Short communications

Inter-rater variability of ultrasound scan measurements: balanced incomplete block design

Chrishantha Abeysena[1] , Pushpa Jayawardana[1]

Abstract

Introduction

To assess inter-rater variability of ultrasound scan measurements for determining period of gestation by three raters applying balanced incomplete block design.

Methodology

Twelve pregnant women who attended the field antenatal clinics were subjected to scan measurements, in terms of bi-parietal diameter (BPD), femur length (FL), abdominal (AC) and head (HC) circumferences of the fetus between 15–24 weeks of gestation. Each pregnant woman was scanned by two of the three raters who were blind to the measurements made by the other using the same machine. Balanced incomplete block design was generated and data were analyzed using ANOVA.

Results

There were no statistically significant variation among raters in measuring BPD ($F = 0.68$; $p = 0.53$), AC ($F = 1.99$; $p = 0.19$) and HC ($F = 0.06$; $p = 0.94$). There was statistically significant variation among raters for measuring FL ($F = 7.4$; $p = 0.01$).

Conclusion

Statistically significant inter-rater differences were observed only for measurements of FL. However, despite the inter-rater differences of mean abdominal and head circumferences being not significant statistically, their variance can have a clinical significance.

Key words

Bi-parietal diameter, femur length, reliability, ultrasound, variability

Introduction

One of the uses of ultrasound scan measurements are to estimate the period of gestation and based on that the expected date of delivery. The usual measurements made with regard to the above are bi-parietal diameter, femur length, abdominal circumference and head circumference. These measurements are then converted into period of gestation by applying the suitable regression model for each measurement. According to the literature more accurate measurements are possible when the ultrasound measurements are done between 15th to 24th weeks of gestation.[1] However; variations in measurements that occur when carried out by several raters may affect the management of pregnancy and its complications adversely.

1. Senior Lecturer, Department of Public Health, Faculty of Medicine, University of Kelaniya, Ragama, Sri Lanka

There are several methods of assessing inter-rater reliability. Latin square design was applied to assess observer variability in anthropometry by 16 field workers using eight children.[5] Another study applied a nested Latin square design to determine the inter/intra-rater reliability of three physiotherapists who independently rated pain by visual analogue scale in 33 subjects on three days in a randomized order.[3]

Balanced incomplete block design was used to assess inter-rater reliability of Vancouver Sedative Recovery Scale by 16 raters using 16 children.[4] This design has an efficiency index of 0.89 relative to a completely crossed design (in which each of 16 raters would rate each of 16 children).[4] Balanced incomplete block design is indicated for comparing the raters' mean levels of rating and whether each mean is estimated with the same precision.[2] The advantage of this method is not having the need to rate all the subjects by every rater.[2] The objective of this study was to assess inter-rater variability of ultrasound scan measurements for determining period of gestation by three raters applying balanced incomplete block design.

Methods

Twelve pregnant women who attended the field antenatal clinics were invited to participate. Each pregnant woman was asked to come for the ultrasound scan measurements twice on two consecutive days during the 15th to 24th weeks of gestation to the Colombo North Teaching Hospital, Ragama. Each pregnant woman was scanned by two of the three raters who were consultant obstetricians. Bi-parietal diameter, femur length, abdominal circumference and head circumference were measured. Second rater was blind to the measurement made by the first rater. All measurements were done using the same ultrasound scan machine.

Balanced incomplete block design was generated[6] (Table 1) with the following features. The three raters (I, II, III) were paired as I and II, II and III and I and III. Each block (participant) was rated by only one pair and the same pair together rated four blocks (For example raters I and II rated together four Blocks namely1, 4 7 and 10). Thus the three pairs covered all the 12 blocks with no overlap between pairs.

Each rater assessed eight blocks which appeared eight times in the design. Statistical analysis was conducted by applying ANOVA to the General Linear Model using Minitab 14.

Subjects	Rater I	Rater II	Rater III

1	A	B	
2		B	C
3	A		C
4	A	B	
5		B	C
6	A		C
7	A	B	
8		B	C
9	A		C
10	A	B	
11		B	C
12	A		C

Table 1- Balanced Incomplete Block Design

Results

Mean age of the participants was 27 (SD±6.7) years, ranging from 19 to 37 years. All were observed to have normal amount of liquor. Mean bi-parietal diameter measurements of three raters were 43 (SD±7.8), 42.3 (SD±6) and 43.5 (SD±9.2) mm respectively (Table 2). Mean abdominal circumference (AC) measurements of the three raters were 122 (SD±29), 144 (SD±35) and 143.5 (SD±33) mm respectively (Table 3). Mean head circumference measurements of three raters were 150.6 (SD±26.6), 167 (SD±33) and 175.6 (SD±28) mm respectively (Table 4). A statistically significant variation was not observed among raters with regard to any of the above three measurements: Bi- parietal diameter ($F = 0.68$; $p = 0.53$); abdominal circumference ($F = 1.99$; $p = 0.19$) and head circumference measurements ($F = 0.06$; $p = 0.94$) by the three raters.

Mean femur length measurements of the three raters were 25 (SD±9.2), 32.5 (SD±6) and 31 (SD±7.2) mm respectively. Table 5 showed that there was a statistically significant variation observed among raters for measuring femur length ($F = 7.4$; $p = 0.01$).

Table 2 - Inter-Rater Variation of Bi-Parietal Diameter

	Sum of Square	Degree of Freedom	Mean Square	F value	P-value
Subjects	1232.8	11	112.07	24.02	0.00
Raters	6.33	2	3.17	0.68	0.53
Error	46.67	10	4.67		

Table 3 - Inter-Rater Variation of Abdominal Circumference

	Sum of Square	Degree of Freedom	Mean Square	F value	P-value
Subjects	20759	11	1187.2	12.01	0.00
Raters	625.6	2	313.3	1.99	0.19
Error	1570.9	10	157.1		

Table 4 - Inter-Rater Variation of Head Circumference

	Sum of Square	Degree of Freedom	Mean Square	F value	P-value
Subjects	17934	11	1630.4	42.32	0.00
Raters	4.7	2	2.4	0.06	0.94
Error	385.2	10	38.5		

Table 5 - Inter-Rater Variation of Femur Length

	Sum of Square	Degree of Freedom	Mean Square	F value	P-value
Subjects	1155.9	11	105.09	17.25	0.00
Raters	90.58	2	45.29	7.44	0.01
Error	60.92	10	6.09		

Discussion

The study showed that bi-parietal diameter, abdominal and head

circumference were more reliable measures of predicting period of gestation than FL. The difference between the lowest and the highest mean bi-parietal diameter of two raters was 1.2mm which is also not clinically significant when converting to period of gestation. Even though there were no statistically significant differences of mean abdominal and head circumference measurements between three raters, the differences between the lowest and the highest mean abdominal and head circumference were 22 mm (144 - 122 mm) and 25 mm (175.6 -150.6 mm) respectively. These differences reflect a difference of two weeks in terms of the period of gestation in respect of each measurement, which may have a greater clinical significance.

Further our study found that there was a statistically significant variation between three raters for measuring femur length. The difference between the lowest and the highest rater of the femur length measurement was 7.5 mm which is approximately two weeks difference by period of gestation.[1] One study found that correlation coefficient of gestational age versus fetal femur length is statistically greater than that of the gestational age versus fetal bi-parietal diameter.[7] This study suggested that the measurement of the fetal femur length was a more precise index of gestational age than the bi-parietal diameter.[7,8] Another study reported that even for mothers between 19 and 32 completed weeks gestation there were no statistically significant differences in femur length vs. gestational age between the various racial categories.[9]

The incomplete block design enabled three raters to assess 12 pregnant mothers and had two major advantages. It avoided the presence of a large group of participants for a reliability study which saves cost. Therefore it minimized the ethical problems and the inconvenience caused by using a larger number of participants who need to be scanned three times within the same week. For a Latin square design at least five mothers should be scanned by five raters. Therefore each participant has to come for the scan five times which would have been inconvenient for both participants and raters.

Conclusion

Statistically significant inter-rater differences were observed only for measurements of FL. However, despite the inter-rater differences of mean abdominal and head circumferences being not significant statistically, their variance can have a clinical significance.

References

1. Chudleigh P, Pearce JM. Obstetric Ultrasound. Second Edition, Churchill Livingstone. 1992; 77-94.

2. Fleiss JL. Balanced incomplete block designs for interrater reliability studies. Applied Psychological Assessment. 1981; 5: 105-12.

3. Pomeroy VM, Frames C, Faragher EB, Hesketh A, Hill E, Watson P, Main CJ. Reliability of a measure of post-

stroke shoulder pain in patients with and without aphasia and/or unilateral spatial neglect. Clin Rehabil. 2000 Dec; 14(6):584-91.

4. Macnab AJ, Levine M, Glick N, Phillips N, Susak L, Elliott M. The Vancouver sedative recovery scale for children: validation and reliability of scoring based on videotaped instruction. CAN J ANAESTH 1994: 41 (10), pp913-8

5. Balasuriya S. Observer variability in anthropometry. Ceylon J. Med. Sci. 1988;31:25-34.

6. Cochran WG, Cox GM. Experimental Designs, 2nd ed, New York: John Wiley & Sons, 1957.

7. Yam MN, Bracero L, Reilly KB, Murtha L, Aboulafia M, Barron BA. Ultrasonic measurement of the femur length as an index of fetal gestational age. Am J Obstet Gynecol. 1982; 144(5):519-22.

8. E. Shalev E, Feldman E, Weiner E, Zuckerman H. Assessment of gestational age by ultrasonic measurement of the femur length. Acta Obstetricia et Gynecologica Scandinavica, 1985;64:71 – 74.

9. Ruvolo KA, Filly RA, Callen PW. Evaluation of fetal femur length for prediction of gestational age in a racially mixed obstetric population. Journal of Ultrasound in Medicine, 1987;l 6(8) 417-419

Research Letters