinical decisions; for example, whether to deliver a fetus with suspected fetal growth restriction (FGR).

At least 60% of neonatal deaths worldwide are associated with low birth weight[6]. Identification of growth-restricted fetuses is therefore clinically important. Inaccurate measurements can lead to erroneous detection of FGR and macrosomia (false positives), and thus to unnecessary intervention, maternal anxiety and iatrogenic perinatal morbidity; or may lead to inadvertently overlooking growth-restricted fetuses and classifying them as normal (false negatives)[7].

It is therefore surprising that relatively few large and robust studies have assessed the variability of ultrasound measurements in fetal biometry by different observers. When antenatal ultrasound examination was being evaluated initially, the accuracy of fetal measurements was investigated in a number of studies[8–11]. However, not all biometric parts were assessed in every study; scans were performed in relatively small numbers (range, 13–106), and the ultrasound equipment used is now obsolete. More recent studies[12,13] included a limited number of pregnancies in the third trimester because their aim was to assess reproducibility in estimating gestational age. Some studies have tried to address the question of accuracy in late gestation by comparing measurements and estimation of fetal weight based on ultrasound examination to

Correspondence to: Dr A. T. Papageorghiou, Nuffield Department of Obstetrics and Gynaecology, University of Oxford, John Radcliffe Hospital, Oxford, OX3 9DU, UK (e-mail: aris.papageorghiou@obs-gyn.ox.ac.uk)

Accepted: 6 August 2011

those obtained postnatally[14]. This is of course a different question, i.e. assessing the accuracy of weight estimation equations. Finally, there are no solid data about the measurement variability for each fetal biometric part throughout pregnancy in relation to biological variability.

In this study we assessed the variability, under standardized conditions throughout pregnancy, of fetal ultrasound measurements within and between observers of the same fetus on the same occasion. A secondary aim was to identify factors contributing to this variability.

METHODS

The International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st) is a large-scale, population-based, multinational observational project including the monitoring of fetal and newborn growth in eight countries across the world (www.intergrowth21.org.uk). One of the component studies, the Fetal Growth Longitudinal Study (FGLS), involves two-dimensional serial fetal growth scans every 5 weeks from approximately $14+0$ to $41+6$ weeks in very low-risk women with certain gestational age estimation. Women participating in the study on observer variability have low-risk pregnancies that fulfil well defined and strict inclusion criteria at recruitment, details of which are available at www.intergrowth21.org.uk (follow link to 'Study Protocol' and download Study Protocol[15]). Briefly, inclusion criteria were maternal age between 18 and 35 years, body mass index (BMI) ≥ 18.5 and < 30 kg/m^2, a singleton pregnancy, a known date of last menstrual period (LMP) and regular cycles (defined as 28 ± 4 days) without hormonal contraceptive use or breastfeeding during the 2 months before pregnancy, natural conception, normal pregnancy history without relevant past medical history, no evidence of socioeconomic constraints likely to impede fetal growth, no use of tobacco or recreational drugs and no heavy alcohol consumption. For eligible women, according to the above screening criteria, an estimation of gestational age is made according to a standardized ultrasound measurement of crown–rump length (CRL) between $9+0$ and $14+0$ weeks. If the difference in gestational age estimation based on CRL[16] and LMP is ≤ 7 days, the women are eligible and the gestational age (deduced from LMP) is considered to be reliable.

Between February and August 2010, intra- and inter-observer variability of fetal ultrasound biometry measurement was assessed in one INTERGROWTH-21st center (Oxford). To assess the variability throughout pregnancy with an equal degree of accuracy, a minimum of 25 fetuses were recruited for every 5-week gestational age window. In total, 175 cases were scanned. All ultrasound examinations were performed using the same commercially available ultrasound machine (Philips HD-9, Philips Ultrasound, Bothell, WA, USA) with curvilinear abdominal transducers (C5-2 and V7-3). For the purposes of the INTERGROWTH-21st study, the software was programmed by the manufacturer so that observers would be blind to the fetal measurements, i.e. values obtained do not appear on screen during the scan. The INTERGROWTH-21st study was approved by the Oxfordshire Research Ethics Committee C and all pregnant women involved in the study gave written informed consent.

Three experienced sonographers (I.S., C.I. and P.C.) performed all ultrasound scans. Sonographers worked in one of three possible pairs; the order and the two sonographers to be paired were determined from a computer-generated randomization list. For each fetus the first sonographer to perform the scan was referred to as Observer 1 and the second as Observer 2 (O1 and O2). The randomization was aimed at ensuring that the three sonographers would scan approximately two thirds of the fetuses, acting as either O1 or O2 for approximately equal times. During each scan visit, the woman was first scanned by O1 and then the scan was repeated by O2. Only one observer was present in the room at any one time, and all observers were blinded to all measurements as these did not appear on screen during the scan. A strict protocol was followed: each observer performed two complete sets of measurements consisting of one head image for recording the biparietal diameter (BPD), occipitofrontal diameter (OFD) and head circumference (HC) using the ellipse facility (after removing the calipers used for the previous measurements), one abdominal image for recording the anteroposterior abdominal diameter (APAD), transverse abdominal diameter (TAD) and abdominal circumference (AC) using the ellipse facility (after removing the calipers used for the previous measurements), and one thigh image for recording the femur length (FL). A complete set of 14 stored measurements for each examination by each observer consisted of six head measurements (two each for BPD and OFD and two for HC using the ellipse), six abdominal measurements (two each for APAD and TAD and two for AC using the ellipse), and two measurements of the femur (for FL). Thus, a total of 28 measurements by both observers were taken for each fetus.

Detailed definitions of the methodology for these measurements are available at www.intergrowth21.org.uk (follow link to 'Study Protocol' and download Ultrasound Manual[15]). Briefly, head measurements were taken in the trans-thalamic plane and measured 'outer to outer', i.e. with the intersection of the calipers placed on the outer border of the parietal (BPD), occipital and frontal bones (OFD) or on the outer border of the skull (HC) using ellipse facility. Abdominal measurements were taken with the umbilical vein in the anterior third of a transverse section of the fetal abdomen at the level of the portal sinus, with the stomach bubble visible and with the intersection of the calipers placed on the outer borders of the body outline (skin) for APAD and TAD (at 90° to the APAD, across the abdomen at the widest point) or AC using the ellipse facility by placing the line of the ellipse on the outer border of the abdomen. For FL, the femur closest to the probe was measured with its long axis as horizontal as possible. Calipers were placed on the outer borders of the diaphysis of the femoral bone ('outer to outer') and excluding the trochanter. For all measurements the area of interest had

to fill at least 30% of the monitor. Only when all of these conditions were met were images acquired and measured.

For each biometric part the blinded measurements were stored electronically directly onto the machine's hard drive along with the corresponding still images. Measurements and images were retrieved for analysis after the end of the data collection. In addition to the ellipse measurement, HC was calculated from the head diameters using the formula 0.5π (BPD + OFD) and AC was calculated from the abdominal diameters using the formula 0.5π (TAD + APAD).

Sources of variability

In addition to inter- and intraobserver variability, we collected data to enable exploration of other possible factors that influence variability in fetal size measurements.

Caliper placement

In order to ascertain the variability associated with caliper placement, images were retrieved a month after completing the collection of all cases for further analysis on the same ultrasound machine. For each of the 175 cases the first of the two head, abdominal and femur images obtained by O1 were retrieved. Calipers were removed from these images and the two observers repeated the complete set of seven biometric measurements in a blinded fashion. Thus, remeasurements by O1, along with the original measurements, provided the data for the intraobserver caliper placement while the new measurements by O2, compared to the original measurements by O1, provided the data for the interobserver caliper placement. For this exercise on caliper remeasurement the observers were blinded to the identity of the sonographer who originally acquired the image.

Other factors

During each scan, each observer was asked to document the fetal presentation and placental position along with giving a subjective assessment of the degree of fetal mobility during the scan (1, active fetus; 2, quiet fetus or 3, unable to comment). Finally, maternal BMI was recorded.

Statistical analysis

The intra- and interobserver comparisons for each fetal biometric part were assessed using the four measurements taken in each fetus (two by O1 and two by O2). Intraobserver variability was assessed by calculating the differences between the two measurements made by the same observer on the same fetus (175 pairs each for O1 and O2). Interobserver variability was assessed by calculating the differences between the means of the two measurements made by the two observers on the same fetus ($n = 175$). The resultant standard deviation (SD) values of the differences of the means were then corrected to obtain the equivalent value for single measurements by using the formula proposed by Bland and Altman[17]. Measurement differences were converted into percentage differences, calculated as the difference between the measurements by the two observers divided by the average of the two measurements multiplied by 100. Measurement differences were also converted into a *Z*-score, using published data, by dividing each one by the corresponding standard deviation of that specific fetal measurement for that gestational age[18–20]. Intra- and interobserver measurement variability was thus expressed as differences between values in measurement units (mm), in percentages and in *Z*-scores, and the corresponding mean differences and limits of agreement are presented graphically using Bland–Altman plots[17].

To ascertain the intra- and interobserver variability of caliper placement, the three values (two for O1 and one for O2) from the caliper placement exercise were used; intra- and interobserver measurement variability was expressed in the same terms as above.

In order to ascertain whether fetal presentation or activity, or maternal BMI, contribute to measurement variability, the pregnancies were divided into the following categories: cephalic vs. non-cephalic, active vs. quiescent and maternal BMI of 18.5–24.9 vs. 25.0–29.9; and the corresponding *Z*-scores were compared using Student's unpaired *t*-test.

All plots and analyses were performed using STATA 11 (StataCorp, College Station, Texas, USA).

RESULTS

A total of 175 consecutive scans of 140 fetuses were included (29 were scanned twice, and three were scanned three times at different gestational ages, at least 5 weeks apart). Mean maternal age was 29.4 (range, 19.2–35.0) years and mean BMI was 23.2 (range, 18.8–29.8). In four fetuses (all > 30 weeks' gestation) head measurements were not obtained by any observer as the fetal presentation and position precluded acquisition of the appropriate planes. A total of 4852 measurements were obtained (684 of BPD, OFD and HC, and 700 of TAD, APAD, AC and FL). There was no statistically significant difference between sonographers in the measurements performed on the same fetus ($P = 0.13$, 0.54 and 0.51 for the three possible pairs; Student's paired *t*-test).

Figure 1 depicts Bland–Altman plots for the intra- and interobserver variability for HC (using the ellipse facility) for differences in measurement units (mm) (Figure 1a and b), for differences expressed in terms of percentage (Figure 1c and d), and for *Z*-scores (Figure 1e and f). The same plots are shown for AC using the ellipse facility (Figure 2) and FL (Figure 3). Plots for HC and AC using the diameters method (BPD and OFD, and APAD and TAD, respectively) were almost identical to those obtained when the ellipse facility was employed (data not shown).

When using measurement units (mm), the variability in intra- and interobserver differences increased with the

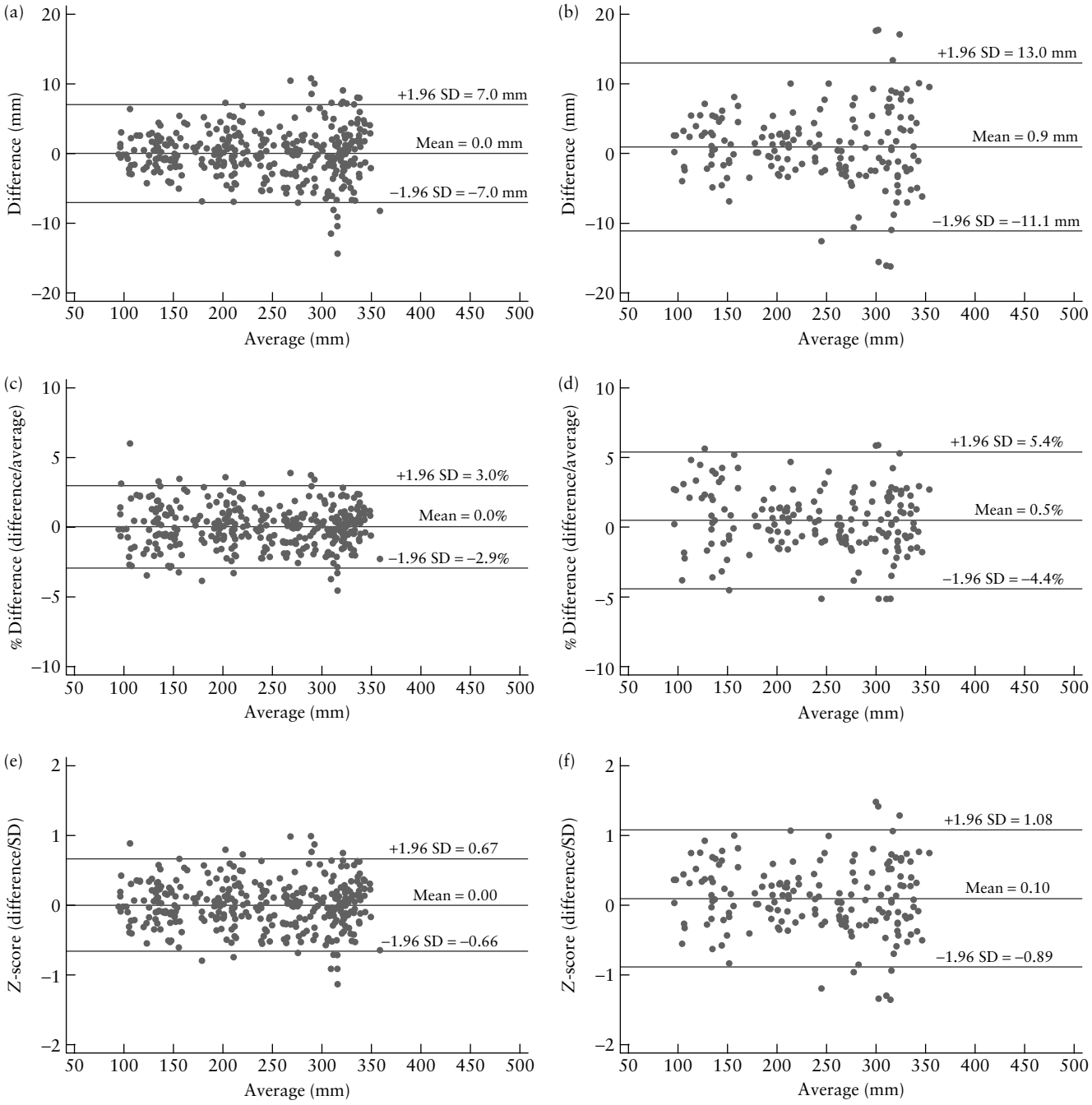

Figure 1 Intraobserver (a,c,e) and interobserver (b,d,f) variability in head circumference measurement (obtained using the ellipse facility), expressed as mm (a,b), percentages (c,d) and Z-scores (e,f).

measured size of all fetal biometric parts (AC more than HC and both more than FL). In contrast, the variability was fairly constant when fetal size or gestational age was corrected for (using percentage of fetal dimensions or *Z*-score differences).

Table 1 summarizes the 95% limits of agreement for all fetal biometric parts and methods of calculation. Agreement was best for HC, with 95% of intra- and interobserver differences being within about ± 3% and ± 5%, respectively, and worst for FL with corresponding values of about ± 7% and ± 11%, respectively. Variability was very similar using the two methods of measuring HC and AC, namely the machine's ellipse facility or calculation of the circumferences from the two diameters (BPD and OFD for HC, and APAD and TAD for AC).

Sources of variability

Concerning caliper placement, as with overall variability, placement variability in measurement units (mm) tended to increase with fetal size for all fetal biometric parts, but was fairly constant when percentages or *Z*-scores were used to correct for fetal size or gestational age as above. The 95% intraobserver limits of agreement for caliper placement in measurement units or percent, respectively, were ± 4.5 mm (2.4%) for HC, ± 9.4 mm (4.1%) for

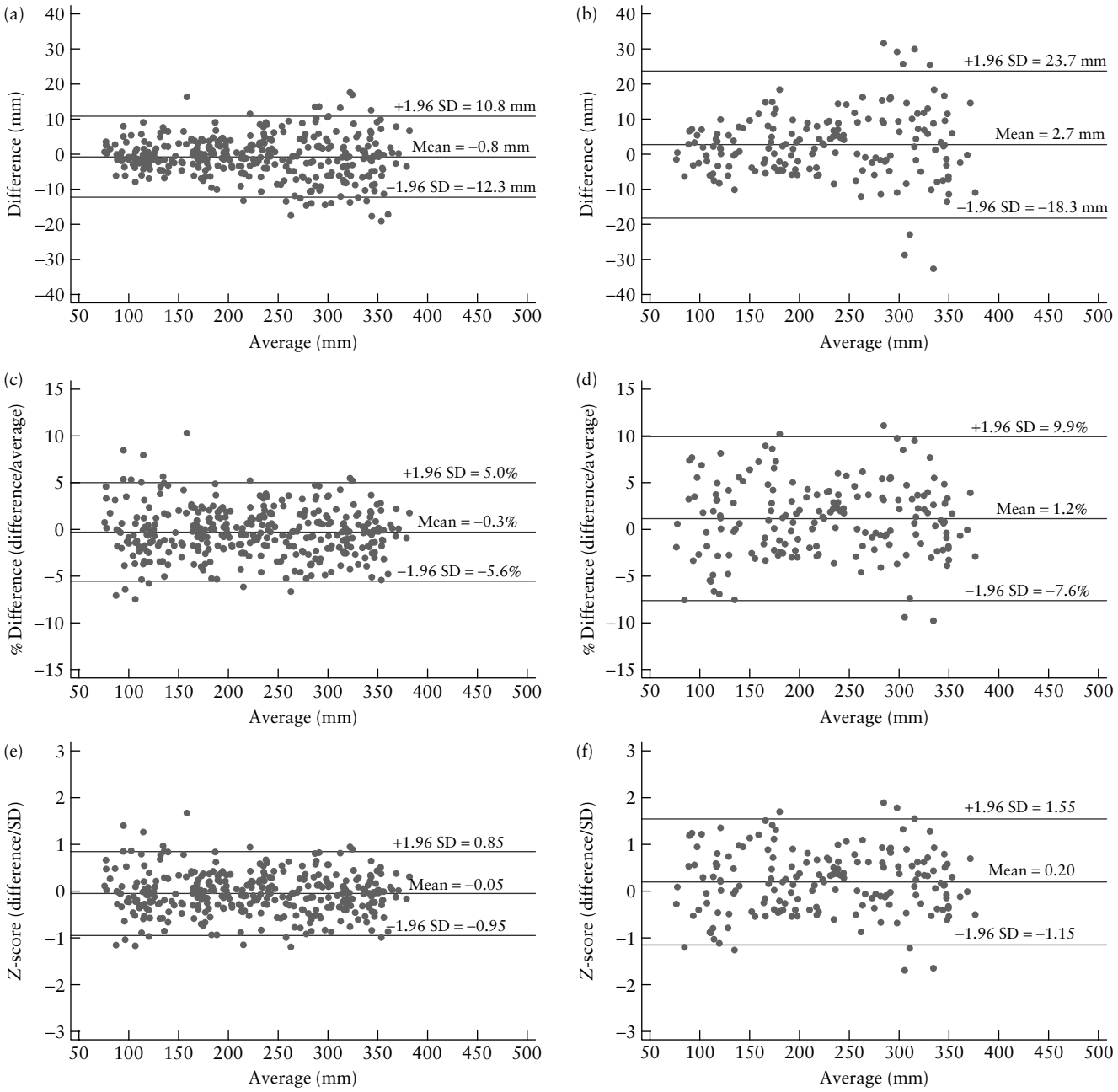

Figure 2 Intraobserver (a,c,e) and interobserver (b,d,f) variability in abdominal circumference measurement (obtained using the ellipse facility), expressed as mm (a,b), percentages (c,d) and Z-scores (e,f).

AC and ± 2.1 mm (4.8%) for FL. The respective values for interobserver variability due to caliper placement for HC, AC and FL were ± 9.8 mm (3.7%), ± 15.5 mm (5.7%) and ± 2.3 mm (5.8%), respectively. When the caliper placement variability values are expressed as a percentage of the values for the overall variability, 52–80% of the observed differences can be accounted for by this step. Caliper placement variability for calculated circumferences was the same as that for the ellipse method.

Concerning other factors, univariate analysis of those that could lead to increased measurement variability showed no statistical difference between active and quiet babies ($P = 0.73$), cephalic and non-cephalic presentation ($P = 0.75$) or maternal BMI of < 24.9 vs. ≥ 24.9 ($P = 0.37$).

DISCUSSION

The usefulness of a screening test depends on its predictive value, which is affected by its reproducibility. Although ultrasound examination has been used as a routine antenatal investigation for over 30 years, the variability in measurements is not well documented when using methods that control the increase in size with advancing gestation. This information is relevant in clinical practice as observers are often different at each evaluation. Previous studies examining reproducibility of fetal biometry measurements are limited in that they are small in number[8,13], included only a narrow range of gestations[13], used ultrasound equipment that is now obsolete[22], did not examine all biometric parts[22,23], used 'non-expert'

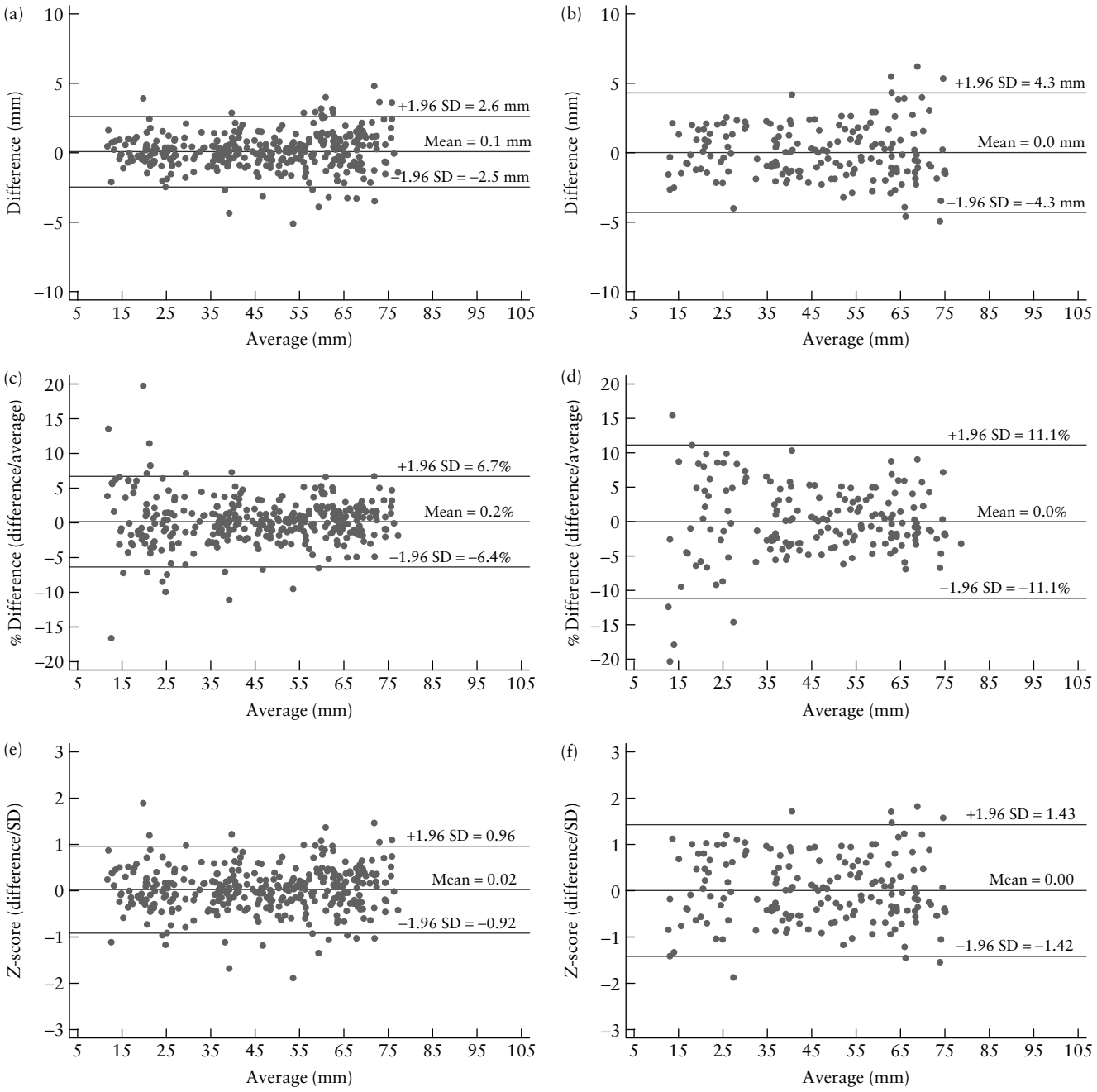

Figure 3 Intraobserver (a,c,e) and interobserver (b,d,f) variability in femur length measurement, expressed as mm (a,b), percentages (c,d) and Z-scores (e,f).

sonographers[21,24], and did not provide definitive numerical values for everyday clinical use since statistical methods yielding data of limited practical interpretive value (such as coefficients of variation and intra-/interclass correlation coefficients)[12,13] were applied. This study addresses these limitations. We recruited women prospectively, using a clearly predefined protocol to include 175 cases evenly distributed from 14–41 weeks' gestation and examined all standard biometric parts including alternate ways of measuring circumferences. We used standardized measurement techniques with modern commercially available ultrasound equipment and equipment specifically programmed to ensure blinding during measurements to remove the possibility of sonographers being biased by seeing previous measurements, which might cause repeat measurements to be artificially similar or biased in line with what would be expected at any given gestation. This is important since evidence from other disciplines indicates that if previous measurements are known, observers frequently record values in line with what they expect them to be[25].

Some studies have suggested an increase in intra- and interobserver variability[11,21] with gestational age while others have not[12]. Visual assessment of the Bland–Altman plots in Figures 1–3 shows that both inter- and intraobserver variability of fetal biometry increases with increasing size and, consequently, gestational age. However, when expressed in relative terms (either as percentage

Table 1 Limits of agreement for intra- and interobserver variability of fetal measurement for all biometric parts, expressed as measurement units, percentages and *Z*-scores. Values in table represent ± 1.96 SD

	Measurement unit (mm)				*Percentage (%)*	*Z-score*
	14–41 weeks	*14–23 weeks*	*24–31 weeks*	*32–41 weeks*	*14–41 weeks*	*14–41 weeks*
Intraobserver variability						
HC (ellipse)	7.0	5.2	7.2	8.8	3.0	0.67
HC (diameter)	7.2	5.1	7.3	10.6	3.1	0.68
AC (ellipse)	11.6	7.9	12.4	15.6	5.3	0.90
AC (diameter)	12.6	8.4	12.8	15.9	5.6	0.92
FL	2.6	2.0	2.5	3.1	6.6	0.94
Interobserver variability						
HC (ellipse)	12.1	6.6	9.0	15.5	4.9	0.99
HC (diameter)	12.0	6.7	9.7	14.5	4.9	0.97
AC (ellipse)	21.0	12.2	13.4	27.8	8.8	1.35
AC (diameter)	20.8	12.4	14.8	26.7	9.1	1.35
FL	4.3	3.2	3.5	4.8	11.1	1.43

AC, abdominal circumference; FL, femur length; HC, head circumference.

of size or as *Z*-scores), variability remains relatively stable throughout pregnancy. As the use of *Z*-scores allows direct comparison of differences across a range of gestational ages, it should be the method of choice.

Although observer variability expressed in measurement units (mm) changes with gestation for all fetal parts, this trend is more pronounced for HC and AC than for FL. A number of reasons may account for this. The femur has edges that are well defined throughout pregnancy and which, with increasing mineralization, become sharper. In contrast, both the head and abdominal margins become more indistinct as gestation increases. For the head, descent into the pelvis and increasing acoustic shadowing at the edges caused by bone mineralization makes measurement more difficult. Measuring the abdomen may be influenced by a relative reduction in amniotic fluid at advanced gestation, and by fetal breathing movements. Finally, these measurements are circumferences, meaning that four points need to be defined (rather than two for the femur, which is a single linear distance). Thus, any errors due to technical limitations are greater.

When variability is expressed in relative terms (i.e. percentages or *Z*-scores), measurement consistency is reversed, with FL being the most variable. This reflects the differences in background biological variability and the absolute total size for each fetal part. Values for HC and AC are consistently larger than those for FL throughout pregnancy, which means that any fixed errors in measurement are relatively more pronounced for the femur.

Many previous studies assessing repeatability have used methods such as coefficient of variation or intra-/interclass correlation coefficient (ICC) determination. We did not consider ICCs to be appropriate since these depend on the range of the measurement values. Consequently, rather than having fixed values, ICCs vary according to the range of gestational ages chosen to study[26]. We therefore did not consider them to be informative and used instead the method proposed by Bland and Altman[17] that has been shown to be more appropriate for assessing repeatability of two measurements[27,28].

Recording of measurements can be viewed as a two-step process of capturing an appropriate image and placing calipers at correct points. Measurement variability can therefore arise from differences in acquisition plane and caliper placement. To ascertain the effect of the latter, we performed an exercise of observer caliper replacement. The relative contribution of caliper placement to overall measurement variability may vary according to background image quality. In this study, all images were acquired to achieve the highest possible quality, and the relative contribution of variability in caliper placement to overall variability is high (52–80% of observed differences). With lower image quality, the relative contribution of caliper placement might be less, as other factors, such as image plane, become more prominent.

Variability caused by slight differences in image plane acquisition can be either random or systematic within and between sonographers. We chose to include three sonographers for all observations in order to ascertain the variability introduced by different operators. A larger number of operators more accurately reflects clinical practice, as most departments have a number of qualified sonographers. In this study, all three were experienced, had strict instructions concerning each measurement and adhered to similar ultrasound practices.

A secondary aim of the study was to identify the effect of factors contributing to variability in fetal biometry measurement. For this purpose fetal presentation, subjectively assessed fetal activity, and maternal BMI were recorded. Our hypothesis was that more active fetuses could make scanning more difficult, increasing the observer-dependent variability compared to that seen with quiescent fetuses, but the results did not support this. In addition, neither fetal presentation nor maternal BMI (within a range of 18.5–29.9) had an effect on variability. Previous studies also found no association between BMI and variability but have documented an association with maternal body fat distribution and, in particular, abdominal wall thickness[11]. It should be noted that in our study the lower limit for BMI was 18.5 and

the upper 29.9; therefore we could not assess the effect of a higher BMI[29].

A limitation of this study is that all newborns were normally grown and amniotic fluid volume had been normal. We cannot ascertain if measurement variability for growth-restricted or macrosomic fetuses would be different, or what the effect of reduced amniotic fluid would be. Another limitation is that the documented variability was obtained under near-optimal conditions, i.e. all sonographers were well trained, utilized the same ultrasound machine and had ample time to complete the examinations. It is possible that under other conditions variability may be different. However, our aim was to report variability under optimal conditions to serve as a standard.

In conclusion, ultrasound variability in fetal biometry performed by trained sonographers under standardized conditions is found to increase with gestational age when expressed in millimeters, but is constant as a percentage of the fetal dimensions or when reported as a *Z*-score. Variability in caliper placement is the major component of overall variability. Measurement variability can be large and should be considered when interpreting fetal growth rates.

ACKNOWLEDGMENTS

We are grateful to the pregnant volunteers who patiently underwent the ultrasound examinations for this study. A.T.P. and C.I. are supported by the Oxford Partnership Comprehensive Biomedical Research Centre with funding from the Department of Health NIHR Biomedical Research Centres funding scheme. 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.

We would like to thank Philips Healthcare for providing the specifically modified HD-9 ultrasound machines and for technical assistance.

REFERENCES

1. Blum L, Kurtz AB. Gestational age: what to measure and when. *Semin Roentgenol* 1990; **25**: 299–308.
2. Bottomley C, Bourne T. Dating and growth in the first trimester. *Best Pract Res Clin Obstet Gynaecol* 2009; **23**: 439–452.
3. Anderson N, Boswell O, Duff G. Prenatal sonography for the detection of fetal anomalies: results of a prospective study and comparison with prior series. *AJR Am J Roentgenol* 1995; **165**: 943–950.
4. Papp Z, Fekete T. The evolving role of ultrasound in obstetrics/gynecology practice. *Int J Gynaecol Obstet* 2003; **82**: 339–346.
5. Dudley NJ, Chapman E. The importance of quality management in fetal measurement. *Ultrasound Obstet Gynecol* 2002; **19**: 190–196.
6. Lawn JE, Cousens S, Zupan J. 4 million neonatal deaths: when? Where? Why? *Lancet* 2005; **365**: 891–900.
7. Mongelli M, Ek S, Tambyrajia R. Screening for fetal growth restriction: a mathematical model of the effect of time interval and ultrasound error. *Obstet Gynecol* 1998; **92**: 908–912.
8. Fescina RH, Ucieda FJ. Reliability of fetal anthropometry by ultrasound. *J Perinat Med* 1980; **8**: 93–99.
9. Hadlock FP, Deter RL, Harrist RB, Park SK. Fetal head circumference: relation to menstrual age. *AJR Am J Roentgenol* 1982; **138**: 649–653.
10. Hadlock FP, Deter RL, Harrist RB, Park SK. Fetal abdominal circumference as a predictor of menstrual age. *AJR Am J Roentgenol* 1982; **139**: 367–370.
11. Harstad TW, Buschang PH, Little BB, Santos-Ramos R, Twickler D, Brown CE. Ultrasound anthropometric reliability. *J Clin Ultrasound* 1994; **22**: 531–534.
12. Perni SC, Chervenak FA, Kalish RB, Magherini-Rothe S, Predanic M, Streltzoff J, Skupski DW. Intraobserver and interobserver reproducibility of fetal biometry. *Ultrasound Obstet Gynecol* 2004; **24**: 654–658.
13. Nakai A, Oya A. Accuracy and reproducibility of ultrasound measurements in obstetric management. *Gynecol Obstet Invest* 2002; **54**: 31–36.
14. Chang TC, Robson SC, Spencer JA, Gallivan S. Ultrasonic fetal weight estimation: analysis of inter- and intra-observer variability. *J Clin Ultrasound* 1993; **21**: 515–519.
15. International Fetal and Newborn Growth Consortium. The International Fetal and Newborn Growth Standards for the 21st Century (INTERGROWTH-21st) Study Protocol. www.intergrowth21.org.uk. 2009.
16. Robinson HP. Sonar measurement of fetal crown-rump length as means of assessing maturity in first trimester of pregnancy. *Br Med J* 1973; **4**: 28–31.
17. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **1**: 307–310.
18. Chitty LS, Altman DG, Henderson A, Campbell S. Charts of fetal size: 2. Head measurements. *Br J Obstet Gynaecol* 1994; **101**: 35–43.
19. Chitty LS, Altman DG, Henderson A, Campbell S. Charts of fetal size: 3. Abdominal measurements. *Br J Obstet Gynaecol* 1994; **101**: 125–131.
20. Chitty LS, Altman DG, Henderson A, Campbell S. Charts of fetal size: 4. Femur length. *Br J Obstet Gynaecol* 1994; **101**: 132–135.
21. Neufeld LM, Wagatsuma Y, Hussain R, Begum M, Frongillo EA. Measurement error for ultrasound fetal biometry performed by paramedics in rural Bangladesh. *Ultrasound Obstet Gynecol* 2009; **34**: 387–394.
22. Deter RL, Harrist RB, Hadlock FP, Carpenter RJ. Fetal head and abdominal circumferences: I. Evaluation of measurement errors. *J Clin Ultrasound* 1982; **10**: 357–363.
23. Tamura RK, Sabbagha RE. Percentile ranks of sonar fetal abdominal circumference measurements. *Am J Obstet Gynecol* 1980; **138**: 475–479.
24. Rijken MJ, Lee SJ, Boel ME, Papageorghiou AT, Visser GH, Dwell SL, Kennedy SH, Singhasivanon P, White NJ, Nosten F, Mcgready R. Obstetric ultrasound scanning by local health workers in a refugee camp on the Thai-Burmese border. *Ultrasound Obstet Gynecol* 2009; **34**: 395–403.
25. Villar J, Repke J, Markush L, Calvert W, Rhoads G. The measuring of blood pressure during pregnancy. *Am J Obstet Gynecol* 1989; **161**: 1019–1024.
26. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput Biol Med* 1990; **20**: 337–340.
27. Sarmandal P, Bailey SM, Grant JM. A comparison of three methods of assessing inter-observer variation applied to ultrasonic fetal measurement in the third trimester. *Br J Obstet Gynaecol* 1989; **96**: 1261–1265.
28. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ* 1992; **304**: 1491–1494.
29. WHO. MONICA Project: Risk Factors. *Int J Epidemiol* 1989; **18**: S46–S55.

