Human Reproduction, Vol.28, No.5 pp. 1361–1368, 2013
Advanced Access publication on March 15, 2013 doi:10.1093/humrep/det062

human reproduction

ORIGINAL ARTICLE *Reproductive endocrinology*

Updated ultrasound criteria for polycystic ovary syndrome: reliable thresholds for elevated follicle population and ovarian volume

Marla E. Lujan[1,*], Brittany Y. Jarrett[1], Eric D. Brooks[1], Jonathan K. Reines[1], Andrew K. Peppin[2], Narry Muhn[2], Ehsan Haider[2], Roger A. Pierson[3], and Donna R. Chizen[3]

[1]Division of Nutritional Sciences, Cornell University, 216 Savage Hall, Ithaca, NY 14853, USA [2]Diagnostic Radiology, McMaster University, Hamilton, ON, Canada [3]Department of Obstetrics, Gynecology & Reproductive Sciences, University of Saskatchewan, Saskatoon, SK, Canada

*Correspondence address. Tel: +1-607-255-3153; Fax: +1-607-255-1033; E-mail: mel245@cornell.edu

Submitted on July 18, 2012; resubmitted on February 7, 2013; accepted on February 18, 2013

STUDY QUESTION: Do the ultrasonographic criteria for polycystic ovaries supported by the 2003 Rotterdam consensus adequately discriminate between the normal and polycystic ovary syndrome (PCOS) condition in light of recent advancements in imaging technology and reliable methods for estimating follicle populations in PCOS?

STUDY ANSWER: Using newer ultrasound technology and a reliable grid system approach to count follicles, we concluded that a substantially higher threshold of follicle counts throughout the entire ovary (FNPO)—26 versus 12 follicles—is required to distinguish among women with PCOS and healthy women from the general population.

WHAT IS KNOWN ALREADY: The Rotterdam consensus defined the polycystic ovary as having 12 or more follicles, measuring between 2 and 9 mm (FNPO), and/or an ovarian volume (OV) >10 cm^3. Since their initial proposal in 2003, a heightened prevalence of polycystic ovaries has been described in healthy women with regular menstrual cycles, which has questioned the accuracy of these criteria and marginalized the specificity of polycystic ovaries as a diagnostic criterion for PCOS.

STUDY DESIGN, SIZE, DURATION: A diagnostic test study was performed using cross-sectional data, collected from 2006 to 2011, from 168 women prospectively evaluated by transvaginal ultrasonography. Receiver operating characteristic (ROC) curve analyses were performed to determine the appropriate diagnostic thresholds for: (i) FNPO, (ii) follicle counts in a single cross section (FNPS) and (iii) OV. The levels of intra- and inter-observer reliability when five observers used the proposed criteria on 100 ultrasound cases were also determined.

PARTICIPANTS/MATERIALS, SETTING, METHODS: Ninety-eight women diagnosed with PCOS by the National Institutes of Health criteria as having both oligo-amenorrhea and hyperandrogenism and 70 healthy female volunteers recruited from the general population. Participants were evaluated by transvaginal ultrasonography at the Royal University Hospital within the Department of Obstetrics, Gynecology and Reproductive Sciences, University of Saskatchewan (Saskatoon, SK, Canada) and in the Division of Nutritional Sciences' Human Metabolic Research Unit, Cornell University (Ithaca, NY, USA).

MAIN RESULTS: Diagnostic potential for PCOS was highest for FNPO (0.969), followed by FNPS (0.880) and OV (0.873) as judged by the area under the ROC curve. An FNPO threshold of 26 follicles had the best compromise between sensitivity (85%) and specificity (94%) when discriminating between controls and PCOS. Similarly, an FNPS threshold of nine follicles had a 69% sensitivity and 90% specificity, and an OV of 10 cm^3 had a 81% sensitivity and 84% specificity. Levels of intra-observer reliability were 0.81, 0.80 and 0.86 when assessing FNPO, FNPS and OV, respectively. Inter-observer reliability was 0.71, 0.72 and 0.82, respectively.

LIMITATIONS, REASONS FOR CAUTION: Thresholds proposed by this study should be limited to use in women aged between 18 and 35 years.

WIDER IMPLICATIONS OF THE FINDINGS: Polycystic ovarian morphology has excellent diagnostic potential for detecting PCOS. FNPO have better diagnostic potential and yield greater diagnostic confidence compared with assessments of FNPS or OV. Whenever possible, images throughout the entire ovary should be collected for the ultrasonographic evaluation of PCOS.

STUDY FUNDING AND COMPETING INTEREST: This study was funded by Cornell University and fellowship awards from the Saskatchewan Health Research Foundation and Canadian Institutes of Health Research. The authors have no conflict of interests to disclose.

Key words: polycystic ovaries / ultrasonography / diagnosis / polycystic ovary syndrome

Introduction

Polycystic ovary syndrome (PCOS) is a complex endocrine condition in which ovulatory dysfunction and androgen excess are cardinal features (Azziz et al., 2009). In 2003, ultrasonographic evidence of polycystic ovaries was included as a third diagnostic criterion, since there was sufficient evidence worldwide supporting polycystic ovaries as a consistent finding in women with clinical and endocrine features of PCOS (The Rotterdam ESHRE/ASRM-sponsored PCOS Consensus Workshop Group, 2004a,b). Polycystic ovarian morphology was defined as the presence of 12 or more follicles, measuring between 2 and 9 mm, throughout the entire ovary (FNPO) and/or an ovarian volume (OV) >10 cm^3 (The Rotterdam ESHRE/ASRM-sponsored PCOS Consensus Workshop Group, 2004a,b). The threshold for OV was based on expert opinion that there was cumulative evidence reporting a larger mean volume of >10 cm^3 for polycystic ovaries (Franks, 2006), while the threshold for FNPO was based on a single study reporting this value to have a 75% sensitivity and 99% specificity to distinguish between controls and women with PCOS (Jonard et al., 2003). The appearance of at least one polycystic ovary on ultrasound was considered sufficient for the diagnosis.

Since their initial proposal in 2003, an increasing number of reports have appeared questioning the utility of polycystic ovaries to serve as a marker of PCOS. A heightened prevalence of polycystic ovaries has been described in healthy women with regular menstrual cycles, which has marginalized the specificity of polycystic ovaries as a diagnostic criterion for PCOS (Duijkers and Klipping, 2010; Johnstone et al., 2010; Kristensen et al., 2010). Some have questioned the accuracy of the ultrasound criteria, citing that thresholds supported by the Rotterdam consensus for follicle counts and OV do not account for the known influence of age (Alsamarai et al., 2009; Duijkers and Klipping, 2010; Kristensen et al., 2010) or a potential effect of ethnicity (Chen et al., 2008; Köşüş et al., 2011). Others have recommended a re-evaluation of the ultrasound criteria based on recent improvements in ultrasound image technology (Allemand et al., 2006; Dewailly et al., 2011) and newly developed methods for reliably estimating follicle populations in polycystic ovaries (Lujan et al., 2010a).

We have previously reported that significant intra- and inter-observer variability exists when counting follicles throughout the entire polycystic ovary (Lujan et al., 2008; Lujan et al., 2009)—unlike estimates of follicle counts in normal ovaries (Scheffer et al., 2002; Jayaprakasan et al., 2007). Since diagnostic thresholds for follicles counts are based on reliable estimates of FNPO in both the normal and irregular condition, it is critical to use reproducible methods for counting follicles and to report rates of reliability among observers to attest to the accuracy of the methodology and interpretative ability of those involved in the analysis. We recently developed a method for counting follicles in polycystic ovaries, which yielded consistent and reliable counts among observers (Lujan et al., 2010a). Our approach involves compartmentalizing the ovary into grid sections and performing focused follicle counts on individual segments of the ovary to generate estimations of FNPO. We demonstrated that the agreement between multiple observers was exceptional when a grid system was used to count follicles in women with polycystic ovaries (Lujan et al., 2010a). Moreover, little to no variation was noted when a single observer assessed the same images, further corroborating the utility of this method for making follicle count measurements in polycystic ovaries (Lujan et al., 2010a).

Since a reliable definition of polycystic ovarian morphology is critical for facilitating the clinical diagnosis of PCOS and paramount for investigating phenotypic variations in this condition, we endeavored to revisit the diagnostic thresholds for polycystic ovaries supported by the Rotterdam consensus. We felt it prudent to revisit diagnostic criteria for follicle counts in a single cross section (FNPS), in addition to FNPO and OV, since many clinicians and scientists continue to use static images in a single cross-sectional view of the ovary to make their diagnosis. We reasoned that a reliable method for counting follicles and the use of newer ultrasound scanner technology would yield higher thresholds for follicle counts in polycystic ovaries. In addition, we hypothesized that providing medical imaging specialists with serial images throughout the entire ovary, rather than static images, would be associated with higher levels of reliability and confidence when making the ultrasonographic diagnosis of PCOS.

Methods

Study subjects

Study participants were recruited from the general population using ads seeking healthy women of reproductive age or women with concerns over outward features of PCOS such as irregular periods, excess hair growth, obesity and/or infertility. Ninety-eight women diagnosed with PCOS by the National Institutes of Health criteria as having both oligo-amenorrhea and hyperandrogenism were recruited to the study. Oligo-amenorrhea was defined as a history of unpredictable menstrual cycles shorter than 21 days or longer than 36 days. Hyperandrogenism was defined as a modified hirsutism score ≥ 7 (internally validated value having a 83% sensitivity and 96% specificity to distinguish between PCOS and controls) and/or an elevated total testosterone value ≥ 3.96 nmol/l (internally validated value having a 87% sensitivity and 100% specificity to distinguish between PCOS and controls). Seventy women from the general population with regular menstrual cycles and no hyperandrogenism served as controls. Volunteers ranged in age from 18 to 35 years and could not have used hormonal contraception, fertility medications or insulin sensitizers in the 3 months prior to enrollment.

Participants were ineligible if they had a previous history of ovarian surgery or current abnormalities in cortisol, prolactin, thyroid hormone, dehydroepiandrosterone sulfate or 17-hydroxyprogesterone secretion.

Ultrasonography procedures and measurements

Participants were evaluated by transvaginal ultrasonography by two experienced ultrasonographers. Control subjects were scanned on Days 2–5 of the menstrual cycle and women with PCOS were scanned at an unspecified time. Images were included for analysis if there was an absence of a dominant follicle ($\geq$10 mm) and corpus luteum. Ovaries were scanned from their inner to outer margins in the longitudinal plane using a 5–9-MHz transducer on an Ultrasonix RP System (Version 2.3.5, Vancouver, BC, Canada) or a 6–12-MHz transducer on a GE Voluson E8 System (GE Healthcare, Milwaukee, WI, USA). Digital cineloops throughout each ovary (DICOM file format) and static images of the largest cross-sectional view of each ovary (JPEG file format) were digitally archived for off-line analysis.

Ultrasound images of each ovary were analyzed using Santesoft DICOM Editor software (©Emmanouil Kanellopoulus, Athens, Greece) for the following parameters: (i) FNPO, (ii) FNPS and (iii) OV. Reliable follicle counts were achieved for each ovary by imposing a programmable grid system onto the viewing window (Fig. 1) as previously described (Lujan *et al.*, 2010a). Based on an intra-class correlation coefficient analysis, the level of inter-observer agreement for FNPO and FNPS by three observers was 0.84 and 0.94, respectively. OV was estimated using the equation: $\pi/6$ (transverse diameter) × (anteroposterior diameter) × (longitudinal diameter). The level of inter-observer agreement for OV by three observers was 0.96. A value for FNPO, FNPS and OV for each participant was designated as the mean recorded values of the left and right ovaries rounded to the nearest whole number.

Reliability analysis

Normal and polycystic ovary case files were randomly selected and duplicated for evaluation by five observers. Observers were asked to judge the presence or absence of polycystic ovaries in 100 case files based on new proposed thresholds for: (i) FNPO, (ii) FNPS and (iii) OV. For the FNPO end-point, observers were provided with a digital cineloop of a longitudinal sweep through the ovary (DICOM file format). For the FNPS end-point, a single cross-sectional view of the ovary was presented for analysis (JPEG file format). For the OV end-point, two digital cineloops, one in the longitudinal plane and the other in the respective orthogonal plane, was provided for analysis. Observers were also asked to rank their subjective level of diagnostic confidence when using each criterion to assess the individual cases for polycystic ovarian morphology. A 5-point ranking scale was used (1 = no confidence in diagnosis to 5 = complete confidence in diagnosis).

Biochemical assays

Total testosterone was measured by isotope dilution liquid chromatography tandem mass spectrometry as previously described (Lujan *et al.*, 2010b). Sex hormone-binding globulin (SHBG) was measured by a commercially available two-site chemiluminescent immunogenic assay (Siemens Healthcare Diagnostics, Deerfield, IL, USA). Intra- and inter-assay coefficients of variation were <7.4% for both assays. The FAI was calculated by dividing the total testosterone level by the SHBG level and multiplying by 100.

Ethical considerations

This study was approved by Cornell University's Institutional Review Board and the University of Saskatchewan Biomedical Research Ethics Review Board. Interactions with human participants occurred at the Royal University Hospital within the Department of Obstetrics, Gynecology and Reproductive Sciences, University of Saskatchewan (Saskatoon, SK, Canada) from 2006 to 2008, and in the Division of Nutritional Sciences' Human Metabolic Research Unit, Cornell University (Ithaca, NY, USA) from 2009 to 2011. Informed, written consent was obtained from all study participants. Ultrasound images used in the analyses were de-identified.

Statistical analysis

JMP 7 Statistical Software (SAS Institute Inc., Cary, NC, USA), MedCalc Version 12.3.0 (MedCalc Software, Mariakerke, Belgium) and GraphPad Prism Version 5.0 (GraphPad Software, San Diego, CA, USA) were used to perform the analyses. Linearity tests were carried out using the Shapiro–Wilk W Goodness-of-Fit test, and logarithmic transformations of the data confirmed unimodal distributions for ultrasonographic end-points (FNPO, FNPS and OV). Descriptive statistics (median and inter-quartile ranges) were tabulated for clinical, hormonal and ultrasonographic features. Mann–Whitney tests were performed to compare parameters between women with PCOS and controls. Accuracy of FNPO, FNPS and OV to discriminate between PCOS and controls was evaluated using a receiver operating characteristic (ROC) curve analysis. Assuming an expected area under the ROC curve of 0.80, a minimum of 36 participants in each group was required to detect a difference from chance alone

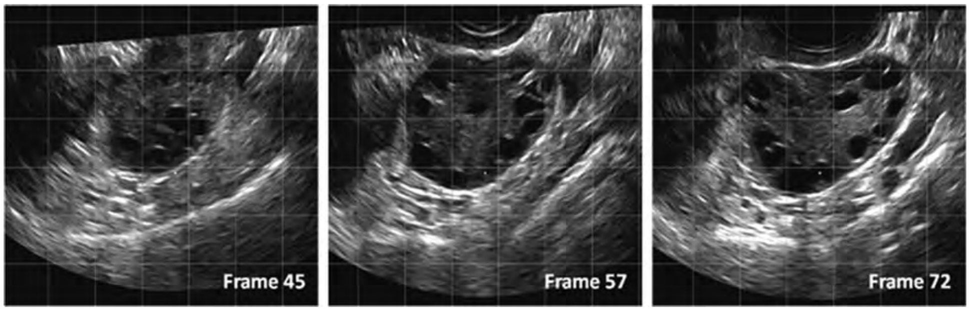

Figure 1 Serial frames of an ultrasound scan through a polycystic ovary during off-line processing. Application of a grid system allowed for systematic counting of follicles within a single grid section. The total number of follicles per ovary or in a single cross section was determined by summing follicle counts made in each grid section.

at an alpha level of 0.05 and a beta level of 0.10 (where chance alone represents an area under the curve equal to 0.50). Based on our current study cohorts, a *post hoc* power analysis determined that our study had 99% power to detect differences in FNPO, FNPS and OV among women with PCOS and controls. Diagnostic thresholds for FNPO, FNPS and OV were proposed based on Youden's index, which balanced maximum test sensitivity and test specificity. The level of agreement when multiple observers used the proposed criteria to diagnose PCOS was determined by Cohen's kappa statistic. Differences in reliability when using the proposed criteria were assessed by Tukey's multiple comparison tests. Frequency tables were generated for the observers' subjective assessment of diagnostic confidence and the percent response was calculated. A P-value <0.05 was regarded as statistically significant.

Results

A comparison of clinical, hormonal and ultrasonographic features in women with PCOS and controls is presented in Table I. Women with PCOS were similar in age to controls ($P = 0.204$), but reported longer intervals between menses ($P < 0.001$), had higher BMI ($P < 0.001$), hirsutism scores ($P < 0.001$), testosterone levels ($P = 0.007$) and free androgen indexes ($P < 0.001$). The mean follicle counts and mean OV in women with PCOS and controls are summarized in Fig. 2. As expected, FNPO were higher in women with PCOS compared with controls ($P < 0.001$), as were FNPS ($P < 0.001$) and OV ($P < 0.001$).

The level of diagnostic accuracy for FNPO, FNPS and OV to distinguish between women with PCOS and controls is summarized in Table II. Diagnostic potential, as judged by the area under the ROC curve, was 0.969 for FNPO, 0.880 for FNPS and 0.873 for OV. An FNPO threshold of 26 follicles had the best compromise between sensitivity (85%) and specificity (94%) when discriminating between controls and women with PCOS (Table II). Likewise, an FNPS threshold of nine follicles had a 69% sensitivity and 90% specificity and an OV threshold of 10 cm^3 had 81% sensitivity and 84% specificity (Table II). The level of diagnostic accuracy calculated using our method for commonly used thresholds for polycystic ovaries is included in Table II for comparison with the newly proposed criteria.

When five observers used the newly proposed thresholds to detect PCOS in 100 case files, the level of intra-observer reliability ranged from 0.72 to 0.92 for FNPO, from 0.68 to 0.96 for FNPS and from 0.75 to 0.96 for OV (Table III). Overall, the level of intra-observer agreement when identifying polycystic ovaries using the proposed criteria was highest for OV followed by FNPO and FNPS, but no

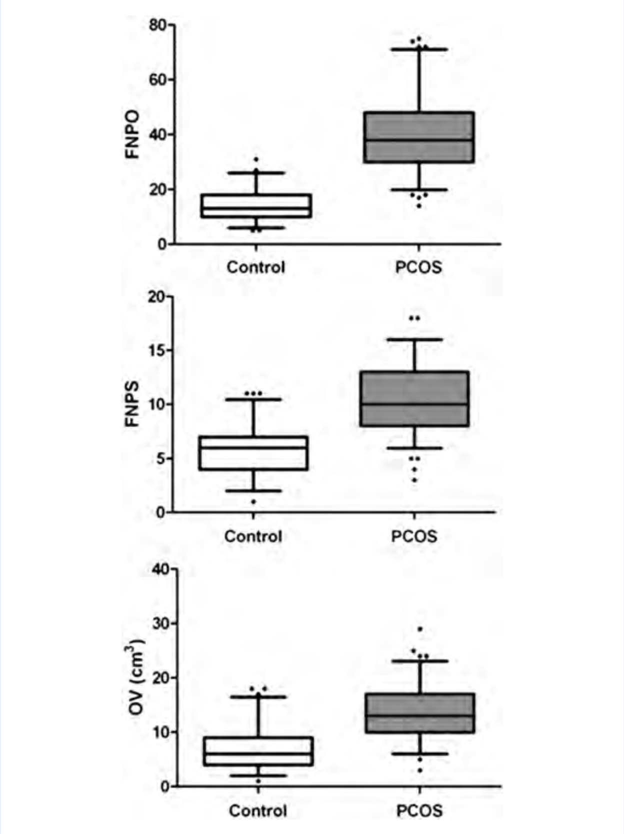

Figure 2 A comparison of FNPO, FNPS and OV in women with PCOS and controls. Box-and-whisker diagrams of FNPO, FNPS and OV are presented for women with PCOS ($N = 98$) and controls ($N = 70$). Boxes represent the 25th and 75th percentile and the horizontal band within the box represents the median. The 5th–95th percentile range is denoted by the vertical bars. OV and follicle counts in both the entire ovary and a single cross section were significantly higher in women with PCOS versus controls. PCOS, Polycystic ovary syndrome.

Table I A comparison of clinical, hormonal and ultrasonographic features in women with PCOS and controls.

	Control (n = 70)	PCOS (n = 98)	P-value
Age (years)	27 (23–35)	28 (25–32)	0.204
BMI (kg/m^2)	23.9 (22.0–27.5)	30.1 (23.7–37.3)	<0.001
Menstrual cycle length (d)	29 (28–30)	74 (46–128)	<0.001
Hirsutism score	2 (0–5)	11 (7–14)	<0.001
Total testosterone (nmol/l)	2.76 (2.21–3.33)	3.32 (2.50–4.50)	0.007
Free androgen index	4 (3–7)	11 (6–18)	<0.001
Mean follicle count per ovary	13 (10–18)	38 (30–48)	<0.001
Mean follicle count per cross-section	6 (4–7)	10 (8–13)	<0.001
Mean OV (cm^3)	6 (4–9)	13 (10–17)	<0.001

Median values are presented with 25–75th quartiles in parentheses.

Table II Sensitivity and specificity of newly proposed diagnostic thresholds for FNPO, FNPS and OV.

Criterion	Area under ROC curve (95% CI)	Threshold	Sensitivity (%)	Specificity (%)
FNPO	0.969* (0.948, 0.990)	12[a]	100	36
		15[b]	99	54
		19[c]	96	77
		20[d]	96	79
		26	**85**	**94**
FNPS	0.880* (0.830, 0.930)	**9**	**69**	**90**
		10[d,e]	58	94
OV (cm^3)	0.873* (0.817, 0.930)	7[c,f]	95	53
		8[g]	93	61
		9[h]	88	71
		10	**81**	**84**
		11[i]	74	86
		13[j]	60	90

ROC curve, receiver operating characteristic curve.
Comparison with previously reported thresholds provided:
[a] Jonard *et al.* (2003);
[b] Fox (1999);
[c] Dewailly *et al.* (2011);
[d] Allemand *et al.* (2006);
[e] Adams *et al.* (1985);
[f] Jonard *et al.* (2005);
[g] Chen *et al.* (2008);
[h] Atiomo *et al.* (2000); [i] van Santbrink *et al.* (1997);
[j] Fulghesu *et al.* (2001).
*$P < 0.0001$ compared with chance alone.

Table III Level of intra-observer agreement among five observers diagnosing polycystic ovaries using newly proposed diagnostic criteria for FNPO, FNPS and OV.

	Cohen's kappa statistic		
Observer	FNPO	FNPS	OV
1	0.92	0.85	0.96
2	0.92	0.96	0.80
3	0.75	0.68	0.96
4	0.72	0.72	0.75
5	0.72	0.77	0.84
Average	0.81[a]	0.80[a]	0.86[a]

Significant differences for within-row comparisons are denoted by different letters ($P < 0.050$).

Table IV Level of inter-observer agreement among five observers diagnosing polycystic ovaries using newly proposed diagnostic criteria for FNPO, FNPS and OV.

	Cohen's kappa statistic		
Observer	FNPO	FNPS	OV
1, 2	0.72	0.72	0.86
1, 3	0.62	0.60	0.84
1, 4	0.64	0.68	0.77
1, 5	0.70	0.70	0.82
2, 3	0.74	0.70	0.84
2, 4	0.84	0.76	0.79
2, 5	0.78	0.86	0.80
3, 4	0.68	0.72	0.84
3, 5	0.66	0.68	0.82
4, 5	0.72	0.72	0.82
Average	0.71[a]	0.72[a]	0.82[b]

Significant differences for within-row comparisons are denoted by different letters ($P < 0.001$).

differences in reliability were detected among criteria ($P = 0.571$). The level of inter-observer reliability ranged from 0.62 to 0.84 for FNPO, 0.60 to 0.86 for FNPS and 0.77 to 0.86 for OV (Table IV). Levels of inter-observer agreement when identifying PCO using the proposed criteria was greater for OV compared with both FNPO ($P < 0.001$) and FNPS ($P < 0.001$).

When observers were asked to rank their subjective level of diagnostic confidence when evaluating 100 randomly selected case files, the use of criteria for FNPO was associated with higher levels of confidence compared with FNPS and OV (Fig. 3). Namely, observers reported being completely confident in their diagnosis 72% of the

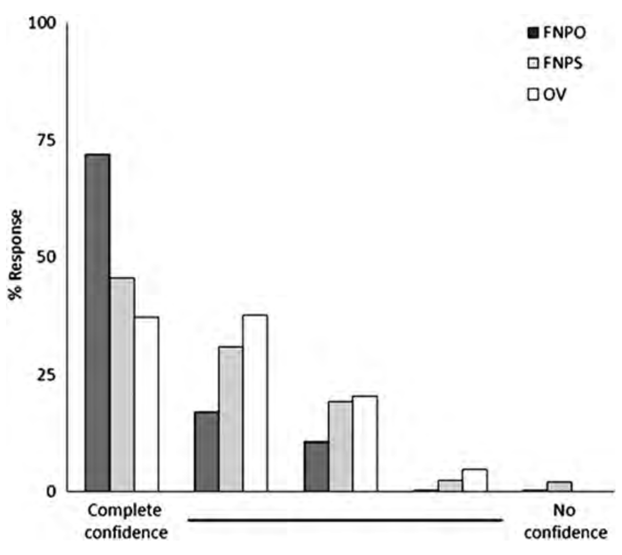

Figure 3 Frequency distribution of diagnostic confidence ratings when using newly proposed diagnostic criteria for FNPO, FNPS and OV. Observers reported higher levels of confidence in their diagnosis when using criteria for FNPO compared with FNPS and OV.

time when using FNPO, compared with only 45 and 37% of the time when using FNPS and OV, respectively.

Discussion

The objective of the current study was to revisit the ultrasonographic criteria for PCOS. Since being introduced as a diagnostic criterion in 2003, numerous reports have questioned the validity of polycystic ovaries to serve as an objective measure of PCOS. Much of the controversy has arisen from reports of unusually high rates of polycystic ovaries in healthy women of reproductive age using the ultrasound-based criteria supported by the Rotterdam consensus (Duijkers and Klipping, 2010; Johnstone *et al.*, 2010; Kristensen *et al.*, 2010). Using newer ultrasound technology and a reliable grid system approach to count follicles, we concluded that a substantially higher threshold of FNPO—26 versus 12 follicles—is required to distinguish among women with PCOS and healthy women from the general population.

The threshold for FNPO proposed by this study is more than double that proposed by the Rotterdam consensus. The difference in follicle counts reflects the fact that we noted substantially more follicles on ultrasonography within both normal and polycystic ovaries. There are three main factors that may have contributed to differences in follicle counts among studies. First, we imaged the ovaries using substantially newer ultrasound technology. The image quality afforded by newer scanners has markedly improved since the time of the Rotterdam consensus and the improved resolution likely allowed for the detection of more follicles in our study, particularly those smaller follicles in the range of 1–2 mm. Secondly, the criteria supported by the Rotterdam consensus were based on a study in which investigators estimated FNPO in real time (Jonard *et al.*, 2003). In our experience, obtaining reproducible follicle counts in polycystic ovaries is exceptionally difficult (Lujan *et al.*, 2008; Lujan *et al.*, 2009). The sheer number and crowding of follicles make it hard to track individual follicles and ensure that one is not overlooking both distant and adjacent follicles. That our counts were higher than those previously reported could be attributed to a reduced likelihood of missing smaller follicles when focused counts are performed off-line on ovaries broken up into smaller, more manageable sections using a grid system, and when the evaluator is equipped with the ability to flag individual follicles as they are counted. Thirdly, our analysis was based on a comparison of follicle counts in healthy women with regular menstrual cycles recruited from the population at large, in contrast to the Jonard *et al*. study in which the control cohort was comprised of subfertile women being treated at a fertility clinic. While derivation of control subjects from referred hospital patients occurs commonly in the literature, their designation as normal is somewhat marred and may not necessarily reflect a true sample of the general population.

Our proposed threshold for FNPO is also higher than more recent reports of revised ultrasound criteria for PCOS (Allemand *et al.*, 2006; Dewailly *et al.*, 2011). In 2006, Allemand *et al*. recommended a threshold of 20 follicles per ovary based on their findings in a small group of subjects using 3D ultrasonography. Their 3D findings have yet to be reproduced and remained tempered by the use of a small sample size, which included only 10 subjects with PCOS. More recently, Dewailly *et al*. (2011) proposed a threshold of 19 follicles per ovary based on a comparison of real-time follicle counts in a much larger group of women using new ultrasound technology ($N = 62$ for PCOS and $N = 66$ for controls). Differences in findings among studies might again be explained by differences in methods for counting follicles and the recruitment/composition of control populations used. On average, we noted 8–10 more follicles per polycystic ovary compared with Dewailly *et al*., which is consistent with the notion that follicle counts tend to be higher when off-line approaches are used in this population. Dewailly *et al*. (2011) obtained control subjects from patients seeking evaluation at their reproductive endocrinology clinic ($N = 105$) and applied a cluster analysis to this control cohort to exclude a population of women who could be considered as having functional evidence of PCOS ($N = 39$ excluded). Since the excluded control subjects had higher follicle counts compared with the remaining controls, it is possible that a higher FNPO threshold would have resulted had their threshold been derived using the original control group whose median follicle counts were more in line with our own.

The use of polycystic ovaries as an inclusion/exclusion criterion for a diagnostic test study is controversial. In our current study, neither the control or PCOS populations were screened for polycystic ovaries prior to inclusion in the study. We felt this approach was warranted for establishing revised thresholds for polycystic ovarian morphology since it avoided the use of any of the contested ultrasound criterion proposed to date. In the case of our control population, the use of the follicle threshold for polycystic ovaries supported by the Rotterdam consensus would have excluded 60% of participants who were otherwise healthy with normal endocrine function and regular menstrual cycles. That such a high number of healthy women would have been excluded reinforces the conclusion that the Rotterdam ultrasound criteria are inaccurate and that implementation of revised criteria for polycystic ovaries is urgently needed in light of advancements in imaging technology. Moreover, we felt it prudent not to screen for polycystic ovaries since this approach eliminated the potential for bias, where there is uncertainty pertaining to both the actual clinical spectrum of PCOS and the physiological range of ovarian follicle populations in unselected women from the general population. As such, there is debate as to whether polycystic ovaries alone represent a variant of PCOS (Adams *et al.*, 2004; Ng *et al.*, 2006; Mortensen *et al.*, 2009; Dewailly *et al.*, 2011; Catteau-Jonard *et al.*, 2012) or whether they reflect natural variations in ovarian function among women and/or across the lifespan (Hassan and Killick, 2003; Murphy *et al.*, 2006; Johnstone *et al.*, 2010). To date, the few studies that have compared endocrine and metabolic function in asymptomatic women with polycystic ovaries to controls have yielded conflicting results pertaining to abnormalities in androgens (Chang *et al.*, 2000; Adams *et al.*, 2004; Mortensen *et al.*, 2009; Johnstone *et al.*, 2010 versus Carmina *et al.*, 1997; Catteau-Jonard *et al.*, 2012), gonadotropins (Johnstone *et al.*, 2010; Catteau-Jonard *et al.*, 2012; versus Carmina *et al.*, 1997; Adams *et al.*, 2004) and cardiometabolic risk factors (Carmina *et al.*, 1997; Chang *et al.*, 2000; Adams *et al.*, 2004 versus Johnstone *et al.*, 2010; Catteau-Jonard *et al.*, 2012). While a recent report of anti-Müllerian hormone levels in asymptomatic women with polycystic ovaries suggests a granulosa cell defect (Catteau-Jonard *et al.*, 2012), these data are tempered by earlier findings that ovarian dysfunction was not uniformly apparent in healthy women with polycystic ovaries (Carmina *et al.*, 1997; Chang *et al.*, 2000; Mortensen *et al.*, 2009), and that the presence of polycystic ovaries did not confer progression to PCOS (Murphy *et al.*, 2006). Taken together, there is a definite need for further evaluation of polycystic-appearing ovaries in asymptomatic women using new-age imaging technology and careful consideration of criteria used to define polycystic ovarian morphology.

Despite major advances in ultrasound image technology, the threshold proposed for FNPS by our current study is remarkably similar to that proposed by Adams *et al.* over 25 years ago using transabdominal ultrasonography (i.e. 9 versus 10 or more FNPS, respectively) (Adams *et al.*, 1985). This similarity supports the notion that improvements in ultrasound image technology alone cannot account for differences in follicle thresholds reported to this point. Very little difference in counts was found among studies. This observation likely relates to the fact that reliability in follicle counts would be expected to be higher when counts are performed in a single cross section and a single experienced imaging specialist is involved in the collection and analysis of the images. While the ‘Adams criteria’ remain in widespread use, our assessment of diagnostic accuracy suggests that a single cross-sectional view of the ovary is not optimal for detecting polycystic ovaries. The single cross-sectional assessment approach relies heavily on the interpretive and technical skills of the sonographer. Since most diagnostic medical sonographers receive little training on PCOS during their gynecological imaging rotation, they may be apt to provide cross-sectional views of the ovary that rule out other pathologies, such as ovarian neoplasia or torsion, but are not optimal for the evaluation of polycystic ovarian morphology. It was, therefore, not surprising that observers in our study reported better diagnostic confidence when presented with the images throughout the entire ovary rather than a single cross-sectional view of the ovary.

The OV threshold proposed in this current study is in-line with the 2003 Rotterdam recommendations (The Rotterdam ESHRE/ASRM-sponsored PCOS Consensus Workshop Group, 2004a,b). Estimates of OV were associated with the highest levels of reliability, which was not entirely surprising since the quotients for OV rely on few measurements while estimates of the follicle counts are more numerous and by virtue, prone to more error. While measurements of OV were associated with the highest levels of inter- and intra-observer reliability, diagnostic accuracy and diagnostic confidence for this parameter were lowest reflecting the greater likelihood of overlap in OV among controls and women with PCOS. Despite this, OV should still be considered a helpful parameter when evaluating ovarian morphology. It can be used in situations in which image quality is reduced and the ability to obtain reliable follicle counts compromised. Since PCOS is often accompanied by obesity, and the excess abdominopelvic adiposity reduces image quality in these patients, it may be possible to gauge the contours/limits of the ovary for measurements of OV and determination of polycystic ovarian morphology.

In summary, ultrasonographic features of ovarian morphology have substantial diagnostic potential to distinguish between women with PCOS and healthy women with regular menstrual cycles. FNPO have better diagnostic potential to distinguish between controls and women with PCOS compared with counts in a single cross section or OV. Clinically, levels of intra- and inter-observer reliability were similar among methods, yet observers reported greater diagnostic confidence when evaluating FNPO. An average value of 26 or more follicles per ovary is a reliable threshold for detecting polycystic ovaries in women with frank manifestation of PCOS. As such, we acknowledge that a lower follicle threshold may be required to detect milder variants of the syndrome. Our proposed thresholds were derived from measurements made off-line since they afforded the opportunity to obtain reproducible follicle counts in polycystic ovaries, which are a requisite for establishing diagnostic thresholds and for research studies evaluating the relevance of polycystic ovarian morphology in PCOS. Because reliability in real-time counts has been shown to decrease with increasing follicle counts (Scheffer *et al.*, 2002), we recommend that off-line measurements involving interpretation of images collected throughout the entire ovary be performed whenever possible.

Authors’ roles

M.E.L. conceived, designed and coordinated the study, analyzed and interpreted the data, drafted the manuscript and provided resources to complete the study. A.K.P., E.D.B., J.K.R., B.Y.J., N.M. and E.H. participated in data collection and assisted in the statistical analyses and interpretation of the data. R.A.P. participated in the conception and design of the study, provided resources to complete the study. D.R.C. clinically evaluated the study volunteers, participated in the conception and design of the study and provided resources to complete the study. All authors evaluated the manuscript and contributed to its content.

Funding

This study was funded by Cornell University and fellowship awards from the Saskatchewan Health Research Foundation and Canadian Institutes of Health Research.

Conflict of interest

None declared.

References

Adams J, Franks S, Polson DW, Mason HD, Abdulwahid N, Tucker M, Morris DV, Price J, Jacobs HS. Multifollicular ovaries: clinical and endocrine features and response to pulsatile gonadotropin releasing hormone. *Lancet* 1985;**70**:1375.

Adams JM, Taylor AE, Crowley WF Jr, Hall JE. Polycystic ovarian morphology with regular ovulatory cycles: insights into the pathophysiology of polycystic ovarian syndrome. *J Clin Endocrinol Metab* 2004;**89**:4343–4350.

Allemand MC, Tummon IS, Phy JL, Foong SC, Dumesic DA, Session DR. Diagnosis of polycystic ovaries by three-dimensional transvaginal ultrasound. *Fertil Steril* 2006;**85**:214–219.

Alsamarai S, Adams JM, Murphy MK, Post MD, Hayden DL, Hall JE, Welt CK. Criteria for polycystic ovarian morphology in polycystic ovary syndrome as a function of age. *J Clin Endocrinol Metab* 2009; **12**:4961–4970.

Atiomo WU, Pearson S, Shaw S, Prentice A, Dubbins P. Ultrasound criteria in the diagnosis of polycystic ovary syndrome (PCOS). *Ultrasound Med Biol* 2000;**26**:977–980.

Azziz R, Carmina E, Dewailly D, Diamanti-Kandarakis E, Escobar-Morreale HF, Futterweit W, Janssen OE, Legro RS, Norman RJ, Taylor AE *et al.* The Androgen Excess and PCOS Society criteria for the polycystic ovary syndrome: the complete task force report. *Fertil Steril* 2009;**91**:456–488.

Carmina E, Wong L, Chang L, Paulson RJ, Sauer MV, Stanczyk FZ, Lobo RA. Endocrine abnormalities in ovulatory women with polycystic ovaries on ultrasound. *Hum Reprod* 1997;**12**:905–909.

Catteau-Jonard S, Bancquart J, Poncelet E, Lefebvre-Maunoury C, Robin G, Dewailly D. Polycystic ovaries at ultrasound: normal variant or silent polycystic ovary syndrome? *Ultrasound Obstet Gynecol* 2012; **40**:223–229.

Chang PL, Lindheim SR, Lowre C, Ferin M, Gonzalez F, Berglund L, Carmina E, Sauer MV, Lobo RA. Normal ovulatory women with polycystic ovaries have hyperandrogenic pituitary-ovarian responses to gonadotropin-releasing hormone agonist testing. *J Clin Endocrinol Metab* 2000;**85**:995–1000.

Chen Y, Li L, Chen X, Zhang Q, Wang W, Li Y, Yang D. Ovarian volume and follicle number in the diagnosis of polycystic ovary syndrome in Chinese women. *Ultrasound Obstet Gynecol* 2008;**32**:700–703.

Dewailly D, Gronier H, Poncelet E, Robin G, Leroy M, Pigny P, Duhamel A, Catteau-Jonard S. Diagnosis of polycystic ovary syndrome (PCOS): revisiting the threshold values of follicle count on ultrasound and of the serum AMH level for the definition of polycystic ovaries. *Hum Reprod* 2011;**26**:3123–3129.

Duijkers IJ, Klipping C. Polycystic ovaries, as defined by the 2003 Rotterdam consensus criteria, are found to be very common in young healthy women. *Gynecol Endocrinol* 2010;**26**:152.

Fox R. Transvaginal ultrasound appearances of the ovary in normal women and hirsute women with oligoamenorrhea. *Aust N Z J Obstet Gynaecol* 1999;**39**:63–68.

Franks S. Diagnosis of polycystic ovarian syndrome: in defense of the Rotterdam criteria. *J Clin Endocrinol Metab* 2006;**91**:786–789.

Fulghesu AM, Ciampelli M, Belosi C, Apa R, Pavone V, Lanzone A. A new ultrasound criterion for the diagnosis of polycystic ovary syndrome: the ovarian stroma/total area ratio. *Fertil Steril* 2001;**76**:326–331.

Hassan MA, Killick SR. Ultrasound diagnosis of polycystic ovaries in women who have no symptoms of polycystic ovary syndrome is not associated with subfecundity or subfertility. *Fertil Steril* 2003; **80**:966–975.

Jayaprakasan K, Walker KF, Clewes JS, Johnson IR, Raine-Fenning NJ. The interobserver reliability of off-line antral follicle counts made from stored three-dimensional ultrasound data: a comparative study of different measurement techniques. *Ultrasound Obstet Gynecol* 2007;**29**:335–341.

Johnstone EB, Rosen MP, Neril R, Trevithick D, Sternfeld B, Murphy R, Addauan-Andersen C, McConnell D, Pera RR, Cedars MI. The polycystic ovary post-Rotterdam: a common, age-dependent finding in ovulatory women without metabolic significance. *J Clin Endocrinol Metab* 2010;**95**:4965–4972.

Jonard S, Robert Y, Cortet-Rudelli C, Pigny P, Decanter C, Dewailly D. Ultrasound examination of polycystic ovaries: is it worth counting the follicles? *Hum Reprod* 2003;**18**:598–603.

Jonard S, Robert Y, Dewailly D. Revisiting the ovarian volume as a diagnostic criterion for polycystic ovaries. *Hum Reprod* 2005;**20**:2893–2898.

Köşüş N, Köşüş A, Turhan NÖ, Kamalak Z. Do threshold values of ovarian volume and follicle number for diagnosing polycystic ovarian syndrome in Turkish women differ from western countries? *Eur J Obstet Gynecol Reprod Biol* 2011;**154**:177–181.

Kristensen SL, Ramlau-Hansen CH, Ernst E, Olsen SF, Bonde JP, Vested A, Toft G. A very large proportion of young Danish women have polycystic ovaries: is a revision of the Rotterdam criteria needed? *Hum Reprod* 2010;**25**:3117.

Lujan ME, Chizen DR, Peppin AK, Kriegler S, Leswick DA, Bloski T, Pierson RA. Improving inter-observer variability in the evaluation of ultrasonographic features of polycystic ovaries. *Reprod Biol Endocrinol* 2008;**8**:30.

Lujan ME, Chizen DR, Peppin AK, Dhir A, Pierson RA. Assessment of ultrasonographic features of polycystic ovaries is associated with modest levels of inter-observer agreement. *J Ovarian Res* 2009;**2** [Epub ahead of print].

Lujan ME, Brooks ED, Kepley AL, Chizen DR, Pierson RA, Peppin AK. Grid analysis improves reliability in follicle counts made by ultrasonography in women with polycystic ovary syndrome. *Ultrasound Med Biol* 2010a;**36**:712–718.

Lujan M, Bloski TG, Chizen DR, Lehotay DC, Pierson RA. Digit ratios do not serve as anatomical evidence of prenatal androgen exposure in clinical phenotypes of polycystic ovary syndrome. *Hum Reprod* 2010b; **25**:204–211.

Mortensen M, Ehrmann DA, Littlejohn E, Rosenfield RL. Asymptomatic volunteers with a polycystic ovary are a functionally distinct but heterogeneous population. *J Clin Endocrinol Metab* 2009;**94**:1579–1586.

Murphy MK, Hall JE, Adams JM, Lee H, Welt CK. Polycystic ovarian morphology in normal women does not predict the development of polycystic ovary syndrome. *J Clin Endocrinol Metab* 2006;**91**:3878–3884.

Ng EH, Chan CC, Ho PC. Are there differences in ultrasound parameters between Chinese women with polycystic ovaries only and with polycystic ovary syndrome? *Eur J Obstet Gynecol Reprod Biol* 2006; **25**:92–98.

Scheffer GJ, Broekmans FJM, Bancsi LF, Habbema JDF, Looman CWN, Te Velde ER. Quantitative transvaginal two- and three-dimensional sonography of the ovaries: reproducibility of antral follicle counts. *Ultrasound Obstet Gynecol* 2002;**20**:270–275.

The Rotterdam ESHRE/ASRM-sponsored PCOS Consensus Workshop Group. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. *Fertil Steril* 2004a; **81**:19–25.

The Rotterdam ESHRE/ASRM-sponsored PCOS Consensus Workshop Group. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome (PCOS). *Hum Reprod* 2004b;**19**:41–47.

van Santbrink EJ, Hop WC, Fauser BC. Classification of normogonadotropic infertility: polycystic ovaries diagnosed by ultrasound verses endocrine characteristics of polycystic ovary syndrome. *Fertil Steril* 1997; **67**:452–458.