Inter-observer variability in fetal biometric measurements

Rami Kilani a, Wesam Aleyadeh a,*, Luay Abu Atieleh b, Abdul Mane Al Suleimat b, Maysa Khadr c, Hassan M. Hawamdeh a

a Hashemite University, Zarqa, Jordan
b Al Bashir Hospital, Amman, Jordan
c University of Jordan, Amman, Jordan

A B S T R A C T

Objective: To evaluate inter-observer variability and reproducibility of ultrasound measurements for fetal biometric parameters.

Materials and methods: A prospective cohort study was implemented in two tertiary care hospitals in Amman, Jordan; Prince Hamza Hospital and Al Bashir Hospital. 192 women with a singleton pregnancy at a gestational age of 18–36 weeks were the participants in the study. Transabdominal scans for fetal biometric parameter measurement were performed on study participants from the period of November 2014 to March 2015. Women who agreed to participate in the study were administered two ultrasound scans for head circumference, abdominal circumference and femur length. The correlation coefficient was calculated. Bland–Altman plots were used to analyze the degree of measurement agreement between observers. Limits of agreement ± 2 SD for the differences in fetal biometry measurements in proportions of the mean of the measurements were derived. Main outcome measures examine the reproducibility of fetal biometric measurements by different observers.

Results: High inter-observer inter-class correlation coefficient (ICC) was found for femur length (0.990) and abdominal circumference (0.996) where Bland–Altman plots showed high degrees of agreement. The highest degrees of agreement were noted in the measurement of abdominal circumference followed by head circumference. The lowest degree of agreement was found for femur length measurement. We used a paired-sample t-test and found that the mean difference between duplicate measurements was not significant (P > 0.05).

Conclusion: Biometric fetal parameter measurements may be reproducible by different operators in the clinical setting with similar results. Fetal head circumference, abdominal circumference and femur length were highly reproducible. Large organized studies are needed to ensure accurate fetal measurements due to the important clinical implications of inaccurate measurements.

© 2018 Taiwan Association of Obstetrics & Gynecology. Publishing services by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

Ultrasound is currently the cornerstone of obstetric imaging. Patients routinely get ultrasound scans during pregnancy. The most common ultrasound scan is usually performed between 18 and 22 weeks of gestation to exclude fetal anomalies [1]. Further ultrasound scans are common in the third trimester, with clinical indications for small-for-date fetuses or reduced fetal movement. The measurements taken during the ultrasound scans, especially during the third trimester, are essential for the management of pregnancy and optimal delivery. Sixty percent of neonatal deaths are associated with low birth weight; the identification of growth-restricted fetuses is thereby important. Inaccurate measurement can lead to unnecessary intervention or missing cases with growth restriction. Incorrect diagnoses of normal fetuses may also lead to perinatal compromise [2].

Few studies have examined inter-observer variation despite the importance of fetal measurement in the clinical decision-making process [3]. There are also no recent studies and none that examine inter-observer variation in developing countries [3–5]. Recent, relevant and extensive studies that examine inter-observer variation and reproducibility of fetal measurements are necessary due to the dynamic and continually changing nature of medical equipment which may increase the risk of diagnostic errors [6].
There is currently a lack of qualified sonographers and a shortage of ultrasound machines in Jordan. This is most likely due to high cost and maintenance of equipment that may make it difficult to administer routine or frequent patient examinations. Most ultrasound scans are performed by obstetricians who identify specific clinical indications for the examination; important clinical decision are made based on the results of a single ultrasonography examination. This is especially significant for evaluating the accuracy and reproducibility of fetal measurements in developing countries like Jordan.

The aim of the present study is to determine the inter-observer reproducibility of ultrasound measurements of fetal biometric parameters in a major hospital in Jordan. We will discuss the methods and the process of patient examination in Materials and methods. Results will highlight the results, while Discussion will discuss current evidence and explore the significance of our results in real life applications and clinical settings. Conclusion will highlight our conclusions.

**Materials and methods**

**Examination process**

Al Bashir Hospital in Amman-Jordan is the main governmental hospital in Amman providing services for over 1,000,000 people and serves as a referral hospital for all other Ministry of Health hospitals in Jordan. Ethical approval was obtained from the Hashemite University Ethical Committee and from Al Bashir Hospital ethical committee.

Between November 2014 and March 2015, one hundred and ninety-two pregnant women were recruited for the study. Participants were women with singleton pregnancies scheduled for delivery at Al Bashir Hospital. Inclusion criteria include:

- Maternal age between 18 and 35 years
- Singleton pregnancy
- A known last menstrual period (LMP)
- Regular cycles and gestational age between 18 weeks and 36 weeks

Exclusion criteria include:

- Pregnancy complications, such as preeclampsia and vaginal bleeding
- Significant past medical history, i.e. maternal or familial hypertension, diabetes, heart disease, history of smoking and a previous pregnancy complicated with intra uterine growth restriction (IUGR)

Women who met the inclusion criteria and agreed to participate in the study were administered two abdominal ultrasound scans performed by two different operators within 10 min of each other. The first scan was performed by “Operator 1” and the second scan was performed by “Operator 2”. Each examiner was given 15 min to complete the exam. Only one operator was present in the exam room during patient examination. All ultrasound tests were performed using the same machine, Samsung Medison R5 (South Korea), with a curvilinear abdominal transducer. The machine was programmed to not show obtained measurements on the screen during examination. The two obstetricians perfuming the examination were certificated obstetricians and gynaecologists with significant experience in performing ultrasound scans.

A strict protocol was followed:

Each examiner started with fetal head measurements after obtaining an image of the head. They then measured head circumference using the ellipse facility followed by an image and measurement of the abdomen using an ellipse facility. They then obtain an image of the femur and record femur length. Each operator completed a set of four stored images for each examination that consisted of a head image, an abdominal image, an image of the femur and a report image showing all three measurements.

Head measurements were taken in the trans-thalamic plane measuring the outer border of the skull at the time of scan using an ellipse region-of-interest (ROI) built into the ultrasound unit. 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 measuring the outer border of the abdomen using an ellipse as well. The femur closest to the probe was measured for femur length with its long axis placed 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.

Estimated fetal body weight was calculated using the most common equation in literature for the EFW, Hadlock formula [11,12].

\[
\text{Hadlock formula: } \log 10 (\text{weight}) = 1.304 + 0.05281^\text{AC} + 0.1938^\text{FL} - 0.004^\text{AC}^2 \text{FL}.
\]

All measurements of estimated fetal weight are supplied in the supporting material section in the appendix. All measurements and image were retrieved at the end of data collection.

**Statistical analysis**

Data used for the descriptive statistics were obtained from clinical. The comparisons for each fetal biometric part were assessed using the six measurements taken in each fetus (three measurements by Operator 1 and three measurements by Operator 2). Variability was assessed by calculating the differences between the two measurements made by the two operators on the same fetus (n = 180). The resultant standard deviation (SD) values of the differences of the measurements were then corrected to obtain the equivalent value for single measurements by using the Bland–Altman formula.

**Results**

A total of 192 women met the inclusion criteria and agreed to participate in the study. It was not possible to obtain acceptable images of the ovaries in 12 participants, i.e. high body mass index, fetal activity and fetal head position. Acceptable images and measurements were obtained and included in the study in 180 participants. Mean maternal age was 24.4 years with a range of 18.0–35.0 years. Mean maternal BMI was 23.2 years with a range of 18–34. Mean gestational age was 27.45 with a range of 18–35 + 6 weeks. Head circumference, abdominal circumference, and femur length measurements can be seen in Table 1.

It can be seen in Figs. 1–3 that fetal biometric measurements between observers has a high degree of agreement and there is random distribution around the equality line indicating lack of bias.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Descriptive statistics for fetal biometric measurements.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fetal Biometry</td>
<td>Number</td>
</tr>
<tr>
<td>HC A</td>
<td>180</td>
</tr>
<tr>
<td>HC B</td>
<td>180</td>
</tr>
<tr>
<td>AC A</td>
<td>180</td>
</tr>
<tr>
<td>AC B</td>
<td>180</td>
</tr>
<tr>
<td>FL A</td>
<td>180</td>
</tr>
<tr>
<td>FL B</td>
<td>180</td>
</tr>
</tbody>
</table>
It can also be seen in Table 2 that there is high inter-observer correlation after measuring the value of ICC. This may be explained by the wide range of fetal measurements made and the small differences between operators.

The Bland–Altman graph is a scatterplot of variable means plotted on the horizontal axis and the differences plotted on the vertical axis which shows the amount of disagreement between the two measurements. This plot includes approximate 95% limits. If differences observed in this plot are not deemed clinically important, this is a confirmation of agreement. Bland–Altman graphs for the differences in follicular number count measurements between observers against the mean with the limits of agreement (±2 Standard Deviation) are plotted in Figs. 4–6. We notice that the 95% limits for the differences between examiners are not clinically significant and there is good agreement in all three measurements. The highest agreement was noted in the measurement of the abdominal circumference followed by the head circumference. The lowest agreement was for the femur length measurement. We used the paired-sample t-test to find that the average difference between duplicate measures was not significant (p < 0.05); there is no systematic difference between the pairs of results, indicating no bias among or between observers.

**Fig. 1.** Femur length measurements by examiner A and B.

**Fig. 2.** Abdominal circumference measurements by examiner A and B.
Discussion

Accurate fetal ultrasound measurements are important for making correct diagnoses and clinical decisions in obstetrics health care. The fetal biometric measurements are used mainly for estimating gestational age and to diagnose intra-uterine growth restrictions. The use of ultrasound in obstetrics is highly operator-dependent. There is a significant risk of diagnostic errors if the ultrasound exam is performed by an unqualified ultrasound operator [6]. Ensuring appropriate training in ultrasound technique and interpretation is an important part in the curriculum of obstetrics and gynecology residency programs.

The aim of this study was to determine the degree of reproducibility of ultrasound fetal biometry measurements at a major hospital in a developing country. We have demonstrated that there is high reproducibility of measurements of head circumference,
abdominal circumference and femur length in singleton gestations using trans abdominal ultrasound. All biometric parameters are highly reproducible by different operators.

Previous studies examining the variability and reproducibility of obstetric ultrasound measurement have not been sufficient. They evaluated fetal biometric measurements and indicated similar findings to ours with good agreement between observers. The first study looking into fetal measurement accuracy dates back to 1977 and compared the biparietal diameter using bistable ultrasonography on 50 study participants. The researchers were able to show good reproducibility of measurements [7]. Another study looking into the reproducibility of ultrasound fetal biometric measurement looked at the reproducibility of biparietal diameter and femur length on 30 participants and also showed good reproducibility of results [4]. However, the aforementioned studies examined a small patient sample population and did not include all fetal measurements. The researchers in Ref. [3] conducted a study on a larger sample of 122 participants and looked at reproducibility of all three important measurements: head, abdomen and femur. This study may not be adequate for reference due to the use of inter/intra class coefficient to assess reproducibility. Rijken et al. [8] aimed to assess reproducibility of fetal measurements in the hands of locally trained
Biparietal diameter (BPD) is obtained using a cross-sectional view of the fetal head at the level of the thalami. Both calipers has been introduced and described in Ref.[9]. It must be noted that has been described. i.e. outer edge to outer edge or outer edge to inner edge. The technique used in the measurement of our study has been introduced and described in Ref. [9]. It must be noted that the BPD is less reliable than head circumference (HC) in determining gestational age especially when there is variation in the skull shape, such as dolichocephaly or brachycephaly. It was also found that HC correlates better to gestational age than BPD [10].

The objective of this study was to evaluate inter-observer variability and reproducibility of measurements. The authors of this study felt that obtaining a cross sectional view of the fetal head at the correct level and measuring the more reliable fetal biometric measurement is sufficient.

This study was necessary because it expands on the sample population and included all parameters. Examinations were also done in a busy tertiary hospital. This is the only study showing reproducibility in a real clinical sitting in a developing country. The high number of participants when compared to previous publications gives us a more reliable measurement of bias, and the short time between each abdominal ultrasound scan allowed for more accurate measurements. This is the only study showing reproducibility in a true clinical sitting. The short time between each scan allowed for better comparison. In our literature search, there were no other prospective studies with a higher number of participants. Sarris and colleagues performed a study with approximately the same number of participants; 175 [5]. The ultimate aim of the study was to show reproducibility of fetal measurement in a clinical setting similar to daily practice.

Differences in measurements between different racial groups have been noted by researchers examining inter- and intra-race variations in measurement of fetal biometric parameters. Shipp and colleagues found that there are significant differences between white, black, and Asian ethnic groups in femur length measurements [13]. Authors of [14] have also found a significant difference in head circumference, abdominal circumference, femur length, and estimated fetal weight between migrant Moroccan and Turkish pregnant mothers and autochthonous Belgian pregnant mothers. Interracial differences thereby exist between fetuses from different maternal race groups. However, intra-racial differences have to be examined further before discerning any significance to measured fetal biometric values in these groups. Jordan presents as a suitable area for measuring biometric parameters of fetuses since it is a relatively a small country that comprises of 98% Arabs [15]. Most of the migrants in the country are from bordering countries such as Syria, Palestine and Iraq, thereby decreasing the possibility of any significant differences in measurement due to racial and ethnic variation.

Limitations of this study include the two-dimensional nature of the imaging involved in the study which is deemed inferior to the more advanced 3-D technology used in other institutions worldwide, and the lack of intra-observer measurement variance. However, the aim was to assess fetal variability in a true clinical sitting in a busy hospital. Assessing intra-observer variation would have increased the time required to perform a single exam. Another limitation is that participants were low-risk women and no growth restricted fetuses or macromesomic fetuses were scanned. This may be examined in the future to calculate any inter- or intra-observer bias and variation in the fetal biometric measurements.

Conclusion

This study shows that inter-observer agreement was high when testing biometric fetal parameter measurements. It is also reproducible by different individuals in the same clinical sitting. Fetal head circumference, abdominal circumference and femur length were highly reproducible with small variation between different operators. Higher variability between different operators is noted at older gestational age. Large well organized studies are always needed to ensure accurate fetal measurements due to the important clinical implications of inaccurate measurements.

Disclosure of interests

The authors report no conflict of interest.

Contribution to authorship

This paper was written by Wesam Aleyadeh and Rami Kilani. Measurements were performed by Luay Abu Atieleh, Abdul Mane Al Suleimat, Maysa Khadra, and Hassan M Hawamdeh.

Details of ethics approval

Ethical approval was obtained from the Hashemite University Ethical Committee and from Al Bashir Hospital ethical committee.

Funding

No external source of funding.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Acknowledgements

This paper would not have been possible without the support of AlBashir Hospital.

Appendix. Master list of measurements.
References


Patient GA Obstetrician 1 Obstetrician 2 Hadlock Eq

<table>
<thead>
<tr>
<th></th>
<th>HC</th>
<th>AC</th>
<th>FL</th>
<th>HC</th>
<th>AC</th>
<th>FL</th>
<th>EPW</th>
</tr>
</thead>
<tbody>
<tr>
<td>164</td>
<td>18.86</td>
<td>158</td>
<td>134</td>
<td>274</td>
<td>157</td>
<td>136</td>
<td>272</td>
</tr>
<tr>
<td>165</td>
<td>24.28</td>
<td>224</td>
<td>192</td>
<td>427</td>
<td>230</td>
<td>194</td>
<td>425</td>
</tr>
<tr>
<td>166</td>
<td>28.72</td>
<td>269</td>
<td>248</td>
<td>552</td>
<td>267</td>
<td>250</td>
<td>550</td>
</tr>
<tr>
<td>167</td>
<td>21.00</td>
<td>172</td>
<td>159</td>
<td>330</td>
<td>170</td>
<td>158</td>
<td>333</td>
</tr>
<tr>
<td>168</td>
<td>34.72</td>
<td>314</td>
<td>303</td>
<td>648</td>
<td>316</td>
<td>300</td>
<td>650</td>
</tr>
<tr>
<td>169</td>
<td>33.56</td>
<td>295</td>
<td>282</td>
<td>630</td>
<td>300</td>
<td>284</td>
<td>628</td>
</tr>
<tr>
<td>170</td>
<td>29.42</td>
<td>276</td>
<td>229</td>
<td>527</td>
<td>275</td>
<td>231</td>
<td>526</td>
</tr>
<tr>
<td>171</td>
<td>28.56</td>
<td>268</td>
<td>240</td>
<td>540</td>
<td>270</td>
<td>237</td>
<td>533</td>
</tr>
<tr>
<td>172</td>
<td>27.86</td>
<td>260</td>
<td>211</td>
<td>502</td>
<td>262</td>
<td>210</td>
<td>500</td>
</tr>
<tr>
<td>173</td>
<td>21.00</td>
<td>173</td>
<td>154</td>
<td>301</td>
<td>168</td>
<td>152</td>
<td>336</td>
</tr>
<tr>
<td>174</td>
<td>30.00</td>
<td>286</td>
<td>264</td>
<td>591</td>
<td>285</td>
<td>266</td>
<td>590</td>
</tr>
<tr>
<td>175</td>
<td>27.14</td>
<td>255</td>
<td>215</td>
<td>420</td>
<td>256</td>
<td>215</td>
<td>520</td>
</tr>
<tr>
<td>176</td>
<td>26.28</td>
<td>236</td>
<td>208</td>
<td>490</td>
<td>233</td>
<td>206</td>
<td>488</td>
</tr>
<tr>
<td>177</td>
<td>23.42</td>
<td>218</td>
<td>178</td>
<td>399</td>
<td>214</td>
<td>175</td>
<td>400</td>
</tr>
<tr>
<td>178</td>
<td>29.00</td>
<td>263</td>
<td>257</td>
<td>564</td>
<td>265</td>
<td>256</td>
<td>560</td>
</tr>
<tr>
<td>179</td>
<td>31.56</td>
<td>303</td>
<td>283</td>
<td>630</td>
<td>305</td>
<td>284</td>
<td>630</td>
</tr>
<tr>
<td>180</td>
<td>32.00</td>
<td>308</td>
<td>281</td>
<td>624</td>
<td>307</td>
<td>282</td>
<td>624</td>
</tr>
</tbody>
</table>