Inter-observer variability in the assessment of ultrasound features of polycystic ovaries

Rami Kilani a, Wesam Aleyadeh a,*, Fida Asali a, Firas Al-Rshouda a, Maysa Khadra b

a Hashemite University, Zarqa, Jordan
b University of Jordan, Amman, Jordan

ABSTRACT

Objective: To evaluate inter-observer variability in the assessment of ultrasound features of polycystic ovaries.

Design: Prospective cohort study.

Population: 60 females with known polycystic ovarian syndrome (PCOS).

Setting: Two tertiary care hospitals in Amman, Jordan; Prince Hamza Hospital and Albashir Hospital.

Methods: A total of 120 transvaginal scans were performed on 60 study participants with known PCOS by two gynecologists. The ovaries were evaluated for the presence or absence of PCOS criteria: The number of follicles and the volume of each ovary. The correlation coefficient was calculated. Bland-Altman plots were used to analyze any discrepancies in the measurement of ovaries between the observers.

Main outcome measures: Identification of ultrasound features of PCOS will be reproducible by different gynecologists.

Results: The mean follicle count and ovarian volume in PCOS patients reported in this study were 21 cm³ follicles and 11.5 cm³ respectively. The Pearson correlation coefficient was calculated and it can be seen that there was low inter-observer correlation in follicular count (0.560) and a moderate correlation in measuring ovarian volume (0.770). Bland-Altman plots show low inter-observer agreement in follicular count and high inter-observer agreement in the measurement of ovarian volume.

Conclusion: The inter-observer agreement in the assessment of ultrasound features of polycystic ovaries is not acceptable. The diagnostic criteria of polycystic ovarian morphology need to be revisited.

1. Introduction

Polycystic ovarian syndrome was first described by Stein and Leventhal as an association of amenorrhea, obesity and an atypical appearance of polycystic and enlarged ovaries at laparotomy [1]. It is the most frequently occurring endocrine pathology among women of reproductive age, with an estimated prevalence of 5–10% of women in reproductive age [2]. The diagnostic criteria for PCOS remains an area of heated debate since the pathophysiology of PCOS is not well understood. The NIH criteria is the first definition that received consensus from the scientific community, which defined PCOS as the combined presence of hyperandrogenism and/or hyperandrogenomena, oligoovulation and exclusion of related disorders such as Cushing’s syndrome, hyperprolactinemia, and congenital adrenal hyperplasia [3]. The Rotterdam criteria for PCOS serves as the second definition to receive consensus and describes PCOS as the presence of at least two of three cardinal features: Oligoovulation, hyperandrogenism and polycystic ovary morphology [4]. Polycystic ovarian morphology was characterized as containing 12 or more follicles measuring 2–9 mm and/or an increased ovarian volume of more than 10 cm³ [5]. The difference between the NIH criteria and the Rotterdam criteria is an emphasis of polycystic ovarian morphology as a separate cardinal feature in the latter.

Only two studies have examined inter- and intra-observer variability when making the ultrasound diagnosis of PCOS. Both studies have shown poor inter-observer agreement in identifying features of PCOS on ultrasound. The first was a prospective observational study undertaken to evaluate the inter-observer reliability of ultrasound diagnosis of PCOS. The authors of this study evaluated 27 women and concluded that there is significant variability between different operators when attempting to diagnose PCOS using ultrasound criteria [6]. A more recent study investigated inter-observer agreement when identifying and quantifying...
individual ultrasonographic features of polycystic ovaries. Four operators evaluated the digital recording of transvaginal ultrasound of thirty women with PCOS and found that the inter-observer agreement between different operators was low [7].

Ultrasound technology has advanced significantly in recent years. The identification of PCOS using ultrasound must be easier to apply and the recorded inter-observer variability must decrease for a reliable diagnosis to be made. In this study we want to measure inter-observer variability in the assessment of ultrasound features of polycystic ovaries and interpret and discuss the results as they are reflected on real life applications and diagnoses of PCOS. We will discuss the methods and the process of patient examination in Section 2. Section 3 will highlight the results, while Section 4 will discuss current evidence and explore the significance of our results in real life applications. Section 5 will highlight our conclusions.

2. Materials and methods

2.1. Examination process

Ethical approval was obtained from the Hashemite University Ethics Committee and from Prince Hamzah Hospital Ethics Committee.

Spanning the period from December 1, 2015 to March 1, 2016, one hundred and twenty women were recruited for this study. Women recruited for the study group were diagnosed with PCOS using the 2003 international consensus guidelines of having two of three characteristics:

- Oligoovulation or anovulation
- Clinical and/or biochemical evidence of hyperandrogenism
- Polycystic ovaries on ultrasound (≥12 follicles measuring 2–9 mm in diameter or an ovarian volume >10 cm³) [4]

Exclusion criteria were those using hormonal contraception, fertility medications and/or valproate in the two months prior to enrolment and the inability to visualize the ovaries on vaginal ultrasound. 81 participants with PCOS were approached and 8 participants were excluded. 6 participants were excluded for the use of fertility medication in the 2 month prior to the study. 2 participants were excluded due to the use of combined birth control pills. All participants were tested for prolactin levels, cortisol levels, thyroid function test, and 17 hydroxy-progesterone levels.

Women who met the inclusion criteria and agreed to participate in the study were administered two transvaginal 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, a Samsung Medison R5 (South Korea), with a transvaginal transducer. The two gynecologists performing the examination were certified obstetricians and gynecologists with significant experience in administering transvaginal ultrasound scans. The scans were performed at a random time during the menstrual cycle of each patient. Each ovary was visualized and the anatomic orientation relative to the utero-ovarian ligament was established. Ovaries were scanned from the inner to outer margins in both transverse and sagittal planes. Gynecologists performing the scans were asked to count the total number of follicles (≥2 mm) in each ovary. Ovarian volume was then calculated using the equation for a spheroid from measurements of the largest and widest diameters of the ovaries in the transverse and sagittal planes [8].

2.2. Statistical analysis

Data used for the descriptive statistics were obtained from clinical and laboratory records. Mean measurements of follicular counts and ovarian volume were compared among observers using the inter-class correlation coefficient (Pearson coefficient). Bland-Altman graphs were used to analyze the degree of agreement between observers. Guidelines for evaluating the level of agreement among scores were: >0.80 for high/good, 0.60–0.80 for moderate/fair, and <0.60 for low/poor [9].

3. Results

A total of 72 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, mainly due to the degree of their obesity. Acceptable images were obtained and included in the study in 60 participants. The mean age of participants was 27.4 years with a range of 18.0–35.0 years. Participants had a mean BMI of 30.2% with a range of 20–40%. The mean menstrual cycle length was 74 days with a range of 31–211 days. Clinical and metabolic features of the women participating in the study can be seen in Table 1.

Descriptive statistics of the follicular counts and ovarian volume can be seen in Table 2. It can be seen in Figs. 1 and 2 that the number of follicles found on each ovary has poor agreement between observers and there is random distribution around the equality line indicating no bias. It can also be seen in Figs. 3 and 4 that ovarian volume measurement has good agreement between observers due to the relation of each value to the equality line. It can also be seen in Table 3 that there is poor inter-observer correlation in obtaining follicular count and moderate inter-observer correlation in measuring ovarian volume after measuring the Pearson correlation coefficient.

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

### Table 1

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean</th>
<th>Range</th>
<th>Normal value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>27.4</td>
<td>18–35</td>
<td>–</td>
</tr>
<tr>
<td>BMI (%)</td>
<td>30.2</td>
<td>20.0–40.0</td>
<td>20.0–25.0</td>
</tr>
<tr>
<td>Menstrual cycle length (days)</td>
<td>74.0</td>
<td>31.0–211</td>
<td>21.0–35.0</td>
</tr>
<tr>
<td>LH:FSH</td>
<td>2.40</td>
<td>0.70–6.70</td>
<td>&lt;2.00</td>
</tr>
<tr>
<td>SHBG (nmol/L)</td>
<td>44.0</td>
<td>16.0–76.0</td>
<td>18.0–114</td>
</tr>
<tr>
<td>Fasting glucose (mmol/L)</td>
<td>4.90</td>
<td>4.00–6.40</td>
<td>&lt;6.10</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Feature</th>
<th>Examiner A Mean</th>
<th>Examiner B Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rt. ovarian follicular count</td>
<td>22.5 ± 7.80</td>
<td>21.0 ± 6.70</td>
</tr>
<tr>
<td>Lt. ovarian follicular count</td>
<td>24.5 ± 7.20</td>
<td>23.3 ± 6.60</td>
</tr>
<tr>
<td>Rt. ovarian volume (cm³)</td>
<td>11.3 ± 1.80</td>
<td>11.4 ± 1.50</td>
</tr>
<tr>
<td>Lt. ovarian volume (cm³)</td>
<td>11.7 ± 1.40</td>
<td>11.8 ± 1.10</td>
</tr>
</tbody>
</table>
observers against the mean with the limits of agreement (±2 Standard Deviation) are plotted in Figs. 5 and 6. We notice that the 95% limits for the differences between examiners are high and clinically significant indicating poor agreement. The Bland-Altman graphs for the inter-observer difference in ovarian volume measurement against the mean indicates better agreement (Figs. 7 and 8). 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.

4. Discussion

Despite advancements in ultrasound technology, agreement in the assessment of ultrasound features of polycystic ovaries remains at an inadequate level. In this study, the mean follicular count and ovarian volume in PCOS patients reported were 21 follicles and 11.5 cm³ respectively which meets the Rotterdam criteria for diagnosing PCOS.

Previous studies looking at follicle count in PCOS showed poor agreement between different examiners [7]. Other studies that
looked at follicular counts in subfertility patients to assess antral follicle count showed better agreement. Scheffer et al. compared two groups of women with regular menstrual cycles where one group consisted of healthy volunteers with proven fertility and the other group consisted of patients visiting the general infertility clinic [10]. In each woman, 2D or 3D vaginal scans was performed in the early follicular phase (days 2, 3 or 4) of the menstrual cycle to measure the number of antral follicles (2–10 mm). The results of this study showed that the reproducibility of follicular count was

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Inter-observer Pearson correlation coefficient for right and left ovarian follicular count and volume measurements.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pearson correlation coefficient</td>
</tr>
<tr>
<td>Rt. ovarian follicular count</td>
<td>0.560</td>
</tr>
<tr>
<td>L.t. ovarian follicular count</td>
<td>0.560</td>
</tr>
<tr>
<td>Rt. ovarian volume</td>
<td>0.793</td>
</tr>
<tr>
<td>L.t. ovarian volume</td>
<td>0.748</td>
</tr>
</tbody>
</table>

Fig. 3. Right ovarian volume as measured by Examiners A and B.

Fig. 4. Left ovarian volume as measured by Examiners A and B.
good but it declined with higher follicle counts. The researchers also found that follicular count reliability declines with higher follicular count. Mercé et al. examined inter-observer agreement and found that there was high correlation between different examiners [11]. Researchers in [12] looked at antral follicular count reliability and found good agreement between different observers. The different findings in PCOS patients is probably due to the high follicular count in those patients compared to general infertility patients. Lujan et al. reported an average of 33 follicles per patient which is much higher than the follicular count reported in the other studies. It is important to recognize that the current ultrasound guidelines only necessitate the ability to reliably count 12 follicles throughout the entire ovary. All reported studies showed that observers were consistent in identifying at least 12 follicles per ovary.
Despite the criteria requiring observing 12 follicles per ovary for successful diagnosis of polycystic ovarian morphology, we are interested in assessing the reliability of total follicle counts. Recent literature suggests that the current criteria of observing 12 follicles per ovary must be amended to 19 follicles per ovary to adequately discriminate between polycystic and normal ovaries [13]. Other studies suggest a higher number of 26 follicles per ovary [14].

Previous studies that evaluated ovarian volume measurement reliability indicate similar findings to ours with good agreement between observers. Lujan et al. reported good agreement between different observers evaluating volume in patients with PCOS. Other studies evaluating ovarian volume in other populations of PCOS patients with infertility also reported good reproducibility of measurements [15,16]. Better agreement when calculating ovarian volume suggests that this criterion may serve as a more reliable indicator of polycystic ovaries than follicular counts. However, there still remains a debate regarding the most appropriate threshold for diagnosing polycystic ovarian morphology in relation to...
ovarian volume. The currently accepted cut-off of >10 cm³ was associated with high specificity and low sensitivity [4] in discriminating between normal and polycystic ovaries. There are current recommendations to change the threshold to 7 cm³ [17]. Other studies suggested changing the threshold to 13 cm³ to improve the sensitivity of ovarian volume measurement as a criterion for polycystic ovarian morphology [18].

The strengths of this study included the prospective nature of the study allowing for adequate and accurate follow-up of participants, the high number of participants when compared to previous publications giving us a more reliable measurement of bias, and the short time between each transvaginal ultrasound scan. Both scans were also live which provides better image quality than reviewing images, videos or volume packages [19].

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.

5. Conclusion

This study shows that inter-observer agreement in the assessment of ultrasound features of polycystic ovaries remains inadequate for the correct diagnosis of PCOS. Rotterdam ultrasound criteria underestimates the appropriate number of follicles and ovarian volume needed to diagnose PCOS. Larger studies are needed to better understand the morphological features of PCOS. Diagnostic criteria of polycystic ovarian morphology need to be revisited.

Disclosure of interests

The authors report no conflict of interest.

Contribution to authorship

This paper was written by Rami Kilani and Wesam Aleyadeh. Measurements were performed by Fida Asali, Firas Rshoud, Maysa Khadr, and Rami Kilani.

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.

Acknowledgements

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

References


Please cite this article in press as: R. Kilani et al., Inter-observer variability in the assessment of ultrasound features of polycystic ovaries, Middle East Fertil Soc J (2017). http://dx.doi.org/10.1016/j.mefs.2017.03.005