Clinical implications of intra- and inter-observer reproducibility of first trimester measurements performed with transvaginal ultrasound between 6 and 9 weeks gestation.

Anne Pexsters[1], Jan Luts[2], Dominique Van Schoubroeck[1], Cecilia Bottomley[3], Ben Van Calster[2], Sabine Van Huffel[2], Yazan Abdallah[4], Thomas D'Hooghe[1], Christoph Lees[5], Dirk Timmerman[1] and Tom Bourne[1,4]

1. Department of Obstetrics and Gynaecology, University Hospitals, KU Leuven, Belgium

2. Department of Electrical Engineering (ESAT), Katholieke Universiteit Leuven, Belgium

3. Chelsea and Westminster Hospital, Fulham Road, London, UK

4. Queen Charlotte's and Chelsea Hospital, Imperial College London, Du Cane Road, London, UK

5. Division of Fetal-maternal Medicine, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Hills Road, Cambridge CB2 2QQ, UK

Correspondence to:

Dr Anne Pexsters
Department of Obstetrics and Gynaecology
University Hospitals, Katholieke Universiteit Leuven,
Herestraat 49, 3000 Leuven, Belgium
Tel: +32-16-344200
Fax: +32-16-344205
Email: annepexsters@hotmail.com

Keywords: crown-rump length; vaginal ultrasound; early pregnancy
Running title: reproducibility of early pregnancy measurements

Abstract

Objectives

To assess intra- and inter-observer agreement of routinely performed measurements used to assess the likelihood of miscarriage in the first trimester of pregnancy: crown-rump length (CRL) and mean gestation sac diameter (MSD) with transvaginal ultrasonography.

Methods

A cross-sectional study of CRL and gestation sac (GS) measurements in first trimester pregnancies was conducted in a fetal medicine referral center with a predominantly Caucasian population. Gestational age (GA) ranged from 6 to 9 weeks. All patients underwent a transvaginal ultrasound examination with high-resolution Voluson E8 equipment. Two measurements of CRL and three diameters of GS were collected from two observers. Agreement within and between observers for CRL and between observers for MSD was analyzed using 95% prediction intervals, 95% limits of agreement in the framework of Bland-Altman plots, and the intra-class correlation coefficient (ICC).

Results

In total 54 patients were included in the study. Intra- and inter-observer ICC's were high for the CRL measurement, ranging between 0.992 and 0.993 for intra-observer agreement on 54 patients and 0.993 for inter-observer agreement on 44 patients. For the MSD, inter-observer ICC was 0.952. Limits of agreement were ±8.91 and ±11.37 percent of the mean of both measurements for intra-observer CRL and ±14.64 percent for inter-observer CRL agreement. For MSD, the limits of agreement were ±18.78%. An MSD measurement of 20 mm by the first observer leads to a range from 16.8 to 24.5 mm for the second observer. For a CRL measurement of 6 mm by the first observer the range for the second observer is 5.4 mm to 6.7 mm.

Conclusion

For dating purposes, there is reasonable reproducibility for CRL measurements in the first trimester of pregnancy before 9 weeks using transvaginal ultrasonography. When diagnosing miscarriage based on measurements of CRL care must be taken for values around any decision boundary. For MSD, the higher inter-observer variability we have observed has implications for the diagnosis of miscarriage based on measurements of MSD in the absence of a visible embryo or yolk sac.

Introduction (reduced to maximum 350 words)

Threshold measurements for the diagnosis of viability or non-viability by transvaginal sonography have been described and incorporated into management guidelines (RCOG, 2006). Current guidelines (RCOG) define a missed miscarriage or early fetal demise based on ultrasound criteria as the visualisation of an embryo equal or greater than six mm with no visible cardiac activity. Similarly an empty sac is defined as a gestation sac of MSD equal or greater than 20mm without a visible embryo or yolk sac. Given the significance of the decisions being made on the basis of these measurements it is essential that both intra- and inter-observer variability are taken into consideration.

The reproducibility of first trimester CRL measurements was described in an early report with static image scanners for transabdominal sonography (Daya, 1993), after which the Robinson curve for dating was constructed (Robinson 1973), using similar equipment. The reproducibility of fetal biometry in the second and third trimester of pregnancy with transabdominal 3.5 MHz linear and curvilinear probes has been established (Perni, 2004). More recently, a good level of reproducibility for measurements taken between 9 and 14 weeks (CRL and BPD, AC and FL) has been described using transabdominal ultrasound (Verburg, 2008).
However, ultrasound equipment technology has rapidly evolved over recent years, allowing better delineation of landmarks in the early first trimester. Moreover, as new reference CRL dating curves are constructed with predominantly transvaginal ultrasound, the reproducibility of CRL measurements at earlier gestations has become more relevant (Verburg, 2008, Pexsters, 2010).

The primary aim of the study is to evaluate agreement in CRL measurements taken by the same observer (intra-observer agreement) and in measurements of CRL and MSD taken by different observers (inter-observer agreement) between 6 and 9 weeks GA with transvaginal sonography, using modern ultrasound equipment. As a secondary goal, we aimed at investigating its implication for accurately determining GA based on CRL before 9 weeks GA and showing whether the level of reproducibility of these measurements has implications for making a diagnosis of miscarriage based on current cut-off levels.

Methods

We conducted a cross-sectional study on the landmarks of single viable embryos at different gestations between six and nine weeks GA. This was carried out in a referral center for fetal medicine with a predominantly Caucasian population between 2008 and 2009. Patients presenting with an intra-uterine pregnancy between 6 and 9 weeks were included and were asked consent to participate in the study after viability was confirmed. Patients were excluded if viability could not be confirmed.

All women underwent an ultrasound assessment using a transvaginal (6–12 MHz) transducer for B mode imaging, with a Voluson E8 (GE Medical Systems, Zipf, Austria). The last menstrual period or known date of conception after infertility treatment was recorded. During the same visit patients were scanned consecutively by the two observers who were both blind to the measurements of each other. The

measurements of the previous observer were always removed from the machine before the second observer entered the examination room.

The CRL was measured from the outer ends of the embryo to the nearest mm (greatest length) as described previously (Pexsters, 2010). To assess the intra-observer agreement of the CRL measurements, the CRL was measured two times. The probe was moved away from the uterus and the ovaries were examined in between measurements of the pregnancy. Measurements of the GS were taken in three orthogonal planes (two in a sagittal plane and one in transverse) from the inner borders of the sac (Bottomley, 2009), and the mean sac diameter (MSD) was derived.

All ultrasound assessments were carried out by the same two gynecologists with specialist training in obstetric and gynecological sonography. Ethics Committee approval was obtained at the University Hospitals Leuven and the patients gave informed consent.

Statistical analysis

Statistical analysis was performed with SAS 9.2 (SAS Institute Inc., Cary, NC, USA) and Matlab 7.4 (The MathWorks Inc., Natick, Massachusetts, USA).

First, scatterplots were created and the line of equality between two measurements was established to visualize the data and assess potential systematic biases between the measurements. The paired *t*-test with an alpha level of 0.05 was used to check for systemic bias by testing whether the average difference between

measurements was different from zero. If so, this would indicate that one measurement was on average higher than the other.

In case of no systemic bias, 95% prediction intervals for the second observer given the measurement of the first were obtained by randomly permuting both measurements to avoid that one of the observers was systematically chosen as the first observer. This was repeated 1000 times, and each time the 95% prediction intervals for the regression of the difference between both measurements divided by their mean (i.e. the proportional difference, expressed in %) on the first measurement were back transformed to obtain 95% prediction intervals for the first measurement versus the second measurement. These prediction intervals were averaged to obtain the final prediction intervals. To assess the strength of the absolute agreement within and between observers, the intra-class correlation coefficient (ICC) was used based on a two-way random effects analysis of variance model (Shrout, 1979). High agreement is evidenced by a high ICC (close to 1).

Next, the difference between measurements was plotted against their mean in a Bland-Altman plot (Bland, 1986, Bland, 2003). The average difference between measurements was calculated, together with the 95% limits of agreement, which are defined as the average difference (which is assumed to be zero in case of no consistent bias, (Bland, 2003)) ± 1.96 times the standard deviation. The normality assumption and the assumptions of constant mean and variance for the limits of agreement were checked. The assumption of constant variance was violated as the magnitude of the difference increased with the mean, such that the limits of agreement become invalid. Therefore, we used the proportional difference rather

than the difference itself in the Bland-Altman plot. The lack of agreement between measurements or observers becomes relevant only when the limits of agreement are wider than what is clinically acceptable.

Results

Fifty-four patients who underwent a first trimester transvaginal ultrasound scan were included in the study to examine intra-observer agreement in the two CRL measurements of observer 1(A) or/and observer 2(B). Inter-observer agreement of CRL and MSD was examined on 44 patients with measurements from the same two observers (A and B). Maternal age ranged from 25 years to 41 years. The largest proportion of scans was performed in women from 30 to 35 years old. From the total of 54 patients, the majority was Caucasian (52/54). One patient of Afro-carribean origin and one of Asian origin was found in the group of 44 patients on which inter-observer variability was investigated. Measurements of all variables were possible in all women. The MSD measurement by the second observer could not be retrieved in one case. The range of gestational age considered was 40 to 68 days.

Table 1 presents descriptive statistics for CRL and MSD. The first and second CRL measurement from observers A and B are reported as CRL A1, CRL A2 and CRL B1, CRL B2, respectively.

Figures 1 to 4 show the lines of equality and 95% prediction intervals within and between observers. All points lie randomly around the lines of equality, therefore there seems to be no systematic bias. The paired *t*-test confirmed that the observed bias is not significantly different from 0 for all four comparisons. The data points are

close to the line of equality for CRL, which reflects a good agreement of the measurements within and between observers, while inter-observer agreement for MSD is lower.

Based on the 95 % prediction intervals this implies that, given a CRL measurement of 6 mm measured by one observer, the measurement for the second observer can range between 5.4 and 6.7 mm (see Table 2). Likewise, given an MSD measurement of 20mm by one observer, the measurement for the second observer can range between 16.8 mm and 24.5 mm for the other observer (see Table 3).

Figures 5-8 show Bland-Altman plots with the 95% limits of agreement (± 1.96*SD) for intra-observer CRL agreement of both observers, inter-observer CRL agreement, and inter-observer MSD agreement. For the intra-observer CRL agreement the 95 % limits of agreement were ±8.91% and ±11.37% (Figures 5-6), while ±14.64% was obtained for inter-observer CRL (Figure 7). Concerning inter-observer agreement for MSD (Figure 8) the 95% limits of agreement were ±18.78%.

Table 4 summarizes the average proportional differences and the 95% limits of agreement for all possible comparisons. All average proportional differences between measurements were close to 0. This indicates that there is no systemic bias in the measurements within or between observers; all t-tests were nonsignificant as mentioned earlier. The mean difference for intra-observer CRL measurement equaled -1.15% and ±0.41%, while +0.25% and +0.93% were obtained for inter-observer CRL and inter-observer MSD, respectively. Table 4 also provides the ICC's

for all measurements. Intra-observer ICC's of 0.992 and 0.993 and an inter-observer ICC of 0.993 were observed for CRL. The inter-observer ICC for MSD was 0.952.

Discussion

We have shown that the intra- and inter-observer agreement for measurements of CRL in early pregnancy before 9 weeks GA are associated with high ICC's and are better than the inter-observer agreement for MSD measurement. However, the measurements of the mean gestation sac diameter that are specifically used when making a diagnosis of miscarriage were found to be less reproducible between two observers in the same gestational age range.

A strength of this study is that it is the first study to report on intra- and inter-observer agreement of landmarks in the early first trimester of pregnancy between 6 and 9 weeks GA with transvaginal ultrasound. Our study was performed with modern ultrasound equipment, and so the high image resolution allowed the best possible delineation of the embryo and gestational structures available to date.

However, we acknowledge that this study has limitations. The population studied was relatively small, especially for very early gestations where the CRL is below 10mm (7 weeks). It appears that the smallest CRL measurements are closest around the lines of equality, and that the reproducibility at six or seven weeks could possibly be better than somewhat later in the early first trimester (eight weeks). At these very early stages differences in intra- and inter-observer CRL measurements amount to tenths of millimetres, which would be unlikely to result in a difference of more than two days

when dating a pregnancy (Pexsters, 2010). It seems obvious that CRL measurements below 8 weeks are more uniformly taken, since the embryo is straight. From 8 weeks onwards, measurement differences are possible due to the curvature of the embryo. Possible measurement errors then range from + or -2 mm in an embryo of 20 mm and from + or -3 mm in an embryo of 30 mm, which corresponds to a difference in dating of 2 or 3 days, which is important when considering possible consequences of inaccurate dating later in pregnancy (Pexsters, 2010). However, in relation to making a diagnosis of early miscarriage, even a small CRL measurement difference of one millimeter can have an impact on clinical decision making.

The issue of the intra- and inter-observer variability of measurements in the first trimester of pregnancy has not been explored widely in the literature to date. Nevertheless a recent paper showed the influence of ethnicity and maternal age on the rate of change in CRL and MSD (Bottomley, 2009). Another study has shown that the requirement to re-date an early pregnancy may be predictive of later growth restriction (Thorsell, 2008). It seems likely that the accuracy of early CRL measurements will become of increasing importance. In order to improve the accuracy of dating from 5.5 weeks GA onwards, we constructed a new reference CRL size curve based on a large number of pregnancies, which showed meaningful differences of several millimeters with the classically used Robinson and Hadlock curves, especially at very early gestations (Pexsters, 2010). This study highlighted that a study investigating intra- and inter-observer variability of CRL measurements using transvaginal ultrasound is necessary.

A key issue regarding the reproducibility of embryo and gestation sac size measurements relates to the definitions used for miscarriage. A miscarriage is defined as an embryo of CRL greater or equal to 6 mm with no FH or an MSD of greater or equal to 20 mm where no yolk sac or embryo can be seen (RCOG, 2006). Accordingly significant inter-observer variability may be associated with a misdiagnosis of miscarriage and lead to inadvertent termination of a pregnancy. If we consider the 95 % prediction intervals from our data, a measured MSD value of 20 mm by the first observer corresponds to a range of 16.8 mm to 24.5 mm for the second observer. This is clinically important. Similarly, for a measured CRL of 6 mm, the second observer's measure ranges from 5.4 to 6.7 mm

Our data show that the variability of CRL measurements both for a single observer and between observers is smaller than the inter-observer variability for MSD measurements. However although small, these differences may have significant clinical consequences. It is clear that whatever single cutoff values may be used to define a miscarriage, great care must be taken when measurements approach the decision boundary. In general little harm is associated with repeating a scan at an interval when deciding about the potential viability of a pregnancy. In the future we suggest that for any proposed cutoff value for CRL or MSD to define miscarriage, possible variations in measurement accuracy are taken into account before diagnosing miscarriage on the basis of one scan. Hence in the UK, an MSD of 20 mm to define miscarriage would become 24.5 mm to take into account possible measurement error. In this way the risk of terminating wanted viable embryos should be minimized.

Acknowledgements

Jan Luts is a postdoctoral researcher from the Research Foundation – Flanders (FWO). Ben Van Calster is a postdoctoral researcher from the Research Foundation – Flanders (FWO). Research supported by Research Council KUL: GOA Ambiorics, GOA MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC); Research Foundation – Flanders: FWO G.0302.07 (SVM), G.0341.07 (Data fusion); IWT: TBM070706-IOTA3; Belgian Federal Science Policy Office IUAP P6/04 (DYSCO, 'Dynamical systems, control and optimization', 2007-2011).

Deleted: supported by a grant from the Agentschap voor Innovatie door Wetenschap en Technologie (IWT)
Deleted:
Deleted:
Deleted:
Deleted:
Deleted:

References

1. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1: 307-310.

2. Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. Ultrasound Obstet Gynecol 2003; 22: 85-93

3. Bottomley C and Bourne T. Dating and growth in the first trimester. Best Practice and research Clin Obstet Gynaecol 2009; xxx: 1-14 (in press)

4. Bottomley C, Daemen A, Mukri F, Papageorghiou AT, Kirk E, Pexsters A, De Moor B, Timmerman D, Bourne T. Assessing first trimester growth: the influence of ethnic background and maternal age. Hum Rep 2009; 24.2: 284-290

5. Daya S. Accuracy of gestational age estimation by means of fetal crown-rump length measurements. Am J Obstet Gynecol 1993; 168: 903-908.

6. Hadlock FP, Shah YP, Kanon OJ, Lindsey JV. Fetal crown-rump length: reevaluation of relation to menstrual age (5-18 weeks) with high-resolution real-time ultrasound. Radiology 1992; 182: 501-5

7. Perni SC, Chervenak FA, Kalish RB, et al. Intraobserver and interobserver reproducibility of fetal biometry. Ultrasound Obstet Gynecol 2004; 24: 654-658

8. Pexsters A, Daemen A, Bottomley C, Van Schoubroeck D, De Catte L, De Moor B, D'Hooghe T, Lees C, Timmerman D and Bourne T. A new crown-rump length CRL size curve based on over 3500 pregancies. Accepted for publication in Ultrasound Obstet Gynecol.

9. 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

10. Robinson HP, Fleming JEE. A critical evaluation of sonar "crown-rump length" measurements. Br J Obstet Gynaecol 1975; 82: 702-710

11. Royal College of Obstetricians and Gynaecologists. The management of early pregnancy loss. RCOG 2006; Green-top guideline number 25

12. Shrout PE, Fleiss JL (1979). Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 86(2): 420-428.

13. Thorsell M, Kaijser M, Almström H, Andolf E. Expected day of delivery from ultrasound dating versus last menstrual period-obstetric outcome when dates mismatch. BJOG. 2008 Apr;115(5):585-9.

14. Verburg BO, Mulder PG, Hofman A, Jaddoe VW, Witteman JC, Steegers EA. Intra- and interobserver reproducibility study of early fetal growth parameters. Prenat Diagn 2008 Apr; 28(4): 323-331

Figures and tables

Table 1: descriptive statistics for CRL and MSD

	N	Mean (mm)	Min (mm)	Max (mm)	SD (mm)
CRL A1	54	15.2	1.3	30.6	8.5
CRL A2	54	15.4	1.3	31.2	8.7
CRL B1	44	16.6	3.4	32.7	8.5
CRL B2	44	16.4	3.2	30.6	8.3
MSD A	43	29.6	9.6	47	9.3
MSD B	43	29.1	10.5	46.9	8.6

CRL= crown-rump length
MSD = mean sac diameter
A = observer 1
B = observer 2
1, 2 = first, second measurement

Table 2: CRL measurement range by the second observer for a CRL measured by the first observer

CRL1 of first observer (mm)	95% PI for CRL1 of second observer (mm)
5	[4.5-5.6]
6	[5.4-6.7]
7	[6.3-7.9]
10	[8.9-11.2]
20	[17.9-22.4]
30	[26.7-33.5]

CRL1= crown-rump length, first measurement
PI = prediction interval

Table 3: MSD measurement range by the second observer for an MSD measured by the first observer

MSD of first observer (mm)	95% PI for MSD of second observer (mm)
10	[8.5-12.6]
15	[12.7-18.6]
16	[13.5-19.8]
17	[14.3-21.0]
18	[15.1-22.2]
19	[16.0-23.4]
20	[16.8-24.5]
21	[17.6-25.7]
22	[18.4-26.9]
23	[19.2-28.0]
24	[20.0-29.2]
25	[20.9-30.4]
30	[24.9-36.1]
40	[32.5-47.6]

MSD = mean sac diameter

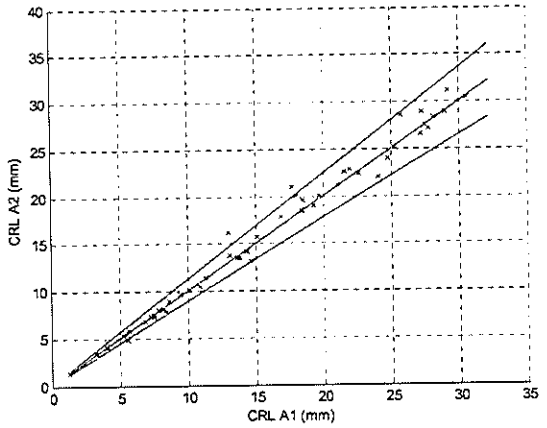
PI = prediction interval

Table 4: proportional differences and 95% limits of agreement

	Mean proportional difference (%)	95% limits of agreement (%)	ICC
CRL A1-A2	-1.15	± 11.37	0.993
CRL B1-B2	0.41	± 8.91	0.992
CRL A-B	0.25	± 14.64	0.993
MSD A-B	0.93	± 18.78	0.952

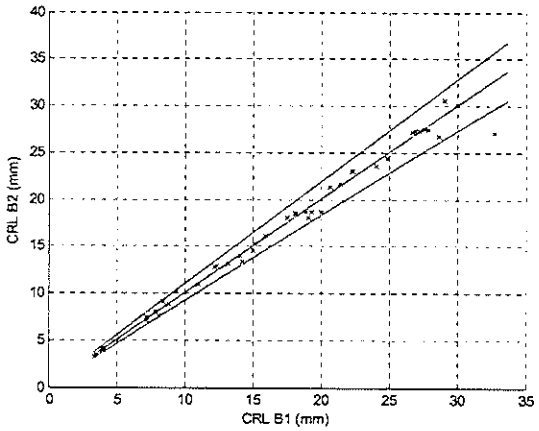
CRL= crown-rump length
MSD = mean sac diameter
A = first observer
B = second observer
1, 2 = first, second measurement
ICC = intraclass correlation coefficient

Figure 1: intra-observer observer A: CRL1 measurement versus CRL2 measurement, line of equality and 95% prediction intervals



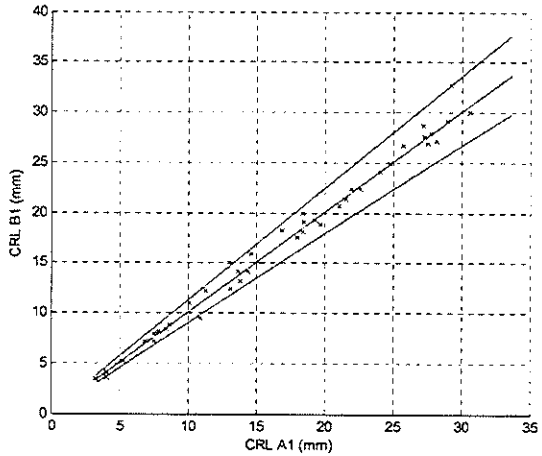
CRL = crown-rump length
CRL A1 = first CRL measurement observer A
CRL A2 = second CRL measurement observer A

Figure 2: intra-observer observer B: CRL1 measurement versus CRL2 measurement, line of equality and 95% prediction intervals



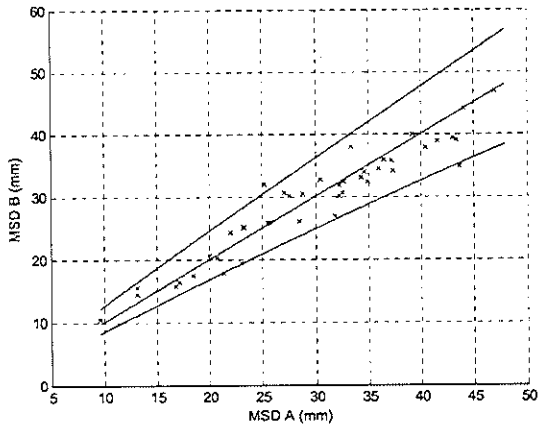
CRL= crown-rump length
CRL B1 = first CRL measurement observer B
CRL B2 = second CRL measurement observer B

Figure 3: inter-observer CRL1 measurement observer A versus observer B, line of equality and 95% prediction intervals



CRL= crown-rump length
CRL A1 = first CRL measurement observer A
CRL B1 = first CRL measurement observer B

Figure 4: inter-observer MSD measurement observer A versus observer B, line of equality and 95% prediction intervals

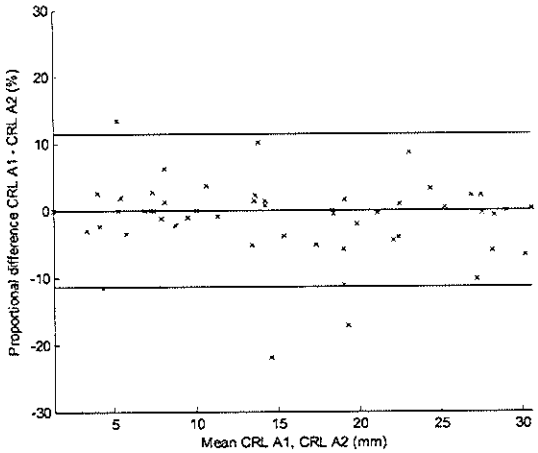


MSD = mean sac diameter

MSD A = MSD observer A

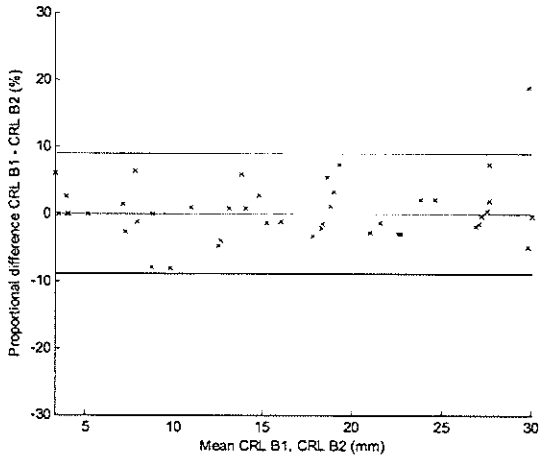
MSD B = MSD observer B

Figure 5: Bland-Altman plot for intra-observer agreement of CRL for observer A



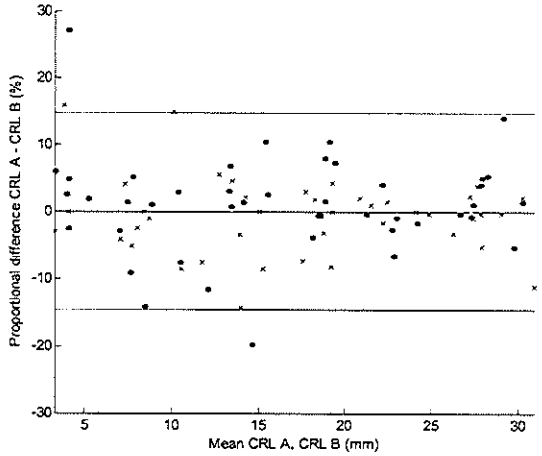
CRL= crown-rump length
CRL A1 = first CRL measurement observer A
CRL A2 = second CRL measurement observer A

Figure 6: Bland-Altman plot for intra-observer agreement of CRL for observer B



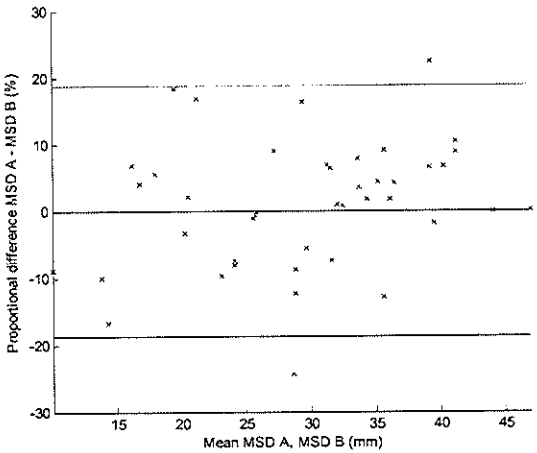
CRL= crown-rump length
CRL B1 = first CRL measurement observer B
CRL B2 = second CRL measurement observer B

Figure 7: Bland-Altman plot inter-observer agreement for the CRL measurement of observer A versus observer B



CRL= crown-rump length
CRL A = CRL observer A
CRL B = CRL observer B
Cross = first measurement
Point = second measurement

Figure 8: Bland-Altman plot inter-observer agreement for the MSD measurement of observer A versus observer B



MSD = mean sac diameter
MSD A = MSD observer A
MSD B = MSD observer B