OXFORD

human reproduction update

Human Reproduction Update, 2024, **30(1),** 109–130

https://doi.org/10.1093/humupd/dmad027

Advance Access Publication Date: October 6, 2023

Ultrasonographic criteria in the diagnosis of polycystic ovary syndrome: a systematic review and diagnostic meta-analysis

Jeffrey Pea[1], Jahnay Bryan[1], Cynthia Wan[1], Alexis L. Oldfield [1], Kiran Ganga[1], Faith E. Carter[1], Lynn M. Johnson[2], and Marla E. Lujan [1,*]

[1]Human Metabolic Research Unit, Division of Nutritional Sciences, Colleges of Human Ecology and Agriculture and Life Sciences, Cornell University, Ithaca, NY, USA

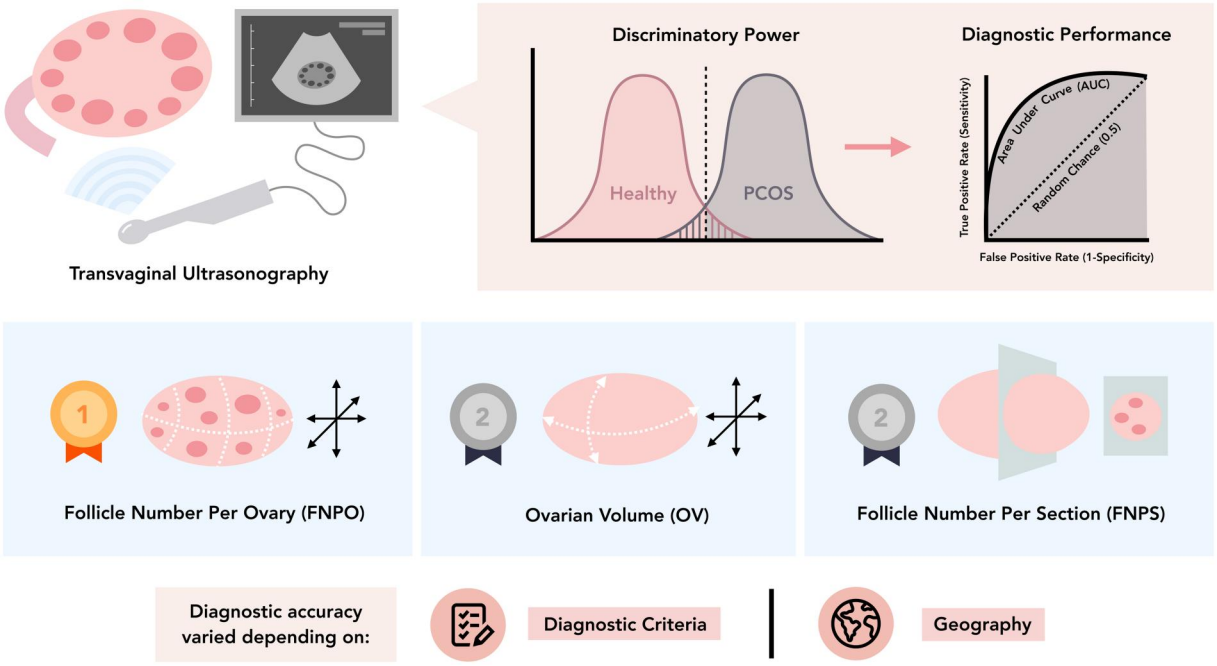
[2]Cornell Statistical Consulting Unit, Cornell University, Ithaca, NY, USA

*Correspondence address. Division of Nutritional Sciences, Colleges of Human Ecology and Agriculture and Life Sciences, Cornell University, Ithaca, NY, USA. E-mail: marla.lujan@cornell.edu https://orcid.org/0000-0002-7203-5814

TABLE OF CONTENTS

Received: May 24, 2022. **Revised:** September 8, 2023. **Editorial decision:** September 18, 2023.

GRAPHICAL ABSTRACT



Systematic review and diagnostic meta-analysis supports follicle number per ovary (FNPO) as the most accurate ovarian ultrasound marker in the diagnosis of PCOS.

ABSTRACT

BACKGROUND: Polycystic ovary morphology (PCOM) on ultrasonography is considered as a cardinal feature of polycystic ovarian syndrome (PCOS). Its relevance as a diagnostic criterion for PCOS was reaffirmed in the most recent International Evidence-Based Guideline for the Assessment and Management of PCOS. However, there remains a lack of clarity regarding the best practices and specific ultrasonographic markers to define PCOM.

OBJECTIVE AND RATIONALE: The aim of this systematic review and diagnostic meta-analysis was to assess the diagnostic accuracy of various ultrasonographic features of ovarian morphology in the diagnosis of PCOS.

SEARCH METHODS: Relevant studies published from 1 January 1990 to 12 June 2023 were identified by a systematic search in PubMed, Web of Science, Scopus, CINAHL, and CENTRAL. Studies that generated diagnostic accuracy measures (e.g. proposed thresholds, sensitivity, specificity) for PCOS using the following ultrasonographic markers met criteria for inclusion: follicle number per ovary (FNPO) or per single cross-section (FNPS), ovarian volume (OV), and stromal features. Studies on pregnant or post-menopausal women were excluded. Risk of bias and applicability assessment for diagnostic test accuracy studies were determined using the QUADAS-2 and QUADAS-C tool for a single index test or between multiple index tests, respectively. Diagnostic meta-analysis was conducted using a bivariate model of pooled sensitivity and specificity, and visualized using forest plots and summary receiver-operating characteristic (SROC) curves.

OUTCOMES: From a total of 2197 records initially identified, 31 studies were included. Data from five and two studies were excluded from the meta-analysis due to duplicate study populations or limited data for the index test, leaving 24 studies. Pooled results of 20 adult studies consisted of 3883 control participants and 3859 individuals with PCOS. FNPO was the most accurate diagnostic marker (sensitivity: 84%, CI: 81–87%; specificity: 91%, CI: 86–94%; AUC: 0.905) in adult women. OV and FNPS had similar pooled sensitivities (OV: 81%, CI: 76–86%; FNPS: 81%, CI: 70–89%) but inferior pooled specificities (OV: 81%, CI: 75–86%; FNPS: 83%, CI: 75–88%) and AUCs (OV: 0.856; FNPS: 0.870) compared to FNPO. Pooled results from four adolescent studies consisting of 210 control participants and 268 girls with PCOS suggested that OV may be a robust ultrasonographic marker for PCOS diagnosis albeit the current evidence remains limited. The majority of the studies had high risk of bias for the patient selection (e.g. lack of randomized/consecutive patient selection) and index test (e.g. lack of pre-proposed thresholds for comparison) domains across all ultrasonographic markers. As such, diagnostic meta-analysis was unable to determine the most accurate cutoff for ultrasonographic markers to diagnose PCOS. Subgroup analysis suggested that stratification based on previously proposed diagnostic thresholds, age, BMI, or technology did not account for the heterogeneity in diagnostic accuracy observed across the studies. Studies that diagnosed PCOS using the Rotterdam criteria had improved sensitivity for FNPO. Studies from North America had lower diagnostic accuracy when compared to Asian studies (FNPO: sensitivity) and European studies (OV: specificity, diagnostic odds ratio and positive likelihood ratio). Geographic differences in diagnostic accuracy may potentially be due to differences in age, BMI, and diagnostic criteria of the PCOS group across regions.

WIDER IMPLICATIONS: This diagnostic meta-analysis supports the use of FNPO as the gold standard in the ultrasonographic diagnosis of PCOS in adult women. OV and FNPS provide alternatives if total antral follicle counts cannot be accurately obtained. Our findings support the potential for ultrasonographic evidence of PCOM in adolescents as more data becomes available. Subgroup analysis suggests the need to investigate any relative contributions of geographical differences on PCOS phenotypes. These findings may provide the basis for the development of strategies and best practices toward a standardized definition of PCOM and a more accurate ultrasonographic evaluation of PCOS.

Keywords: ultrasonography / polycystic ovary syndrome / ovary / ovarian follicle / diagnostic imaging

Introduction

Polycystic ovary syndrome (PCOS) is a complex endocrine disorder that impacts up to 13% of women of reproductive age, as a leading cause of anovulatory infertility (Azziz *et al.*, 2016). Symptoms and signs of PCOS include a combination of biochemical, clinical, and morphological indicators of androgen excess and ovulatory dysfunction (Teede *et al.*, 2018a). In particular, polycystic ovary morphology (PCOM) was established as a key feature of PCOS since the earliest descriptions of the condition (Stein and Leventhal, 1935). PCOM is commonly evaluated using pelvic ultrasonography and described as an ovarian enlargement (i.e. increased ovarian volume (OV)) and/or an excess of small antral follicles, either within the entire ovary (FNPO) or within a single cross-sectional image (FNPS) (Dewailly *et al.*, 2014b). Other morphological features, including the peripheral organization of follicles within the ovary and the echogenicity and relative abundance of the ovarian stroma, have been investigated in the context of PCOS but their utility to define PCOM is less defined (Christ *et al.*, 2014; Dewailly *et al.*, 2014b). Despite its namesake, controversy has persisted surrounding the relevance of PCOM to serve as a diagnostic criterion for PCOS due to factors such as its prevalence in regular cycling women (Johnstone *et al.*, 2010; Kristensen *et al.*, 2010), the emphasis on biochemical or clinical assessments of PCOS by previous consensus diagnostic criteria (Zawadski and Dunaif, 1992; Azziz *et al.*, 2006), as well as a lack of standardization across methodology and technology used to assess and define PCOM (Dewailly *et al.*, 2014b).

In 2018, the International Evidence-based Guideline for the Assessment and Management of PCOS reaffirmed the inclusion of PCOM in its recommended diagnostic criteria for PCOS (Teede *et al.*, 2018a). However, the authors noted that evidence-based recommendations to define PCOM were not available and further research to comprehensively assess its diagnostic performance in detecting PCOS was needed (Teede *et al.*, 2018b).

Therefore, the primary purpose of this systematic review and meta-analysis was to evaluate the diagnostic accuracy of ultrasonographic ovarian markers in the diagnosis of PCOS. In addition, a secondary aim was to identify potential strategies and best practices that would work toward a standardized definition of PCOM.

Methods

The systematic review was conducted in accordance to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for Diagnostic Test Accuracy (McInnes *et al.*, 2018). The PICO criteria were defined before the literature search and detailed in Supplementary Table S1. Concisely, our study question was, in comparing women with or without PCOS (P), what is the most accurate ultrasonographic criteria (I) to diagnose PCOS (O)?

The study protocol was registered and is available at the International Prospective Register of Systematic Reviews (PROSPERO) (Registration ID: CRD42021259118).

Search strategy

A systematic search of published literature was initially conducted in the electronic databases of MEDLINE, Institute for Scientific Information (ISI) Web of Science, Scopus, Cochrane Central Register of Controlled Trials (CENTRAL), and Cumulative Index to Nursing and Allied Health Literature (CINAHL) from 31 December 2020 through 8 January 2021 using a search strategy based on the PICO framework (Supplementary Table S1). The search was continuously updated to identify the newest relevant studies until 12 June 2023. Details of the search strategy are available (Supplementary Table S2). Further, manual searches of the reference list of included studies supplemented the electronic database searches. Non-English studies, animal studies, or studies published before 1990 were excluded. The last adjustment was chosen to allow the inclusion of studies where the PCOS diagnosis was compiled by the 1990 NIH (Zawadski and Dunaif, 1992), 2003 Rotterdam (The Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group, 2004), 2006 Androgen Excess and PCOS Society (AE-PCOS) (Azziz *et al.*, 2006), or 2018 International Guideline (Teede *et al.*, 2018a) criteria.

Study selection

Studies were included if they met the PICO criteria described in Supplementary Table S1. Briefly, observational (cross-sectional, case–control, cohort) studies or cross-sectional analysis of baseline measures from non-randomized or randomized control trials on women of reproductive age with and without PCOS were included wherein transvaginal (adult) or transabdominal/transrectal (adolescent) ultrasonographic markers were used in the diagnosis of PCOS. Adult women were defined as individuals between 18 and 50 years old. Adolescent girls were defined as individuals <20 years old and at least 1-year post-menarche in line with the recommendations from the 2018 International Guideline (Teede *et al.*, 2018a; Peña *et al.*, 2020).

Systematic reviews, evidence-based guidelines, non-peer-reviewed studies, case series, and animal studies were excluded. In addition, studies fulfilling the following criteria were excluded: studies limited to pregnant and post-menopausal (>50 years old) women; studies wherein the diagnosis of PCOS did not comply with the NIH, Rotterdam, AE-PCOS, or International Guideline criteria; studies that exclusively used other imaging methods, such as MRI or computerized tomography (CT); and studies where data were irretrievable after contacting their corresponding authors. Our primary outcome was diagnostic test accuracy measures (i.e. proposed thresholds, sensitivity, specificity, AUC) of the following ultrasound ovarian markers: follicle number per whole ovary (FNPO) or within a single cross-section/slice (FNPS), OV, and stromal features. Proposed thresholds were defined as those recommended by authors based on their chosen methodology (e.g. Youden's index in ROC analysis, 95th percentile of a healthy population, previously recommended thresholds). The screening process was completed by four investigators (J.P., J.B., C.W., and K.G.) independently using the double-blind coding assignment function of the online Covidence systematic review platform (Covidence.org, Alfred Health, Melbourne, Australia) and Excel. Review of all citations identified by the search strategy was conducted first by title and abstract followed by full text to determine eligibility. All discrepancies were resolved by consensus with an additional investigator (M.E.L.).

Risk of bias and applicability assessment

Methodological quality items and types of bias were evaluated using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool and the QUADAS-C. QUADAS-2 is a validated tool developed specifically for primary diagnostic accuracy studies that addresses four domains: patient selection, index test, reference standard, and flow and timing (Whiting *et al.*, 2011). The first three domains are also assessed for applicability concerns. Risk of bias and applicability concerns are scored as 'low', 'high', or 'unclear'. QUADAS-C is an extension tool used to assess risk of bias between multiple index tests that are compared within diagnostic accuracy studies (Yang *et al.*, 2021). Three investigators

(J.P., A.L.O., and C.W.) independently evaluated the included studies, with any discrepancies being resolved through consensus or discussion with an additional investigator (M.E.L.).

Data extraction

The following data were extracted using a standardized protocol, including: (i) first author's name; (ii) study publication year; (iii) study design and setting; (iv) country of origin; (v) PCOS diagnostic criteria used; (vi) ultrasound machine and ultrasound transducer frequency range; (vii) participants' characteristics, including total sample size and sample size of cases and controls; (viii) participants' age; (ix) participants' BMI; (x) sonographic ovarian measures assessed; and (xi) diagnostic test accuracy measures, including proposed diagnostic thresholds, sensitivity, specificity and AUC values with their corresponding 95% confidence intervals, in the diagnosis of PCOS (per patient). Mean differences and SD of sonographic measures, age, and BMI were collected for both control women and women with PCOS. Where SEM were only reported, SD were calculated using a formula SD = SEM × square root (n), where n is the number of participants (Higgins *et al.*, 2019). When medians and percentile ranges were reported instead of means and SD, we used the median in place of mean and appropriate formulas were used to calculate SD from percentile ranges (Wan *et al.*, 2014; Higgins *et al.*, 2019). In the case of any missing or unclear data, two attempts were made to contact the corresponding author by email to request data or clarify methods. Data extraction was completed by five investigators (J.P., J.B., F.E.C., C.W., and K.G.) independently and was reviewed by all authors (A.L.O., L.M.J., and M.E.L.) for any potential extraction error.

Statistical analysis

Summary statistics for diagnostic accuracy, namely the proposed diagnostic threshold, sensitivity, and specificity values, were generated from included studies. If multiple studies included the same study population, the study with the greatest sample size was included for meta-analysis. A bivariate generalized linear mixed model was used to determine pooled sensitivity and specificity and their 95% confidence intervals. This approach models between-study heterogeneity in the sensitivity and specificity and the correlation between them directly using random effects (Reitsma *et al.*, 2005). When bivariate models did not converge, the correlation between the sensitivity and specificity was set to zero or one of the random effects was removed from the model. The bivariate generalized linear mixed model also determined additional diagnostic measures such as diagnostic odds ratio (DOR) and likelihood ratio (LR). DOR is a single indicator of test performance, similar to accuracy, but is independent of disease prevalence. Similarly, LR is independent of disease prevalence and based on the ratio of sensitivity and specificity. The positive LR (LR+) determines the likelihood of an individual having the disease upon testing whereas a negative LR (LR−) determines the likelihood of an individual not having the disease upon testing.

To describe inter-study heterogeneity, we constructed forest plots and summary receiver-operating characteristic (SROC) curves with 95% prediction regions estimated using bivariate model meta-analysis. Given that meta-analysis of diagnostic test accuracy requires the knowledge of bivariate data in the form of sensitivity and specificity, statistical approaches for systematic reviews of interventions, such as Cochran's Q and I^2 statistic, were avoided. Subgroup analyses were performed to investigate potential sources of heterogeneity determined a priori and as appropriate based on the availability of data. Bivariate models were fitted with the following covariates to explore the influence of: (i) participants' mean age, (ii) participants' mean BMI, (iii) ultrasound transducer frequency, (iv) PCOS diagnostic criteria, (v) country of origin, (vi) ovarian measurement methodology, and (vii) risk of bias. Given that conventional funnel plot asymmetry uses DORs and even sample sizes of diseased and non-diseased groups to determine publication bias, they may be misleading in diagnostic test accuracy studies which have expectedly high DORs and often uneven sample sizes due to case–control study designs or prevalence of disease. Therefore, a modified funnel plot asymmetry method for diagnostic accuracy studies was constructed using the association between log DOR (lnDOR) and 'effective sample size' (ESS), which is a simple function of the number of diseased and non-diseased individuals (Deeks *et al.*, 2005). All analyses were performed (J.P. and L.M.J) using Review Manager (RevMan) 5 (The Nordic Cochrane Center, Copenhagen, Norway), R (R Core Team, 2020), RStudio Version 1.3.959 (Rstudio Team, 2020), and the Hmisc (Harrell, 2021), the ggplot2 (Wickham, 2016), the mada (Doebler, 2020), the metafor (Viechtbauer, 2010), and the msm (Jackson, 2011) packages. Results were considered significant at $P \leq 0.05$.

Results

Search results and characteristics of included studies

Overall, 2197 records were identified through the systematic database searches. In total, 1353 records were excluded during title and abstract screening and 130 records were excluded after full text review (Fig. 1). Finally, 31 studies (Fulghesu *et al.*, 2001; Jonard *et al.*, 2005; Allemand *et al.*, 2006; Chen *et al.*, 2008a,b; Alsamarai *et al.*, 2009; Dewailly *et al.*, 2011, 2014a; Diamanti-Kandarakis *et al.*, 2011; Köşüş *et al.*, 2011a,b; Lujan *et al.*, 2013; Villa *et al.*, 2013; Bili *et al.*, 2014; Christ *et al.*, 2014; Köninger *et al.*, 2014; Çiraci *et al.*, 2015; Villarroel *et al.*, 2015; Carmina *et al.*, 2016; Kim *et al.*, 2017; Lie Fong *et al.*, 2017; Kar and Swoyam, 2018; Wongwananuruk *et al.*, 2018; Ahmad *et al.*, 2019; Ertekin *et al.*, 2019; Jarrett *et al.*, 2019; Khashchenko *et al.*, 2020; Le *et al.*, 2021; Giménez-Peralta *et al.*, 2022; Özay *et al.*, 2022; Vanden Brink *et al.*, 2023) that proposed diagnostic thresholds for sonographic ovarian markers to diagnose PCOS were deemed eligible for inclusion.

General characteristics of the included studies are presented in Table 1. The selected studies included 9144 participants, with 4464 control women and 4680 women with PCOS. Among the adult studies, 18 proposed diagnostic thresholds for FNPO, 21 studies, for OV, and seven studies, for FNPS. In addition, two studies proposed diagnostic thresholds for ovarian area (OA) (Jonard *et al.*, 2005; Christ *et al.*, 2014) and one for ovarian contour (Bili *et al.*, 2014). Five studies proposed diagnostic thresholds using stromal features, including stromal area (SA) (Christ *et al.*, 2014), stromal volume (Kar and Swoyam, 2018), stromal area over total ovarian area (S/A) (Fulghesu *et al.*, 2001), stromal thickness (Özay *et al.*, 2022), and stromal strain ratio (Çiraci *et al.*, 2015). Several studies proposed multiple diagnostic thresholds, including phenotype-specific thresholds (Köninger *et al.*, 2014), left and right ovary thresholds (Jarrett *et al.*, 2019; Özay *et al.*, 2022), three-dimensional (3D) thresholds (Allemand *et al.*, 2006; Kar and Swoyam, 2018), and age-specific thresholds (Kim *et al.*, 2017; Lie Fong *et al.*, 2017; Ahmad *et al.*, 2019). A range of diagnostic thresholds were proposed across each ultrasonographic marker. For FNPO, the included studies proposed a range of cutoffs from 8 to 28 follicles per ovary. For OV, the included studies proposed a range of cutoffs from 6 to 13 cm^3. For FNPS, the included studies proposed a range of cutoffs from 7 to 13 follicles

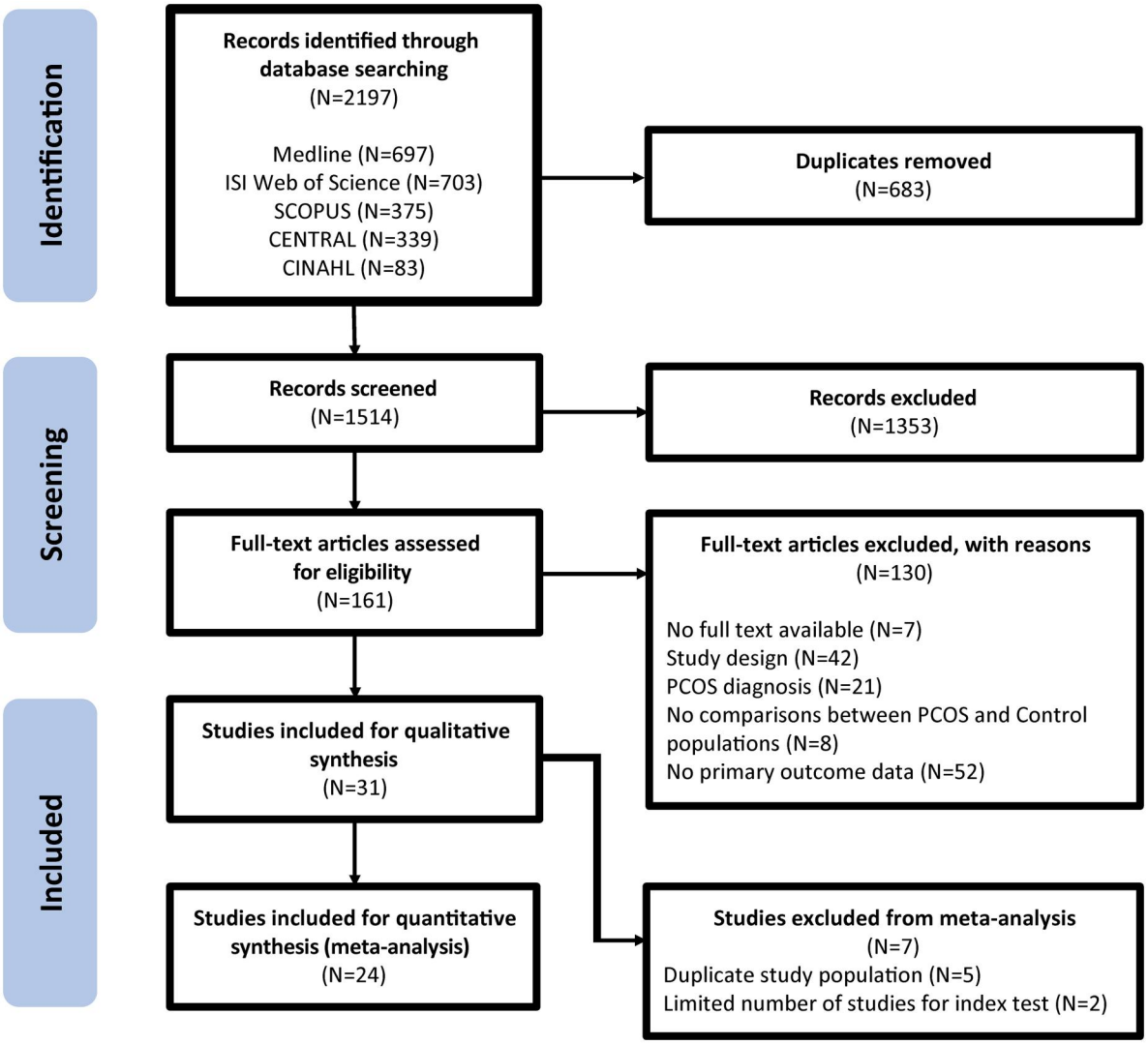


Figure 1. **Preferred Reporting Items for Systematic Review and Meta-Analyses (PRISMA) flow diagram of the systematic search and study selection process.**

per cross-sectional image. The included studies primarily presented ovarian morphology measurements using the average value between the two ovaries (83% for FNPO, 71% for OV, 57% for FNPS) whereas the remainder presented the maximum measurement across both ovaries. Diagnosis of PCOS was mostly conducted using the NIH criteria (18/31 studies) followed by the Rotterdam (12/31 studies) and AE-PCOS criteria (1/31 studies). Four adolescent studies were identified, of which all proposed diagnostic thresholds for OV (Chen *et al.*, 2008b; Villa *et al.*, 2013; Villarroel *et al.*, 2015; Khashchenko *et al.*, 2020) as well as one, for FNPO (Villarroel *et al.*, 2015) and one, for ovarian to uterine index (OUI) (Khashchenko *et al.*, 2020).

Risk of bias and applicability assessments

The proportions of risk of bias and applicability assessments based upon the QUADAS-2 and QUADAS-C tools are presented in Fig. 2. Only one study (Dewailly *et al.*, 2011) met a low risk of bias for the Patient Selection domain. The remaining studies were graded as high risk of bias due to lack of randomized patient selection or consecutive sampling when evaluating individual index tests or comparing diagnostic test accuracy across multiple tests. All studies were also graded for high risk of bias in the Index Test domain primarily due to the methodology for proposed thresholds. None of the proposed thresholds were prespecified and were instead determined using either Youden's index (which balances sensitivity and specificity) or using the 95th percentile of the control group. Both approaches have been used to define cutoffs for PCOM (Dewailly *et al.*, 2014b) but generate a 'threshold effect' that introduces additional bias when comparing diagnostic accuracy. In addition, the 95th percentile cutoff is arbitrary and may not accurately reflect meaningful long-term health outcomes for PCOS patients. A low risk of bias was determined across index tests for most studies for the Reference Standard (FNPO: 58%; OV: 60%; FNPS: 71%; Stroma: 60%) and for Flow and Timing (FNPO: 79%; OV: 80%; FNPS: 71%; Stroma: 80%) domains. Given that the QUADAS-C tool incorporates QUADAS-2 scoring in its assessment of bias between index tests within a study, all studies were graded as high risk of bias in the Index Test domain. However, low risk of bias when comparing index tests within a study was determined for the majority of other three domains (Patient Selection: 16/17 (94%); Reference Standard: 12/17 (71%); Flow and Timing: 16/17 (94%)). For most index tests, concerns for applicability were primarily graded as low in the Patient Selection (FNPO: 79%; OV: 80%; FNPS: 100%)

Table 1. General characteristics of included studies examining the role of ovarian ultrasound markers in PCOS diagnosis.

First author, year of publication	Participants' N	Country	Setting	PCOS diagnosis	Ultrasound machine	Transducer frequency (MHz)	Age (y)	BMI (kg/m^2)	Reported markers evaluated	Proposed diagnostic thresholds
Ahmad *et al.*, 2019	Total: 1001 Control: 756 PCOS: 245	USA	Control: Community PCOS: Academic Medical Center	1990 NIH	Shimadzu SDU-450XL, General Electric Voluson s8 E8C-RS	4–8 (Shimadzu), 4–10 (Voluson)	Control: N/A PCOS: N/A	Control: N/A PCOS: N/A	FNPO, OV (Max)	FNPO ≥15 (25 to <30 yo) FNPO ≥14 (30 to <35 yo) FNPO ≥12 (35 to <40 yo) OV ≥8.50 cm^3 (25 to <30 yo) OV ≥7.00 cm^3 (30 to <35 yo) OV ≥6.25 cm^3 (35 to <40 yo)
Allemand *et al.*, 2006	Total: 39 Control: 29 PCOS: 10	USA	Academic Medical Center	1990 NIH	Philips ATL HDI 5000	4–8	Control: 30.90 ± 3.50 PCOS: 31.20 ± 3.90	Control: 24.00 ± 5.50 PCOS: 32.20 ± 10.80	FNPO, OV, FNPS	FNPO ≥20.1 OV ≥ 13.00 cm^3 FNPS ≥10
Alsamarai *et al.*, 2009	Total: 876 Control: 382 PCOS: 494	USA	Academic Medical Center	1990 NIH	Philips ATL HDI 1500	5	Control: 28.10 ± 6.40 (Younger) 48.10 ± 6.60 (Older) PCOS: 27.80 ± 5.70 (Younger) 46.30 ± 4.50 (Older)	Control: 24.40 ± 5.10 (Younger) 26.70 ± 5.40 (Older) PCOS: 30.60 ± 8.70 (Younger) 31.30 ± 8.50 (Older)	OV (Max), FNPS (Max)	OV ≥7 cm^3 FNPS ≥12
Bili *et al.*, 2014	Total: 83 Control: 40 PCOS: 43	Greece	Academic Medical Center	Rotterdam	General Electric Voluson 730	5–7	Control: 30.80 ± 4.30 PCOS: 28.90 ± 5.00	Control: 22.50 ± 3.70 PCOS: 24.90 ± 5.90	OV, Ovarian Contour (OC)	OV ≥9.64 cm^3 OC ≥8.75 cm
Carmina *et al.*, 2016	Total: 160 Control: 47 PCOS: 113	Italy	Academic Medical Center	Rotterdam	Biosound Esaote MyLab 50 Xvision	8–10	Control: 23.10 ± 4.00 PCOS: 23.00 ± 4.30	Control: 27.00 ± 4.00 PCOS: 27.60 ± 6.00	FNPO, OV	FNPO ≥22 OV ≥8.00 cm^3
Chen *et al.*, 2008a	Total: 585 Control: 153 PCOS: 432	China	Academic Medical Center	1990 NIH	Toshiba Sonolayer SSA-220A (Transvaginal or Transrectal)	6	Control: 27.15 ± 2.33 PCOS: 26.25 ± 2.01	Control: N/A PCOS: N/A	FNPO, FNPO (Max), OV, OV (Max),	FNPO ≥10 FNPO (Max) ≥12 OV ≥6.40 cm^3 OV (Max) ≥7.90 cm^3
Chen *et al.*, 2008b [adolescent]	Total: 95 Control: 69 PCOS: 26	China	Academic Medical Center	1990 NIH	Toshiba Sonolayer SSA-220A (Transrectal)	6	Control: 13.00 ± 1.23 PCOS: 12.57 ± 1.25	Control: 21.77 ± 4.60 PCOS: 20.12 ± 2.17	OV, OV (Max)	OV ≥6.74 cm^3 OV (Max) ≥7.82 cm^3
Christ *et al.*, 2014	Total: 142 Control: 60 PCOS: 82	USA/Canada	Control: Community PCOS: Academic Medical Center	1990 NIH	Ultrasonix RP, General Electric Voluson E8	5–9 (Ultrasonix) 6–12 (Voluson)	Control: 27.00 ± 5.19 PCOS: 28.00 ± 5.19	Control: 23.70 ± 3.93 PCOS: 31.20 ± 10.44	FNPO, OV, OA, SA, FNPS	FNPO ≥28 OV ≥10.00 cm^3 OA ≥5.00 cm^2 SA ≥3.00 cm^2 FNPS ≥9

(continued)

Table 1. Continued

First author, year of publication	Participants' N	Country	Setting	PCOS diagnosis	Ultrasound machine	Transducer frequency (MHz)	Age (y)	BMI (kg/m^2)	Reported markers evaluated	Proposed diagnostic thresholds
Çiraci *et al.*, 2015	Total: 96 Control: 48 PCOS: 48	Turkey	Academic Medical Center	Rotterdam	General Electric Logiq E9	6.5	Control: 27.10 ± 5.20 PCOS: 25.70 ± 4.20	Control: N/A PCOS: N/A	FNPO, OV, Stromal Strain Ratio (SSR)	FNPO ≥12 OV ≥10.00 cm^3 SSR ≥3.80
Dewailly *et al.*, 2011	Total: 128 Control: 66 PCOS: 62	France	Academic Medical Center	1990 NIH	General Electric Voluson E8 Expert	5–9	Control: 30.00 ± 3.86 PCOS: 27.60 ± 4.22	Control: 24.00 ± 5.74 PCOS: 28.00 ± 6.99	FNPO, OV	FNPO ≥19 OV ≥7.00 cm^3
Dewailly *et al.*, 2014a	Total: 716 Control: 621 PCOS: 95	Croatia	Academic Medical Center	1990 NIH	Toshiba Nemio	5–7	Control: 32.50 ± 3.86 PCOS: 29.80 ± 4.29	Control: 23.00 ± 3.34 PCOS: 27.00 ± 6.08	FNPO	FNPO ≥12
Diamanti-Kandarakis *et al.*, 2011	Total: 97 Control: 47 PCOS: 50	Greece	Academic Medical Center	1990 NIH	N/A	N/A	Control: 27.15 ± 6.72 PCOS: 26.46 ± 5.86	Control: 26.27 ± 5.30 PCOS: 26.49 ± 5.00	FNPO	FNPO ≥19.5
Ertekin *et al.*, 2019	Total: 53 Control: 16 PCOS: 37	Turkey	Academic Medical Center	Rotterdam	Samsung RS80	6	Control: 23.0 ± 5.0 PCOS: 21.5 ± 3.7	Control: 25.5 ± 3.8 PCOS: 24.5 ± 4.8	OV	OV ≥7.5
Fulghesu *et al.*, 2001	Total: 83 Control: 30 PCOS: 53	Italy	Academic Medical Center	Rotterdam (Frank)	General Electric Logiq 500	6.5	Control: N/A PCOS: N/A	Control: 23.15 ± 4.49 PCOS: 23.61 ± 3.88	S/A	S/A ≥ 0.34
Giménez-Peralta *et al.*, 2022	Total: 311 Control: 111 PCOS: 200	Spain	Academic Medical Center	Rotterdam	Aplio 500	7	Control: 31.60 ± 0.59 PCOS: N/A	Control: 22.90 ± 0.55 PCOS: N/A	FNPO	FNPO ≥12
Jarrett *et al.*, 2019	Total: 154 Control: 67 PCOS: 87	USA/Canada	Control: Community PCOS: Academic Medical Center	1990 NIH	Ultrasonix RP, General Electric Voluson E8	5–9 (Ultrasonix), 6–12 (Voluson)	Control: 27.00 ± 5.93 PCOS: 27.00 ± 5.19	Control: 23.60 ± 4.22 PCOS: 32.00 ± 10.74	OV	OV ≥10.00 cm^3 (Right) OV ≥9.00 cm^3 (Left)
Jonard *et al.*, 2005	Total: 155 Control: 57 PCOS: 98	France	Academic Medical Center	1990 NIH	General Electric Logiq 400	7	Control: 29.00 ± 4.10 PCOS: 27.20 ± 5.27	Control: 22.90 ± 4.88 PCOS: 27.90 ± 8.07	FNPO, OV, OA	FNPO ≥12 OV ≥7.00 cm^3 OA ≥5.00 cm^2
Kar and Swoyam, 2018	Total: 131 Control: 45 PCOS: 86	India	Tertiary Care Hospital	Rotterdam	General Electric Voluson E8	6–12	Control: 28.45 ± 4.62 PCOS: 26.03 ± 3.52	Control: 23.02 ± 3.58 PCOS: 25.71 ± 4.87	FNPO, OV, SV	FNPO ≥12 OV ≥6.15 cm^3 SV ≥6 cm^3
Khashchenko *et al.*, 2020 [adolescent]	Total: 160 Control: 30 PCOS: 130	Russia	Academic Medical Center	Rotterdam	General Electric Vivid-q	1.8–6.0	Control: 16.0 ± 1.48 PCOS: 16.0 ± 1.48	Control: 20.2 ± 2.52 PCOS: 22.4 ± 5.41	OV, OUI	OV ≥10.7 cm^3 OUI ≥3.95

(continued)

Table 1. Continued

First author, year of publication	Participants' N	Country	Setting	PCOS diagnosis	Ultrasound machine	Transducer frequency (MHz)	Age (y)	BMI (kg/m^2)	Reported markers evaluated	Proposed diagnostic thresholds
Kim *et al.*, 2017	Total: 1210 Control: 666 PCOS: 544	USA	Academic Medical Center	1990 NIH	ATL HDI 1500 Ultrasound, Phillips HD11 XE	4–8	Control: 22.1 ± 1.8 (18–24 yo) 27.2 ± 1.3 (25–29 yo) 32.3 ± 1.3 (30–34 yo) 37.3 ± 1.4 (35–39 yo) PCOS: 22.0 ± 1.9 (18–24 yo) 27.3 ± 1.4 (25–29 yo) 32.4 ± 1.4 (30–34 yo) 37.8 ± 1.5 (35–39 yo)	Control: 23.8 ± 4.4 (18–24 yo) 24.3 ± 4.8 (25–29 yo) 23.8 ± 4.7 (30–34 yo) 26.3 ± 5.8 (35–39 yo) PCOS: 29.4 ± 8.0 (18–24 yo) 30.4 ± 9.5 (25–29 yo) 32.0 ± 9.0 (30–34 yo) 33.5 ± 7.5 (35–39 yo)	OV (Max), FNPS (Max)	OV ≥12 cm^3 (18–24 yo) OV ≥10 cm^3 (25–29 yo) OV ≥9 cm^3 (30–34 yo) OV ≥8 cm^3 (35–39 yo) OV ≥10 cm^3 (40–44 yo) OV ≥6 cm^3 (>44 yo) FNPS ≥13 (18–24 yo) FNPS ≥14 (25–29 yo) FNPS ≥10 (30–34 yo) FNPS ≥10 (35–39 yo) FNPS ≥9 (40–44 yo) FNPS ≥7 (>44 yo)
Köninger *et al.*, 2014	Total: 128 Control: 48 PCOS: 80	Germany	Academic Medical Center	Rotterdam	General Electric Voluson E8 Expert	3–9	Control: 34.00 ± 5.50 PCOS (Severe): 27.10 ± 5.80 PCOS (Mild): 29.30 ± 5.80	Control: 24.30 ± 4.40 PCOS (Severe): 29.10 ± 7.40 PCOS (Mild): 26.70 ± 7.00	FNPO (Max), OV (Max)	FNPO ≥9.5 (Severe PCOS) FNPO ≥8.5 (Mild PCOS) OV ≥10.50 cm^3 (Severe PCOS) OV ≥10.20 cm^3 (Mild PCOS)
Köşüş *et al.* 2011a	Total: 310 Control: 100 PCOS: 210	Turkey	Academic Medical Center	Rotterdam	General Electric Logic 200 Pro	6.5	Control: 26.70 ± 5.60 PCOS: 26.30 ± 5.40	Control: 20.80 ± 2.40 PCOS: 26.50 ± 5.30	OV	OV ≥6.43 cm^3
Köşüş *et al.* 2011b	Total: 316 Control: 65 PCOS: 251	Turkey	Academic Medical Center	AE-PCOS	General Electric Logic 200 Pro	6.5	Control: 26.70 ± 5.60 PCOS: 24.90 ± 6.10	Control: 20.80 ± 2.40 PCOS: 27.10 ± 6.20	FNPO, OV	FNPO ≥8 OV ≥6.43 cm^3
Le *et al.*, 2021	Total: 392 Control: 273 PCOS: 119	Vietnam	Academic Medical Center	Rotterdam	ALOKA ProSound SSD-3500	9	Control: 33.99 ± 4.78 PCOS: 32.66 ± 4.10	Control: 20.82 ± 2.56 PCOS: 21.31 ± 2.80	OV	OV ≥6.03 cm^3
Lie Fong *et al.*, 2017	Total: 945 Control: 297 PCOS: 648	Nether-lands/ USA	Academic Medical Center	1990 NIH	N/A	N/A	Control: N/A PCOS: N/A	Control: N/A PCOS: N/A	FNPO	FNPO ≥12.25 (Young) FNPO ≥10.75 (Old)

(continued)

Table 1. Continued

First author, year of publication	Participants' N	Country	Setting	PCOS diagnosis	Ultrasound machine	Transducer frequency (MHz)	Age (y)	BMI (kg/m^2)	Reported markers evaluated	Proposed diagnostic thresholds
Lujan *et al.*, 2013	Total: 168 Control: 70 PCOS: 98	USA/Canada	Control: Community PCOS: Academic Medical Center	1990 NIH	Ultrasonix RP, General Electric Voluson E8	5–9 (Ultrasonix), 6–12 (Voluson)	Control: 27.00 ± 8.89 PCOS: 28.00 ± 5.19	Control: 23.90 ± 4.07 PCOS: 30.10 ± 10.07	FNPO, OV, FNPS	FNPO ≥26 OV ≥10.00 cm^3 FNPS ≥9
Özay *et al.*, 2022	Total: 152 Control: 46 PCOS: 106	Cyprus	Academic Medical Center	Rotterdam	General Electric Voluson 730 Expert	7–9	Control: 22.00 ± 3.25 PCOS: 21.00 ± 3.00	Control: 22.00 ± 3.25 PCOS: 21.00 ± 3.00	ST (Max)	N/A
Vanden Brink *et al.*, 2023	Total: 117 Control: 51 PCOS: 66	USA	Academic Medical Center, Community	1990 NIH	General Electric Voluson	N/A	Control: 28.00 ± 5.40 PCOS: 26.00 + 5.30	Control: 27.10 ± 6.70 PCOS: 31.20 ± 9.30	FNPO, OV, FNPS	FNPO ≥26 OV ≥8.2 cm^3 FNPS ≥8
Villa *et al.*, 2013 [adolescent]	Total: 134 Control: 48 PCOS: 86	Italy	Academic Medical Center	1990 NIH	Biosound Esaote MyLab 25 Gold	3.5–5	Control: 15.3 ± 1.7 PCOS: 15.7 ± 1.4	Control: 23.9 ± 4.9 PCOS: 25.7 ± 5.4	OV	OV ≥5.596 cm^3
Villarroel *et al.*, 2015 [adolescent]	Total: 89 Control: 63 PCOS: 26	Chile	Academic Medical Center	1990 NIH	Medison SonoAce 6000C	5	Control: 16.6 ± 1.5 PCOS: 17.3 ± 1.9	N/A	FNPO, OV	FNPO ≥12 OV ≥10 cm^3
Wongwananuruk *et al.*, 2018	Total: 118 Control: 63 PCOS: 55	Thailand	Academic Medical Center	1990 NIH	Aloga Alpha 6 (Transvaginal or Transrectal)	8	Control: 29.70 ± 7.20 PCOS: 25.10 ± 5.30	Control: 23.50 ± 5.10 PCOS: 25.30 ± 6.30	FNPO (Max), OV (Max), FNPS (Max)	FNPO ≥15 OV ≥6.50 cm^3 FNPS ≥7

PCOS: polycystic ovary syndrome; NIH: National Institutes of Health; AE-PCOS: Androgen Excess & PCOS Society; FNPS: follicle number per cross-section; SA: stromal area; S/A: stromal area/ovarian area; ST: stromal thickness; SV: stromal volume; OA: ovarian area; OUI: ovarian to uterine index; OV: ovarian volume; FNPO: follicle number per ovary.

and Reference Standard (FNPO: 100%; OV: 92%; FNPS: 100%; Stroma: 100%) domains. The only exception was Stroma in which majority had high concerns regarding applicability for Patient Selection (40%). For the Index Test domain, most studies were also given low concerns regarding applicability for FNPO (63%), OV (80%), and Stroma (100%). However, there were high concerns regarding applicability for FNPS (29%). Full details of QUADAS-2 and QUADAS-C scores for each study and index test are presented in Supplementary Table S3 for adults and Supplementary Table S4 for adolescents.

Diagnostic meta-analysis

Five studies (Alsamarai *et al.*, 2009; Köşüş *et al.*, 2011a; Christ *et al.*, 2014; Jarrett *et al.*, 2019; Vanden Brink *et al.*, 2023) used the same study population when evaluating diagnostic accuracy of ultrasonographic ovarian markers and were omitted from the meta-analysis. One study included in the meta-analysis used the same control group when proposing diagnostic thresholds for two PCOS phenotypes (Severe, Mild) based upon the Rotterdam criteria (Köninger *et al.*, 2014). Given the limited number of studies and the variation in features measured, meta-analysis was not conducted on certain ovarian (OA, ovarian contour) and all stromal features. This process excluded two studies (Fulghesu *et al.*, 2001; Özay *et al.*, 2022) that evaluated stromal features exclusively from the meta-analysis. In the end, 8338 unique study participants (4156 control participants and 4182 women with PCOS), including 7442 adult women (3883 control and 3859 PCOS) and 478 adolescent girls (210 control and 268 PCOS), were included as part of the diagnostic meta-analysis.

Diagnostic accuracy of ultrasonographic markers

The sensitivity and specificity forest plots for the ultrasonographic ovarian markers are presented in Fig. 3 and a summary of the pooled diagnostic measures is presented in Table 2. Pooled sensitivity was similar across all ultrasonographic markers whereas pooled specificity was greater for FNPO compared to OV. The improved pooled specificity for FNPO contributed to its higher diagnostic accuracy compared to OV as defined by its AUC value when visualizing all markers on SROC curves (Fig. 4). Although FNPO also had elevated specificity compared to FNPS, the two markers overlapped in their 95% confidence region, contributing to an intermediate AUC value for FNPS (0.870) that was lower than FNPO (0.905) but higher than OV (0.856). Similarly, FNPO had a higher DOR compared to OV and FNPS but exhibited overlap in their 95% confidence regions. FNPO had an elevated LR+ compared to OV and FNPS whereas the LR− was similar across all markers. As judged by the forest plots and SROC curves, FNPO had less within-study and between-study variability in its diagnostic accuracy, respectively, compared to OV and FNPS (Figs 3 and 4). Diagnostic meta-analysis was only possible for OV in adolescent studies due to the limited number of studies available. However, pooled diagnostic measures for OV in adolescents suggest that they may offer similar accuracy as OV in adults (Table 2). Statistical tests evaluating the association between lnDOR and ESS showed no evidence of funnel plot asymmetry for FNPO ($P = 0.37$), FNPS ($P = 0.10$), and OV in adolescents ($P = 0.71$). In contrast, there was significant funnel plot asymmetry for OV in adults ($P = 0.01$) (Supplementary Fig. S1). Given that heterogeneity in test accuracy is expected in diagnostic test accuracy reviews, it is unclear whether the observed funnel plot asymmetry for OV is due to publication bias alone (Macaskill *et al.*, 2010).

Subgroup analysis

Stratified meta-analyses using pre-determined subgroups were conducted to explore potential sources of heterogeneity across studies and are presented in Table 3. For FNPO, studies were stratified based on whether their proposed thresholds were above or below previously recommended diagnostic thresholds, such as the ≥12 follicles threshold (The Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group, 2004) and the ≥20 follicles threshold (Teede *et al.*, 2018a). The proposed threshold of ≥25 follicles previously recommended by the AE-PCOS society (Dewailly *et al.*, 2014b) was not evaluated since only one study (Lujan *et al.*, 2013) met that cutoff. There was no significant threshold effect for FNPO. Similarly, stratification based upon the previously proposed ≥10 cm^3 threshold did not improve diagnostic measures for OV.

Age-specific stratifications separating adult studies with mean PCOS age of ≥30 years old and those with mean age <30 years old did not improve diagnostic measures for both FNPO and OV. Similar results were observed when stratifying age for the control population (data not shown). There was an improvement in sensitivity when stratifying for study populations ≥30 years old for FNPS. However, there was considerable overlap in the 95% confidence region and a limited number of studies between age groups for FNPS (Table 3; Supplementary Fig. S2). From studies that reported the mean age of their PCOS population, the weighted mean age for FNPO, OV, and FNPS were 26.28 ± 1.63, 26.91 ± 3.42, and 27.90 ± 4.69 years old, respectively. Similar weighted means were observed for the age of the control population across markers.

BMI-specific stratifications separating studies with mean PCOS BMI of ≥30 kg/m^2 ('obese') and <30 kg/m^2 ('non-obese') did not improve diagnostic measures in both FNPO, OV, or FNPS. Similar findings were also observed when stratifying for BMI in the control population (data not shown).

Geographic stratifications separated adult studies into three general regions: North America, Europe and Asia. Studies from North America had lower diagnostic accuracy compared to studies from Europe and Asia for both FNPO and OV. Specifically, Asian studies had improved sensitivity for FNPO and European studies had improved specificity, DOR and LR+ for OV compared to North American studies (Table 3; Supplementary Fig. S2). Comparison of the PCOS populations across geographic regions showed that the weighted mean for FNPO was highest in North American studies (27.10 ± 7.28 follicles), intermediate in European studies (22.69 ± 7.49 follicles), and lowest in Asian studies (13.22 ± 4.13 follicles) (Supplementary Fig. S3). In contrast, the weighted mean for OV was similar across all geographic regions (North America: 10.82 ± 1.38 cm^3; Europe: 11.19 ± 2.36 cm^3; Asia: 10.39 ± 2.01 cm^3) (Supplementary Fig. S3). For the control populations, North American studies also had a slightly higher weighted mean FNPO (10.24 ± 1.89 follicles) and OV (6.48 ± 0.61 cm^3) compared to European (FNPO: 7.78 ± 1.98 follicles; OV: 5.71 ± 1.10 cm^3) and Asian (FNPO: 6.94 ± 1.69 follicles; OV: 5.58 ± 0.91 cm^3) studies (Supplementary Fig. S3).

Although age and BMI did not directly account for heterogeneity observed across adult studies, they may also underlie the observed geographic differences in diagnostic accuracy. The weighted mean age in the PCOS population was younger in Asian (25.74 ± 0.61 years old) studies compared to North American studies (28.30 ± 0.93 years old) and European (26.80 ± 2.32 years old) studies for FNPO (Supplementary Fig. S4). In addition, weighted mean age for the control population in European studies was older (31.53 ± 2.38 years old) compared to North American (28.14

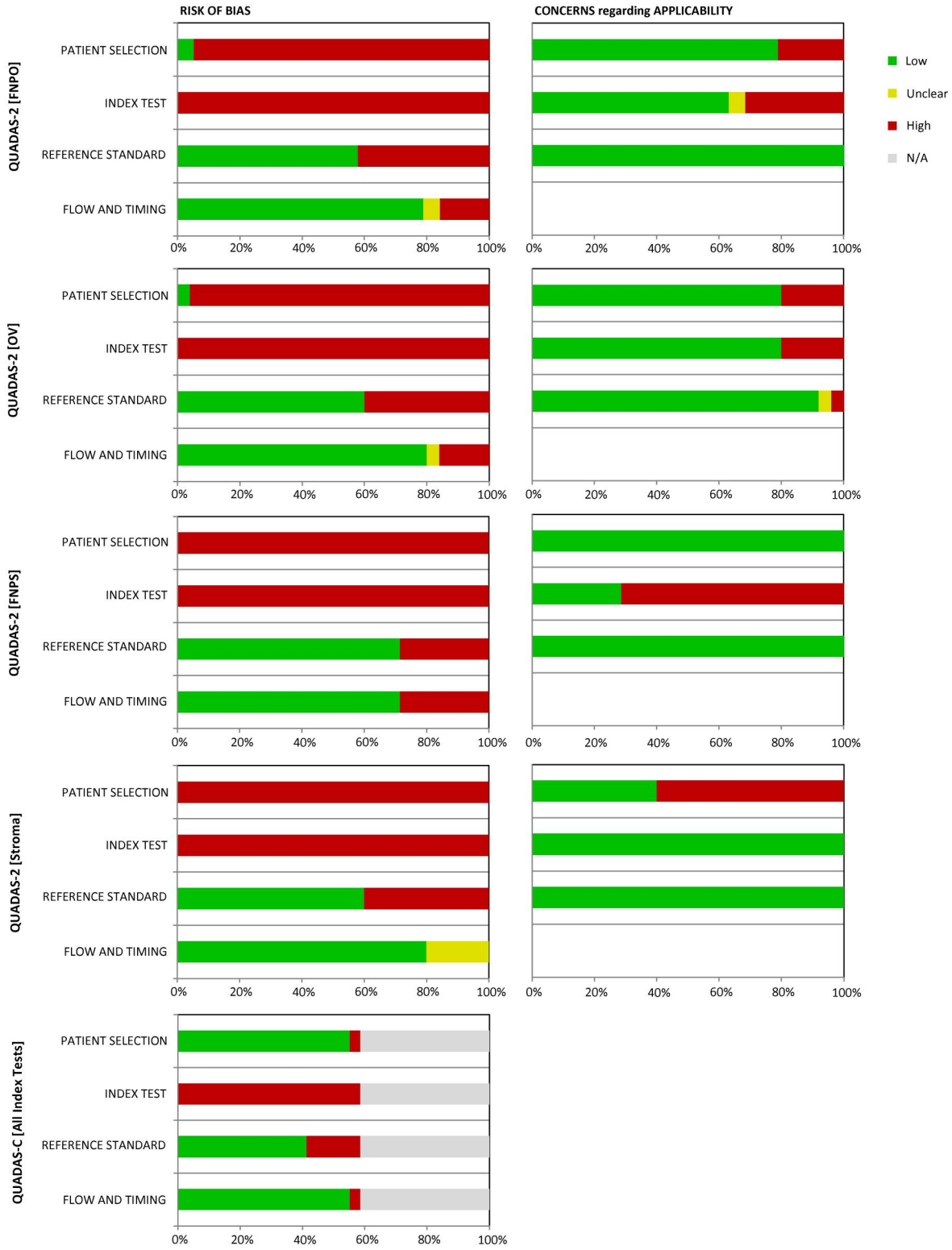


Figure 2. Graphical display of QUADAS-2 and QUADAS-C judgments for risk of bias as percentages across included studies. FNPO: follicle number per ovary; OV: ovarian volume; FNPS: follicle number per cross-section.

± 1.78 years old) and Asian (27.65 ± 1.04 years old) studies for FNPO (Supplementary Fig. S4). For FNPO and OV, the weighted mean BMI categorized the PCOS populations for North America studies as obese (FNPO: 30.29 ± 0.61 kg/m^2; OV: 30.76 ± 1.33 kg/m^2) whereas PCOS populations for studies from Europe (FNPO: 27.62 ± 0.73 kg/m^2; OV: 27.62 ± 1.10 kg/m^2) and Asia (FNPO: 26.54 ± 0.75 kg/m^2; OV: 25.27 ± 2.24 kg/m^2) were categorized as overweight (Supplementary Fig. S5). For the control population, the weighted mean BMI for North American studies (FNPO: 23.93 ± 0.05 kg/m^2; OV: 24.17 ± 0.71 kg/m^2) was also higher than Asian studies (FNPO: 22.36 ± 1.23 kg/m^2; OV: 21.56 ± 1.29 kg/m^2) (Supplementary Fig. S5).

Stratifications based on PCOS diagnostic criteria separated adult studies into either those diagnosed using the 1990 NIH

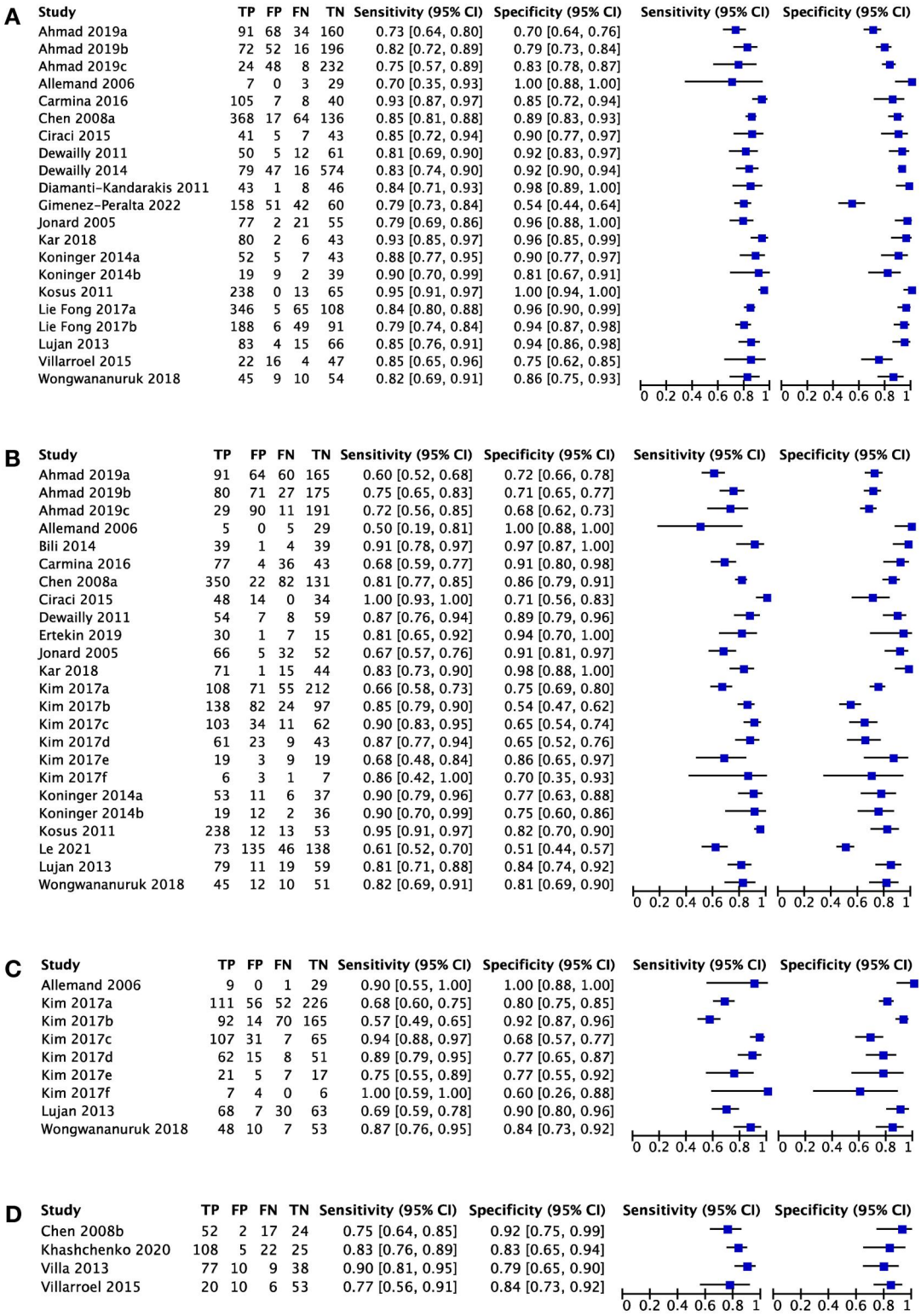


Figure 3. Sensitivity and specificity forest plots for ultrasonographic ovarian markers in the diagnosis of polycystic ovary syndrome (PCOS). (**A**) Follicle number per ovary (FNPO), (**B**) ovarian volume (OV), (**C**) follicle number per cross-section (FNPS), and (**D**) OV (adolescents). TP: true positives; FP: false positives; FN: false negatives; TN: true negatives; CI: confidence interval.

criteria or those diagnosed using the Rotterdam criteria. One study diagnosed PCOS using the AE-PCOS criteria (Köşüş *et al.*, 2011b) and was added to the Rotterdam group. Sensitivity was improved for the Rotterdam group versus the 1990 NIH group in FNPO (Table 3; Supplementary Fig. S2) whereas there was no significant difference in diagnostic accuracy measures between groups for OV. Stratification of PCOS diagnostic criteria was partially aligned by geographic regions. North American studies diagnosed the condition using only the 1990 NIH criteria (100% (5/5)) whereas it was infrequently used in Asian studies (29% (2/7)) in preference for the Rotterdam criteria. Notably, the increased pooled sensitivity observed in the Rotterdam group

Table 2. Summary diagnostic accuracy measures of ultrasonographic markers to detect PCOS.

Index Test	N	Sensitivity % (95% CI)	Specificity % (95% CI)	AUC	DOR (95% CI)	Positive LR (95% CI)	Negative LR (95% CI)
FNPO	16	84.32 (81.27–86.95)	**91.06 (86.44–94.21)***	0.905	54.76 (29.69–101.02)	**9.43 (7.40–12.03)***	0.17 (0.11–0.26)
OV (adult)	16	81.48 (76.05–85.90)	**81.04 (74.66–86.11)***	0.856	18.80 (11.62–30.41)	**4.30 (3.28–5.63)***	0.23 (0.17–0.31)
OV (adolescent)	4	81.84 (75.74–86.68)	83.54 (77.06–88.46)	0.892	22.87 (13.16–39.74)	**4.97 (3.62–6.82)***	0.22 (0.15–0.31)
FNPS	4	81.07 (70.10–88.67)	82.70 (75.15–88.31)	0.870	20.47 (12.69–33.02)	**4.69 (2.99–7.35)***	0.23 (0.17–0.31)

FNPO: follicle number per ovary; OV: ovarian volume; FNPS: follicle number per cross-section; DOR: diagnostic odds ratio; LR: likelihood ratio.
* Significant difference between markers (bolded).

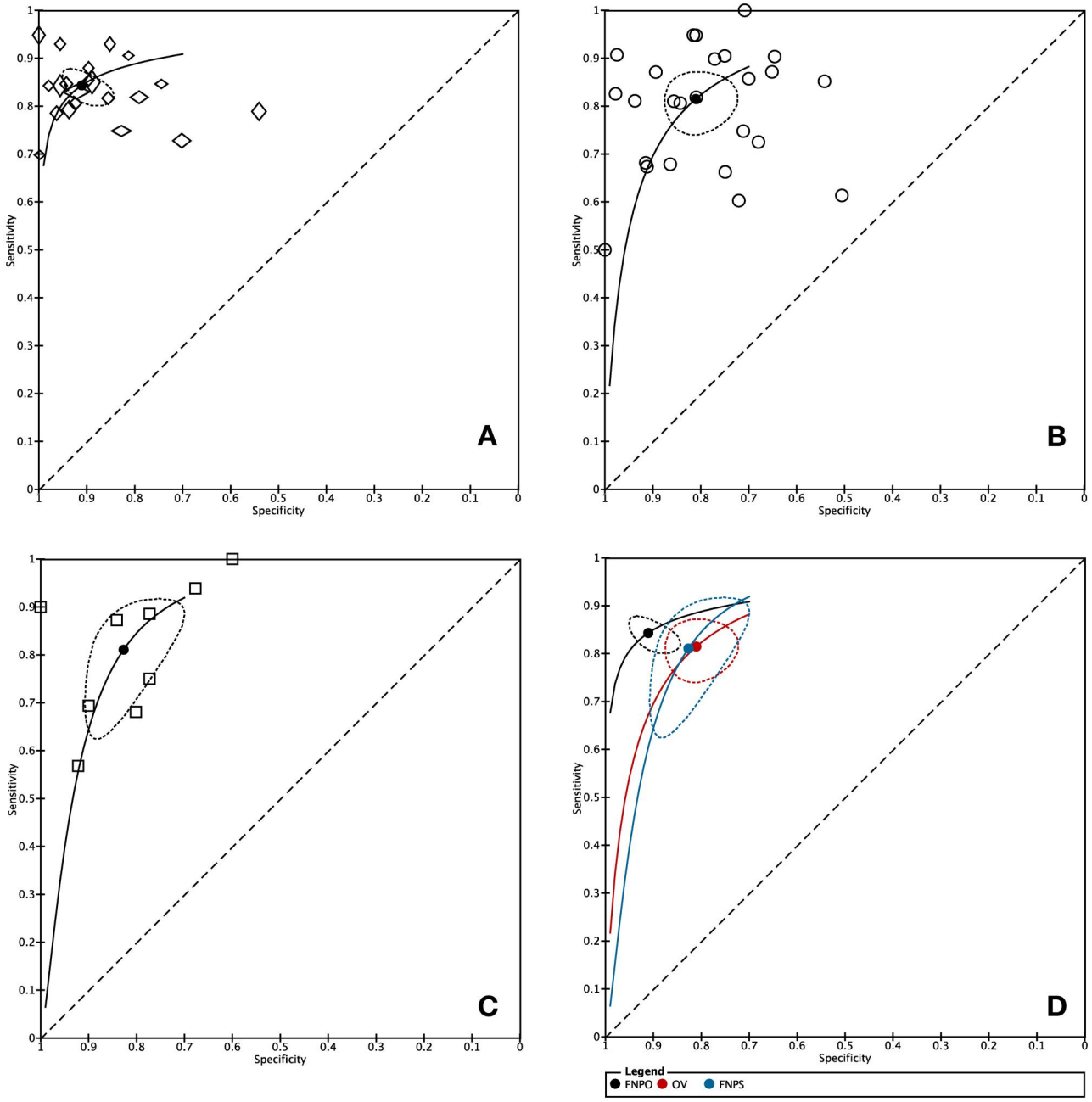


Figure 4. Summary receiver-operating characteristic (SROC) curve for ultrasonographic ovarian markers in the diagnosis of polycystic ovary syndrome (PCOS) in adult women. (**A**) Follicle number per ovary (FNPO), (**B**) ovarian volume (OV), (**C**) follicle number per cross-section (FNPS), and (**D**) all ovarian markers. Each symbol represents a single study. The black dot represents the summary point and the dotted region represents the 95% confidence region. The diagonal dotted line represents AUC = 0.50 (random chance).

versus 1990 NIH group was similar to results comparing Asian versus North American studies. European studies applied both PCOS diagnostic criteria equally (44% (4/9)).

Ultrasound transducer frequency was stratified based on adult studies that used either ≥8 MHz or <8 MHz, as ≥8 MHz is the recommended transducer frequency to maximal resolution

Table 3. Subgroup analysis of diagnostic accuracy of ultrasonographic markers to detect PCOS.

	Index test	N	Sensitivity % (95% CI)	Specificity % (95% CI)	DOR (95% CI)	Positive LR (95% CI)	Negative LR (95% CI)
Follicle number per ovary (FNPO)							
Cutoff value	≥20 follicles	3	87.64 (78.78–93.12)	93.74 (82.40–97.96)	106.18 (28.06–401.70)	14.00 (7.51–26.11)	0.13 (0.05–0.37)
	<20 follicles	13	83.54 (80.29–86.35)	90.71 (85.08–94.35)	49.54 (24.48–100.24)	8.99 (6.88–11.74)	0.18 (0.11–0.29)
	≥12 follicles	12	82.98 (79.85–85.71)	90.20 (84.67–93.87)	44.85 (23.72–84.81)	8.46 (6.65–10.77)	0.19 (0.12–0.29)
	<12 follicles	4	87.18 (80.01–92.03)	93.04 (83.34–97.28)	90.89 (23.57–350.50)	12.53 (7.07–22.20)	0.14 (0.06–0.33)
Age groups	≥30 years old	4	79.29 (74.84–83.13)	89.93 (76.86–96.00)	34.18 (12.32–94.85)	7.87 (5.80–10.69)	0.23 (0.11–0.50)
	<30 years old	14	85.80 (82.38–88.65)	92.08 (88.03–94.84)	70.29 (37.54–131.64)	10.84 (8.23–14.26)	0.15 (0.10–0.23)
BMI groups	≥30 kg/m^2	2	83.33 (75.09–89.24)	95.96 (89.73–98.48)	118.75 (38.70–364.36)	20.63 (12.34–34.46)	0.17 (0.08–0.40)
	<30 kg/m^2	9	87.78 (83.27–91.21)	92.92 (88.48–95.73)	94.32 (46.75–190.30)	12.40 (8.66–17.75)	0.13 (0.08–0.21)
Geography	North America	3	**78.63 (72.91–83.41)***	87.24 (73.06–94.52)	25.17 (9.48–66.83)	6.16 (4.40–8.63)	0.24 (0.12–0.51)
	Europe	8	83.14 (79.46–86.27)	90.96 (84.73–94.80)	49.61 (25.23–97.55)	9.19 (7.12–11.88)	0.19 (0.11–0.31)
	Asia	5	**89.33 (83.70–93.18)***	93.05 (85.44–96.83)	112.13 (42.96–292.67)	12.86 (8.08–20.44)	0.11 (0.05–0.24)
Diagnostic criteria	1990 NIH	10	**81.76 (79.23–84.05)***	91.06 (86.40–94.24)	45.69 (26.17–79.76)	9.15 (7.51–11.16)	0.20 (0.14–0.30)
	Rotterdam	6	**89.81 (84.11–93.62)***	89.79 (77.46–95.75)	77.55 (26.62–225.86)	8.80 (5.51–14.05)	0.11 (0.05–0.26)
Transducer frequency	≥8 MHz	9	84.27 (80.07–87.72)	86.42 (79.62–91.20)	34.09 (16.95–68.55)	6.21 (4.59–8.38)	0.18 (0.12–0.28)
	<8 MHz	5	86.94 (80.48–91.49)	94.13 (87.07–97.45)	106.83 (31.68–360.20)	14.82 (8.75–25.08)	0.14 (0.06–0.30)
Methodology	2D real time	14	84.30 (80.97–87.13)	90.24 (85.19–93.70)	49.63 (26.30–93.67)	**8.64 (6.67–11.19)***	0.17 (0.12 0.26)
	Offline	2	83.33 (75.09–89.24)	95.96 (89.73–98.48)	118.75 (38.70–364.36)	**20.63 (12.34–34.46)***	0.17 (0.08–0.40)
Risk of bias	High risk	9	86.52 (82.03–90.03)	91.71 (85.33–95.47)	71.04 (30.07–167.83)	10.44 (7.27–15.00)	0.15 (0.08–0.26)
	Low risk	7	81.75 (77.80–85.14)	89.74 (82.09–94.34)	39.18 (18.20–84.36)	7.97 (6.00–10.58)	0.20 (0.12–0.35)
	Index test	**N**	**Sensitivity % (95% CI)**	**Specificity % (95% CI)**	**DOR (95% CI)**	**Positive LR (95% CI)**	**Negative LR (95% CI)**
Ovarian volume (OV)							
Cutoff value	≥10 cm^3	5	82.56 (68.83–91.03)	78.95 (67.26–87.25)	17.75 (8.65–36.44)	3.92 (2.24–6.87)	0.22 (0.14–0.34)
	<10 cm^3	12	80.74 (74.54–85.71)	82.01 (74.01–87.95)	19.11 (10.01–36.46)	4.49 (3.24–6.21)	0.23 (0.16–0.35)
Age groups	≥30 years old	4	75.58 (64.62–83.98)	73.31 (59.19–83.88)	8.50 (4.26–16.96)	2.83 (1.96–4.10)	0.33 (0.21–0.53)
	<30 years old	14	83.63 (77.19–88.52)	83.77 (77.40–88.61)	26.36 (15.11–45.99)	5.15 (3.65–7.28)	0.20 (0.14–0.28)
BMI groups	≥30 kg/m^2	3	84.78 (79.75–88.73)	77.82 (54.55–91.11)	19.53 (6.33–60.31)	3.82 (2.79–5.24)	0.20 (0.08–0.49)
	<30 kg/m^2	11	81.91 (74.32–87.63)	85.78 (77.39–91.40)	27.31 (12.33–60.52)	5.76 (3.79–8.75)	0.21 (0.13–0.34)
Geography	North America	4	75.84 (67.50–82.59)	**72.98 (65.03–79.69)***	**8.48 (5.82–12.35)***	**2.81 (2.16–3.65)***	0.33 (0.26–0.42)
	Europe	5	83.30 (72.83–90.27)	**87.96 (80.58–92.79)***	**36.44 (19.59–67.77)***	**6.92 (4.16–11.52)***	0.19 (0.12–0.30)
	Asia	7	86.53 (74.74–93.31)	82.68 (69.05–91.08)	30.67 (9.95–94.59)	5.00 (2.59–9.64)	0.16 (0.08–0.32)
Diagnostic criteria	1990 NIH	8	77.48 (71.67–82.39)	78.41 (71.18–84.22)	12.50 (8.33–18.74)	3.59 (2.86–4.51)	0.29 (0.22–0.38)
	Rotterdam	8	87.61 (77.62–93.52)	85.45 (72.88–92.77)	41.54 (14.59–118.26)	6.02 (3.26–11.13)	0.14 (0.07–0.29)
Transducer frequency	≥8 MHz	9	79.17 (73.65–83.79)	79.38 (72.08–85.16)	14.63 (9.39–22.80)	3.84 (3.02–4.87)	0.26 (0.19–0.36)
	<8 MHz	7	86.70 (72.76–94.09)	84.54 (70.90–92.47)	35.66 (10.34–122.97)	5.61 (2.60–12.09330)	0.16 (0.08–0.32)
Methodology	2D real time	11	82.53 (75.27–87.99)	80.11 (71.80–86.43)	19.02 (11.03–32.82)	4.15 (2.95–5.83)	0.22 (0.15–0.31)
	Other	5	78.99 (69.64–86.04)	83.12 (72.41–90.24)	18.52 (6.53–52.51)	4.68 (2.80–7.82)	0.25 (0.14–0.44)
Risk of bias	High risk	11	83.16 (76.88–88.00)	78.17 (70.70–84.16)	17.68 (9.83–31.81)	3.81 (2.76–5.26)	0.22 (0.15–0.31)
	Low risk	5	76.88 (70.06–82.53)	86.56 (82.70–89.66)	21.41 (13.51–33.95)	5.72 (4.19–7.81)	0.27 (0.21–0.34)
Adolescent studies	Total	4	81.84 (75.74–86.68)	83.54 (77.06–88.46)	22.87 (13.16–39.74)	4.97 (3.62–6.82)	0.22 (0.15–0.31)
	TAUS	3	83.77 (77.90–88.31)	81.88 (74.56–87.45)	23.32 (13.10–41.54)	4.62 (3.36–6.37)	0.20 (0.14–0.29)
	Index test	**N**	**Sensitivity % (95% CI)**	**Specificity % (95% CI)**	**DOR (95% CI)**	**Positive LR (95% CI)**	**Negative LR (95% CI)**
Follicle number per cross-section (FNPS)							
Age groups	≥30 years old	2	**89.54 (81.83–94.21)***	79.98 (62.36–90.59)	34.21 (11.50–101.71)	4.47 (2.65–7.54)	0.13 (0.06–0.29)
	<30 years old	3	**70.14 (58.51–79.64)***	86.58 (79.71–91.37)	15.15 (8.87–25.87)	5.22 (3.50–7.80)	0.34 (0.25–0.48)
BMI groups	≥30 kg/m^2	3	82.37 (65.78–91.91)	88.00 (74.57–94.83)	34.29 (9.57–122.88)	6.87 (3.09–15.25)	0.20 (0.09–0.43)
	<30 kg/m^2	2	77.80 (60.74–88.81)	80.87 (76.37–84.68)	14.81 (6.27–35.01)	4.07 (2.09–7.90)	0.27 (0.21–0.36)

(continued)

Table 3. Continued

	Index test	N	Sensitivity % (95% CI)	Specificity % (95% CI)	DOR (95% CI)	Positive LR (95% CI)	Negative LR (95% CI)
Methodology	2D real time	2	83.17 (70.52–91.08)	79.27 (70.99–85.66)	18.89 (11.33–31.49)	4.01 (2.38–6.77)	0.21 (0.16–0.28)
	Offline	2	71.30 (62.08–79.03)	95.10 (65.16–99.51)	48.17 (4.48–518.13)	14.54 (6.67–31.70)	0.30 (0.06–1.60)
Risk of bias	Total	4	82.70 (75.15–88.31)	81.07 (70.10–88.67)	20.47 (12.69–33.02)	4.37 (2.80–6.81)	0.21 (0.15–0.29)
	Low risk	3	80.55 (65.70–89.95)	90.92 (78.95–96.40)	41.49 (11.90–144.68)	8.88 (4.29–18.36)	0.21 (0.10–0.47)

DOR: diagnostic odds ratio; LR: likelihood ratio.
* Significant difference between stratified groups (bolded).

of antral follicles (Dewailly et al., 2014b; Teede et al., 2018a). However, there were no differences in diagnostic accuracy due to transducer frequency for both FNPO and OV. Test methodology separated studies that used conventional real time approaches versus offline analysis. Offline analysis for FNPO improved the LR+ compared to real time approaches, however, there are only a limited number of studies using offline methods (Table 3). In addition, stratification by methodology also did not improve diagnostic accuracy for OV. For adolescent studies, one study (Chen et al., 2008b) used both transabdominal and transrectal ultrasonography to evaluate ovarian morphology. Subgroup analysis for studies that exclusively used transabdominal ultrasonography did not change the diagnostic accuracy of OV.

Given that almost all studies had high risk of bias for the Patient Selection and Index Test domains of QUADAS-2, risk of bias was stratified with studies at 'High Risk' if they also had a high risk of bias grade for the Reference Standard and/or Flow and Timing domains. Exclusion of 'High Risk' studies did not improve diagnostic accuracy for FNPO, OV, or FNPS.

Discussion

Main findings

The present systematic review and diagnostic meta-analysis of 20 studies in adult women, comprising of 3883 control participants and 3859 women with PCOS, indicate that FNPO is the most accurate marker for the diagnosis of PCOS on ultrasonography. OV and FNPS have comparable diagnostic accuracy for PCOS yet show poorer performance compared to FNPO. As such, OV and FNPS should be used as alternative diagnostic markers when accurate measurement of FNPO is not possible. In addition, the systematic review identified four adolescent studies consisting of 210 control participants and 268 girls with PCOS. Pooled diagnostic measures suggest that OV in adolescents may offer comparable accuracy to OV in adults. However, the currently available data are limited and lack the ability to conduct diagnostic meta-analysis on other ovarian markers, such as follicle counts. Subgroup analysis suggests that stratification based on previously proposed diagnostic thresholds, age, BMI, and transducer frequency did not improve diagnostic accuracy for FNPO and OV. However, diagnostic accuracy for FNPO improved when stratified for studies diagnosing PCOS using the Rotterdam criteria versus the 1990 NIH criteria or when stratified for studies using offline follicle counting versus real time methods. There were substantial differences in diagnostic accuracy between geographic regions, with North American studies having poorer diagnostic accuracy compared to Asian and European studies for FNPO and OV, respectively. Comparisons of study characteristics across geographic regions suggest that differences in age, BMI, and diagnostic criteria may indirectly underlie the observed differences in diagnostic accuracy. Overall, these observations highlight the utility of various ultrasonographic markers of PCOM in the diagnosis of PCOS. Nevertheless, it should be noted that most included studies were at high risk of bias owing primarily to the non-randomized, observational study design, and lack of comparisons of diagnostic accuracy with previously proposed thresholds. Standardization and refinement in the conduct, assessment and comparison of PCOM is critical in facilitating a more accurate evaluation of PCOS and investigating phenotypic variations in the pathogenesis of the condition.

Strengths and limitations

Strengths of this study included a comprehensive database search and the use of recommended statistical approaches for evaluating diagnostic test accuracy studies. In addition, we employed analyses on various study-level variables known to influence ovarian ultrasound markers and their diagnostic accuracy to identify potential sources of heterogeneity that could be leveraged for future research or guide improvement in current practice. Risk of bias and concerns regarding applicability were evaluated using a quality assessment tool specifically designed for diagnostic test accuracy studies. A primary limitation of this diagnostic meta-analysis was the lack of proposed summary thresholds for any ultrasonographic marker. The inability to determine a common threshold to define PCOM was primarily due to the heterogeneity across studies and limited data comparing previously recommended thresholds and those proposed within studies. In addition, we were unable to conduct a diagnostic meta-analysis for stromal features given the limited number of studies and variety of features proposed. Our subgroup meta-analysis integrated aggregated data made available publicly or upon request by investigators but remained limited in its inability to comprehensively evaluate the impact of confounders on diagnostic accuracy. Our search strategy included English-only studies and did not include any grey literature databases. Therefore, relevant diagnostic test accuracy studies written in other languages or non-peer reviewed data may have been omitted from the list of included studies.

Comparison with existing literature

Our observations affirm the recommendations set by the 2018 International Guideline (Teede et al., 2018a) and the 2014 AE-PCOS Task Force Report (Dewailly et al., 2014b) for ultrasonographic ovarian features in the context of PCOS diagnosis in adult women. Our findings support the use of FNPO as the gold standard ovarian marker for the diagnosis of PCOS. Furthermore, we corroborated previous recommendations of OV as a robust alternative to FNPO. Current evidence attributes PCOM to an excess of antral follicles, driven primarily by the accumulation of small 2–5 mm follicles (Jonard et al., 2003; Webber et al., 2003; Maciel et al., 2004; Christ et al., 2015). Although enlarged OV has been widely observed in women with PCOS (Balen et al., 2003; Carmina et al., 2005; de Guevara et al., 2013), OV has not been shown to consistently reflect the severity of reproductive dysfunction in PCOS (Christ et al., 2015) which aligns with its reduced discriminatory power. However, OV still offers strong diagnostic potential in the context of varying technology across routine practice and in situations when poor ultrasound image quality prevents accurate follicle counting (Dewailly et al., 2014b). In addition, we found that FNPS offers similar diagnostic accuracy compared to OV. FNPS was not previous recommended in the PCOM diagnostic criterion (Dewailly et al., 2014b; Teede et al., 2018a) but our findings show promise of using FNPS as an alternative to detect follicle excess when counting across the entire ovary is unavailable. Both OV and FNPS offer similar or better inter-rater reliability compared to FNPO, highlighting their value as alternative markers (Lujan et al., 2008). That said, the limited number of studies evaluating the diagnostic utility of FNPS indicate a need for further research. Our findings also suggest that OV may potentially be a robust ovarian marker for PCOS diagnosis in adolescent girls. However, current evidence remains limited owing to variations in diagnostic criteria (1990 NIH, Rotterdam) and ultrasonography methods (transabdominal, transrectal) used across the studies. In addition, the 2018 International Guideline does not recommend the use of ovarian ultrasonography in individuals <8 years post-menarche owing to the high incidence of PCOM and increasing ovarian size during this life stage (Teede et al., 2018a; Peña et al., 2020). With the absence of large longitudinal

studies to validate normative ranges of ovarian development during the adolescent transition, it may be premature to propose the use of ultrasonography in evaluation of adolescents with suspicion for PCOS.

The use of a transducer frequency ≥8 MHz has been previously recommended to ensure maximal detection of 2–9 mm antral follicles (Dewailly *et al.*, 2014b; Teede *et al.*, 2018a). We had anticipated that studies whose transducer frequency included ≥8 MHz would have improved diagnostic accuracy for FNPO. However, our data found that diagnostic accuracy for FNPO was unaffected after stratification for presumably older (<8 MHz) versus newer (≥8 MHz) imaging technology. This discrepancy may be due to the lack of reliability of conventional real-time counting methods between observers. Previous studies have shown that counting the high number of follicles in polycystic ovaries in real time has moderate to poor reliability between raters (Lujan *et al.*, 2008, 2009) and can significantly misclassify ovaries as having PCOM (Vanden Brink *et al.*, 2021) even when higher resolution transducers are used. As such, detection of more follicles with newer imaging technology may not necessarily yield a more accurate analysis of PCOS and standardized methodology in follicle counting is required for greater reliability across clinicians and researchers.

Modeling from cross-sectional and longitudinal follow up studies has shown that the prevalence of PCOM and measurements of follicle counts and ovarian size decline with age in women with PCOS (Hudecova *et al.*, 2009; Glintborg *et al.*, 2012; Wiser *et al.*, 2013; Ahmad *et al.*, 2018; Jacewicz-Święcka *et al.*, 2021; van Keizerswaard *et al.*, 2022). As such, development of age-specific diagnostic thresholds for PCOM in women of reproductive age have been recommended by the 2018 International Guideline (Teede *et al.*, 2018a) with several having been proposed to date (Kim *et al.*, 2017; Lie Fong *et al.*, 2017; Ahmad *et al.*, 2019). However, our diagnostic meta-analysis found that age did not play a substantial role in the heterogeneity of diagnostic accuracy across studies for FNPO and OV when studies were stratified into age groups of ≥30 or <30 years old. It should be noted that the weighted mean age for the PCOS and control populations across all studies for FNPO (PCOS: 26.30 ± 1.68 years old; Control: 30.43 ± 2.82 years old), OV (PCOS: 27.06 ± 3.44 years old; Control: 28.86 ± 4.47 years old), and FNPS (PCOS: 27.90 ± 4.69 years old; Control: 27.20 ± 4.80 years old) were relatively young, and few studies were available to capture individuals ≥30 years old. Consequently, this limited range of age across studies likely accounted for the failure to detect an impact of chronological age on diagnostic accuracy.

Similarly, we also found that stratification based on BMI did not directly account for any heterogeneity in diagnostic accuracy across ultrasonographic markers. Previous studies on healthy women with obesity indicate that FNPO did not differ when compared to their leaner counterparts (Roth *et al.*, 2014; Moslehi *et al.*, 2018) despite evidence of a negative association between BMI and ovarian morphology markers (Moslehi *et al.*, 2018; Peigné *et al.*, 2018; Kazemi *et al.*, 2020; Neubronner *et al.*, 2021). The relationship between adiposity and FNPO in the context of PCOS is more controversial, with studies indicating a significant association but uncertainty as to its directionality and specificity in follicle subpopulations (Jonard *et al.*, 2003; Christ *et al.*, 2015; Moslehi *et al.*, 2018; Peigné *et al.*, 2018; Neubronner *et al.*, 2021). However, our observations support previous evidence that OV is not influenced by BMI in either healthy women (Malhotra *et al.*, 2013; Neubronner *et al.*, 2021) or women with PCOS (Christ *et al.*, 2015; Neubronner *et al.*, 2021).

We found that geographic differences may underlie some of the heterogeneity in diagnostic accuracy of ultrasonographic markers observed across studies. We noted variation in the manifestation of PCOM across regions that could potentially impact discriminatory power. In the case of FNPO, we noted higher FNPO in North American studies relative to European and Asian studies suggestive of greater potential for follicle excess in this geographic region. Our findings of higher follicle count in North American PCOS populations versus European populations aligns with previously reported differences in FNPS between Caucasian women with PCOS from Boston compared with those from Iceland (Welt *et al.*, 2006). Likewise, our finding of lower FNPO in Asian PCOS populations corroborate other reports from East Asia (Lee *et al.*, 2015; Han *et al.*, 2017) as well as comparisons of FNPS between North American-based Asian populations and other ethnicities (Welt *et al.*, 2006). In contrast, our data showed no differences in OV across geographic regions and differs from previously reported findings (Welt *et al.*, 2006; Lee *et al.*, 2015; Han *et al.*, 2017). It should be noted that the observed geographic differences in diagnostic accuracy may be partially due to differences in the PCOS diagnostic criteria. North American studies exclusively used the 1990 NIH criteria whereas the Asian studies preferably used the Rotterdam criteria. We found that the increased sensitivity in Asian versus North American studies aligned with similar results observed when comparing Rotterdam versus 1990 NIH subgroups for FNPO. However, the increased diagnostic accuracy measures in European versus North American studies for OV did not share the same pattern.

Age and BMI may have also contributed to the observed differences in diagnostic accuracy between geographic groups. We noted that the PCOS population in Asian studies was younger and leaner than those in North American studies when comparing FNPO. This observation aligns with previous comparisons between Caucasian and Asian women with PCOS (Welt *et al.*, 2006; Chan *et al.*, 2017). Furthermore, one study that matched for age and BMI to compare ethnicity alone between North American Caucasian and Asian women reported no differences in prevalence of PCOM (Wang *et al.*, 2013). These findings suggest that geographic variation, which may include a variety of potential factors, contributes to differences in ovarian morphology and, by extension, diagnostic accuracy in PCOM.

Implications for clinical practice

Our findings reinforce the use of FNPO as the superior ultrasonographic marker in the diagnosis of PCOS (Teede *et al.*, 2018b). However, there remains room for improvement regarding standardization of best practices in obtaining FNPO to ensure that accuracy is maintained across clinicians and researchers, settings and technology. We also found OV and FNPS to be alternatives with good diagnostic performance if total follicle counts cannot be accurately evaluated (e.g. in the case of poor image quality or low transducer frequency). In particular, FNPS offers promising utility as a proxy for FNPO that can be obtained quickly using a single slice of the ovary; however, further evidence is required to validate these findings. Likewise, there may be potential to establish a PCOM criteria for adolescents as more data becomes available. Despite the clinical heterogeneity and limited quality across studies, the sensitivity and specificity of the observed ovarian ultrasound markers remained generally consistent and highlights their utility for accurately capturing PCOS.

Implications for future research

Given the heterogeneity across studies, standardization in the evaluation of ultrasonographic ovarian features is required to

ensure a more accurate diagnosis of PCOS. Although real time follicle counting remains the conventional method, it has been shown to exhibit moderate to poor inter-rater reliability across clinicians when evaluating polycystic ovaries (Lujan et al., 2008, 2009). Indeed, our subgroup analysis suggests that alternative approaches, such as offline analysis, may have a greater likelihood of detecting PCOS although there remains a limited number of studies. Several approaches in follicle counting have been explored toward improving standardization and reproducibility, including formal training workshops across clinicians (Lujan et al., 2008) or the use of a grid overlay in either offline (Lujan et al., 2010) or real time follicle counting (Vanden Brink et al., 2021). With the advent of new ultrasound technology, 3D follicle counting has also been strongly considered, albeit few data are available in the context of PCOS diagnosis (Allemand et al., 2006; Battaglia et al., 2012; Kar and Swoyam, 2018). Compared to 2D approaches, inter-observer reliability is improved when using 3D follicle counting methods such as multiplanar view (MPV) allowing for simultaneous visualization of follicles across three perpendicular axes (Mercé et al., 2005; Jayaprakasan et al., 2007, 2008; Deb et al., 2009) or automated 3D volume reconstruction software (e.g. SonoAVC^{TM}) (Jayaprakasan et al., 2008; Deb et al., 2009). Real time and offline MPV may offer a clinically feasible method as they provide similar FNPO for both PCOM and non-PCOM ovaries compared to gold standard approaches (Vanden Brink et al., 2021). In contrast, semi-automated follicle count software, such as SonoAVC^{TM}, remains a challenge given its requirement of post-automation processing and systematic undercounting of follicles (Deb et al., 2009; Vanden Brink et al., 2021). Future studies should determine standardized approaches in 3D follicle counting of polycystic ovaries to evaluate their inter-observer reliability and subsequent diagnostic accuracy compared with conventional 2D methods.

Although we found that stratification based on age and BMI did not account for heterogeneity in diagnostic accuracy across markers, future research should identify opportunities to investigate diagnostic test accuracy of ovarian ultrasound markers in a large population of women with PCOS across the age and adiposity spectrum. Thus far, studies that propose age-specific diagnostic thresholds for PCOM are based on a small number of women older than 35 years in their study population (Alsamarai et al., 2009; Kim et al., 2017; Lie Fong et al., 2017; Ahmad et al., 2019). In addition, future research should evaluate how ultrasonographic ovarian markers differ across the BMI spectrum and their implications on the diagnosis of PCOS via BMI-stratified thresholds. Thus far, adjustment for BMI in the detection in PCOS has only been evaluated with anti-Müllerian hormone (AMH) (Palomaki et al., 2020) and has not been assessed with ultrasonographic markers of PCOM. Although several studies have compared prevalence of PCOM across different ethnicities and geographic regions (Glintborg et al., 2010; Wang et al., 2013; Chan et al., 2017), only one study has directly compared ultrasonographic ovarian markers (Welt et al., 2006). Further research should prospectively conduct comparisons of age- and BMI-matched PCOS populations across different geographical regions to evaluate differences in ovarian morphology. Specifically, integration of raw data across multiple clinical and research sites around the world would allow for individual patient data meta-analysis that can directly inform the impact of underlying confounders and their independent associations to diagnostic accuracy.

Ultrasonographic ovarian measurements were presented either as the maximum value or average value between the two ovaries. It has been shown that there is little variation in left-right differences in FNPO (Jarrett et al., 2019) and either would be a concordant marker if only one ovary is visible. However, there are significant left-right differences in OV and FNPS leading to ovary-specific diagnostic thresholds (Jarrett et al., 2019). Consensus is therefore required across researchers and clinical settings on ways to present ovarian ultrasound data to improve standardization and reliability. In addition, nomenclature of ultrasonographic markers should be addressed. Although FNPO is broadly accepted to denote follicle counts across the entire ovary, antral follicle count (AFC) has been used as an equivalent term. However, AFC can also indicate the total sum of follicles in both ovaries and is commonly used in the context of reproductive medicine, including for measurement of the ovarian reserve (Rosen et al., 2012), in vitro fertilization outcomes (Jayaprakasan et al., 2012) and predicted response to ovarian stimulation (Nastri et al., 2015). Consensus is required as to which term should be used in the context of PCOS diagnosis and evaluation.

Most of the included studies were graded high risk of bias due in their patient selection and index test methodology. In particular, all studies were observational in design and only one used consecutive sampling (Dewailly et al., 2011). Further evaluations in diagnostic test accuracy should utilize random sampling from a larger study population or consecutive sampling across clinical or research settings. Raters interpreting ultrasound ovarian markers should be blinded of the participant's phenotype. Our findings indicated that including PCOM as part of the Rotterdam criteria could exaggerate the diagnostic accuracy of FNPO. However, using the 1990 NIH criteria to diagnose PCOS only captures those with the most severe manifestation of PCOS in terms of reproductive and metabolic dysfunction (Diamanti-Kandarakis and Panidis, 2007; Kauffman et al., 2008; Clark et al., 2014). Therefore, ultrasound image analysis blinded and conducted prior to phenotyping is critical in reducing bias associated with the inclusion of PCOM in the PCOS diagnosis of the study population. Lastly, the presence of varying thresholds across all studies provided additional heterogeneity in determining diagnostic accuracy (threshold effect) and subsequently prevented a determination of the most accurate cutoff to diagnose PCOS. Future studies should report diagnostic accuracy measures across multiple thresholds, including previously recommended cutoffs, and pool individual patient data across large datasets to facilitate more robust diagnostic meta-analysis.

Conclusion

This systematic review and diagnostic meta-analysis confirms that FNPO is the most accurate ultrasonographic marker in the diagnosis of PCOS in adults, with OV and FNPS as alternatives when accurate total follicle counts are not possible. Standardization of best practices in follicle counting is required to ensure accurate measurements across users, settings, and technology. Although stratification based on age and BMI did not substantially account for the heterogeneity observed in diagnostic accuracy across markers, geographic variation across studies may influence differences in diagnostic accuracy in FNPO and OV. Weaknesses in study designs, such as patient selection and index test methodology, limit the strength of the evidence, and conclusions. In addition, more data are needed to support any ultrasonographic criterion for PCOS in adolescents. However, our findings serve as a foundation for well-designed studies toward a standard definition of PCOM and will inform future guideline recommendations. These efforts are essential for a more accurate evaluation of PCOS and for future investigations of the variable

pathogenic mechanisms that underlie phenotypic differences in this condition.

Supplementary data

Supplementary data are available at *Human Reproduction Update* online.

Data availability

The data underlying this article are available in the article and its online supplementary material.

Acknowledgements

None.

Authors' roles

J.P. conceived and designed the study. J.P., J.B., C.W., and K.G. conducted systematic database searches. J.P., A.L.O., and C.W. performed the risk of bias and applicability assessment. J.P., J.B., F.E.C., C.W., and K.G. extracted data and J.P., J.B., A.L.O., F.E.C., L. M.J., and M.E.L. reviewed the collated data. J.P. conducted the statistical analysis with contributions from L.M.J. J.P., L.M.J., and M. E.L. contributed to the data interpretation. J.P. and M.E.L. wrote the manuscript and J.B., A.L.O., F.E.C., C.W., K.G. and L.M.J. edited the manuscript.

Funding

This study was supported through the National Institutes of Health (R01HD093748).

Conflict of interest

The authors have no conflict of interest to disclose.

References

Ahmad AK, Kao CN, Quinn M, Lenhart N, Rosen M, Cedars MI, Huddleston H. Differential rate in decline in ovarian reserve markers in women with polycystic ovary syndrome compared with control subjects: results of a longitudinal study. *Fertil Steril* 2018;**109**:526–531.

Ahmad AK, Quinn M, Kao CN, Greenwood E, Cedars MI, Huddleston HG. Improved diagnostic performance for the diagnosis of polycystic ovary syndrome using age-stratified criteria. *Fertil Steril* 2019;**111**:787–793.e2.

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; **94**:4961–4970.

Azziz R, Carmina E, Chen Z, Dunaif A, Laven JSE, Legro RS, Lizneva D, Natterson-Horowtiz B, Teede HJ, Yildiz BO. Polycystic ovary syndrome. *Nat Rev Dis Primers* 2016;**2**:16057–16018.

Azziz R, Carmina E, Dewailly D, Diamanti-Kandarakis E, Escobar-Morreale HF, Futterweit W, Janssen OE, Legro RS, Norman RJ, Taylor AE *et al.*; Androgen Excess Society. Criteria for defining polycystic ovary syndrome as a predominantly hyperandrogenic syndrome: an androgen excess society guideline. *J Clin Endocrinol Metab* 2006;**91**:4237–4245.

Balen AH, Laven JSE, Tan S-L, Dewailly D. Ultrasound assessment of the polycystic ovary: international consensus definitions. *Hum Reprod Update* 2003;**9**:505–514.

Battaglia C, Battaglia B, Morotti E, Paradisi R, Zanetti I, Meriggiola MC, Venturoli S. Two- and three-dimensional sonographic and color doppler techniques for diagnosis of polycystic ovary syndrome: the stromal/ovarian volume ratio as a new diagnostic criterion. *J Ultrasound Med* 2012;**31**:1015–1024.

Bili AE, Dampala K, Iakovou I, Tsolakidis D, Giannakou A, Tarlatzis BC. The combination of ovarian volume and outline has better diagnostic accuracy than prostate-specific antigen (PSA) concentrations in women with polycystic ovarian syndrome (PCOs). *Eur J Obstet Gynecol Reprod Biol* 2014;**179**:32–35.

Carmina E, Campagna AM, Fruzzetti F, Lobo RA. AMH measurement versus ovarian ultrasound in the diagnosis of polycystic ovary syndrome in different phenotypes. *Endocr Pract* 2016;**22**:287–293.

Carmina E, Orio F, Palomba S, Longo RA, Lombardi G, Lobo RA. Ovarian size and blood flow in women with polycystic ovary syndrome and their correlations with endocrine parameters. *Fertil Steril* 2005;**84**:413–419.

Chan JL, Kar S, Vanky E, Morin-Papunen L, Piltonen T, Puurunen J, Tapanainen JS, Maciel GAR, Hayashida SAY, Soares JM *et al.* Racial and ethnic differences in the prevalence of metabolic syndrome and its components of metabolic syndrome in women with polycystic ovary syndrome: a regional cross-sectional study. *Am J Obstet Gynecol* 2017;**217**:189.e1–189.e8.

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* 2008a; **32**:700–703.

Chen Y, Yang D, Li L, Chen X. The role of ovarian volume as a diagnostic criterion for Chinese adolescents with polycystic ovary syndrome. *J Pediatr Adolesc Gynecol* 2008b;**21**:347–350.

Christ JP, Vanden Brink H, Brooks ED, Pierson RA, Chizen DR, Lujan ME. Ultrasound features of polycystic ovaries relate to degree of reproductive and metabolic disturbance in polycystic ovary syndrome. *Fertil Steril* 2015;**103**:787–794.

Christ JP, Willis AD, Brooks ED, Vanden Brink H, Jarrett BY, Pierson RA, Chizen DR, Lujan ME. Follicle number, not assessments of the ovarian stroma, represents the best ultrasonographic marker of polycystic ovary syndrome. *Fertil Steril* 2014;**101**:280–287.e1.

Çiraci S, Tan S, Özcan AŞ, Aslan A, Keskin HL, Ateş ÖF, Akçay Y, Arslan H. Contribution of real-time elastography in diagnosis of polycystic ovary syndrome. *Diagn Interv Radiol* 2015;**21**:118–122.

Clark NM, Podolski AJ, Brooks ED, Chizen DR, Pierson RA, Lehotay DC, Lujan ME. Prevalence of polycystic ovary syndrome phenotypes using updated criteria for polycystic ovarian morphology: an assessment of over 100 consecutive women self-reporting features of polycystic ovary syndrome. *Reprod Sci* 2014; **21**:1034–1043.

de Guevara AL, Crisosto N, Echiburú B, Preisler J, Vantman N, Bollmann J, Pérez-Bravo F, Sir-Petermann T. Evaluation of ovarian function in 35–40-year-old women with polycystic ovary syndrome. *Eur J Obstet Gynecol Reprod Biol* 2013;**170**:165–170.

Deb S, Jayaprakasan K, Campbell BK, Clewes JS, Johnson IR, Raine-Fenning NJ. Intraobserver and interobserver reliability of automated antral follicle counts made using three-dimensional ultrasound and SonoAVC. *Ultrasound Obstet Gynecol* 2009;**33**:477–483.

Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol* 2005; **58**:882–893.

Dewailly D, Alebić M, Duhamel A, Stojanović N. Using cluster analysis to identify a homogeneous subpopulation of women with polycystic ovarian morphology in a population of non-hyperandrogenic women with regular menstrual cycles. *Hum Reprod* 2014a;**29**:2536–2543.

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.

Dewailly D, Lujan ME, Carmina E, Cedars MI, Laven J, Norman RJ, Escobar-Morreale HF. Definition and significance of polycystic ovarian morphology: a task force report from the androgen excess and polycystic ovary syndrome society. *Hum Reprod Update* 2014b;**20**:334–352.

Diamanti-Kandarakis E, Livadas S, Katsikis I, Piperi C, Mantziou A, Papavassiliou AG, Panidis D. Serum concentrations of carboxylated osteocalcin are increased and associated with several components of the polycystic ovarian syndrome. *J Bone Miner Metab* 2011;**29**:201–206.

Diamanti-Kandarakis E, Panidis D. Unravelling the phenotypic map of polycystic ovary syndrome (PCOS): a prospective study of 634 women with PCOS. *Clin Endocrinol (Oxf)* 2007;**67**:735–742.

Doebler P. mada: Meta-Analysis of Diagnostic Accuracy 2020. https://cran.r-project.org/package=mada.

Ertekin E, Turan OD, Tuncyurek O. Is shear wave elastography relevant in the diagnosis of polycystic ovarian syndrome?. *Med Ultrason* 2019;**21**:158–162.

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

Giménez-Peralta I, Lilue M, Mendoza N, Tesarik J, Mazheika M. Application of a new ultrasound criterion for the diagnosis of polycystic ovary syndrome. *Front Endocrinol (Lausanne)* 2022;**13**:915245.

Glintborg D, Mumm H, Hougaard D, Ravn P, Andersen M. Ethnic differences in Rotterdam criteria and metabolic risk factors in a multiethnic group of women with PCOS studied in Denmark. *Clin Endocrinol (Oxf)* 2010;**73**:732–738.

Glintborg D, Mumm H, Ravn P, Andersen M. Age associated differences in prevalence of individual Rotterdam criteria and metabolic risk factors during reproductive age in 446 Caucasian women with polycystic ovary syndrome. *Horm Metab Res* 2012;**44**:694–698.

Han YS, Lee AR, Song HK, Choi JI, Kim JH, Kim MR, Kim MJ. Ovarian volume in Korean women with polycystic ovary syndrome and its related factors. *J Menopausal Med* 2017;**23**:25–31.

Harrell Jr FE. Hmisc: Harrell Miscellaneous. 2021. https://hbiostat.org/R/Hmisc/.

Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, Welch V (eds), *Cochrane Handbook for Systematic Reviews of Interventions*, 2nd edn. Chirchester, UK: John Wiley & Sons. 2019.

Hudecova M, Holte J, Olovsson M, Sundstrom Poromaa I. Long-term follow-up of patients with polycystic ovary syndrome: reproductive outcome and ovarian reserve. *Hum Reprod* 2009;**24**:1176–1183.

Jacewicz-Święcka M, Wołczyński S, Kowalska I. The effect of ageing on clinical, hormonal and sonographic features associated with PCOS—a long-term follow-up study. *J Clin Med* 2021;**10**:2101.

Jackson CH. Multi-state models for panel data: the msm package for R. *J Stat Soft* 2011;**38**:1–29.

Jarrett BY, Vanden Brink H, Brooks ED, Hoeger KM, Spandorfer SD, Pierson RA, Chizen DR, Lujan ME. Impact of right–left differences in ovarian morphology on the ultrasound diagnosis of polycystic ovary syndrome. *Fertil Steril* 2019;**112**:939–946.

Jayaprakasan K, Campbell BK, Clewes JS, Johnson IR, Raine-Fenning NJ. Three-dimensional ultrasound improves the interobserver reliability of antral follicle counts and facilitates increased clinical work flow. *Ultrasound Obstet Gynecol* 2008;**31**:439–444.

Jayaprakasan K, Chan Y, Islam R, Haoula Z, Hopkisson J, Coomarasamy A, Raine-Fenning N. Prediction of in vitro fertilization outcome at different antral follicle count thresholds in a prospective cohort of 1,012 women. *Fertil Steril* 2012;**98**:657–663.

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, Reijo Pera R, 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, Dewailly D. Revisiting the ovarian volume as a diagnostic criterion for polycystic ovaries. *Hum Reprod* 2005;**20**:2893–2898.

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.

Kar S, Swoyam S. 2D and 3D trans-vaginal sonography to determine cut-offs for ovarian volume and follicle number per ovary for diagnosis of polycystic ovary syndrome in Indian women. *J Reprod Infertil* 2018;**19**:146–151.

Kauffman RP, Baker TE, Baker VM, Dimarino P, Castracane VD. Endocrine and metabolic differences among phenotypic expressions of polycystic ovary syndrome according to the 2003 Rotterdam consensus criteria. *Am J Obstet Gynecol* 2008;**198**:670e.e1.

Kazemi M, Jarrett BY, Brink HV, Lin AW, Hoeger KM, Spandorfer SD, Lujan ME. Obesity, insulin resistance, and hyperandrogenism mediate the link between poor diet quality and ovarian dysmorphology in reproductive-aged women. *Nutrients* 2020;**12**:1953.

van Keizerswaard J, Dietz de Loos ALP, Louwers YV, Laven JSE. Changes in individual polycystic ovary syndrome phenotypical characteristics over time: a long-term follow-up study. *Fertil Steril* 2022;**117**:1059–1066.

Khashchenko E, Uvarova E, Vysokikh M, Ivanets T, Krechetova L, Tarasova N, Sukhanova I, Mamedova F, Borovikov P, Balashov I et al The relevant hormonal levels and diagnostic features of polycystic ovary syndrome in adolescents. *J Clin Med* 2020;**9**:1831.

Kim HJ, Adams JM, Gudmundsson JA, Arason G, Pau CT, Welt CK. Polycystic ovary morphology: age-based ultrasound criteria. *Fertil Steril* 2017;**108**:548–553.

Köninger A, Koch L, Edimiris P, Enekwe A, Nagarajah J, Kasimir-Bauer S, Kimmig R, Strowitzki T, Schmidt B. Anti-Mullerian hormone: an indicator for the severity of polycystic ovarian syndrome. *Arch Gynecol Obstet* 2014;**290**:1023–1030.

Köşüş N, Köşüş A, Turhan NÖ. Relationship of ovarian volume with mean platelet volume and lipid profile in patients with polycystic ovary syndrome. *Exp Ther Med* 2011a;**2**:1141–1144.

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* 2011b;**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**:3117–3122.

Le NSV, Le MT, Nguyen ND, Tran NQT, Nguyen QHV, Cao TN. A cross-sectional study on potential ovarian volume and related factors in women with polycystic ovary syndrome from infertile couples. *Int J Womens Health* 2021;**13**:793–801.

Lee DE, Park SY, Lee SR, Jeong K, Chung HW. Diagnostic usefulness of transrectal ultrasound compared with transvaginal ultrasound assessment in young Korean women with polycystic ovary syndrome. *J Menopausal Med* 2015;**21**:149–154.

Lie Fong S, Laven JSE, Duhamel A, Dewailly D. Polycystic ovarian morphology and the diagnosis of polycystic ovary syndrome: redefining threshold levels for follicle count and serum anti-Müllerian hormone using cluster analysis. *Hum Reprod* 2017;**32**:1723–1731.

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* 2010;**36**:712–718.

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**:6.

Lujan ME, Chizen DR, Peppin AK, Kriegler S, Leswick DA, Bloski TG, Pierson RA. Improving inter-observer variability in the evaluation of ultrasonographic features of polycystic ovaries. *Reprod Biol Endocrinol* 2008;**6**:1–11.

Lujan ME, Jarrett BY, Brooks ED, Reines JK, Peppin AK, Muhn N, Haider E, Pierson RA, Chizen DR. Updated ultrasound criteria for polycystic ovary syndrome: reliable thresholds for elevated follicle population and ovarian volume. *Hum Reprod* 2013;**28**:1361–1368.

Macaskill P, Gatsonis C, Deeks J, Harbord R, Takwoingi Y. *Chapter 10: analysing and presenting results*. In: Deeks J, Bossuyt P, Gatsonis C (eds), *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*. The Cochrane Collaboration, 2010 [Internet]. http://srdta.cochrane.org/.

Maciel GAR, Baracat EC, Benda JA, Markham SM, Hensinger K, Chang RJ, Erickson GF. Stockpiling of transitional and classic primary follicles in ovaries of women with polycystic ovary syndrome. *J Clin Endocrinol Metab* 2004;**89**:5321–5327.

Malhotra N, Bahadur A, Singh N, Kalaivani M, Mittal S. Does obesity compromise ovarian reserve markers? A clinician's perspective. *Arch Gynecol Obstet* 2013;**287**:161–166.

McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, Clifford T, Cohen JF, Deeks JJ, Gatsonis C, Hooft L *et al.*; the PRISMA-DTA Group. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. *JAMA* 2018;**319**:388–396.

Mercé LT, Gómez B, Engels V, Bau S, Bajo JM. Intraobserver and interobserver reproducibility of ovarian volume, antral follicle count, and vascularity indices obtained with transvaginal 3-dimensional ultrasonography, power doppler angiography, and the virtual organ computer-aided analysis imaging program. *J Ultrasound Med* 2005;**24**:1279–1287.

Moslehi N, Shab-Bidar S, Ramezani Tehrani F, Mirmiran P, Azizi F. Is ovarian reserve associated with body mass index and obesity in reproductive aged women? A meta-analysis. *Menopause* 2018;**25**:1.

Nastri CO, Teixeira DM, Moroni RM, Leitão VMS, Martins WP. Ovarian hyperstimulation syndrome: pathophysiology, staging, prediction and prevention. *Ultrasound Obstet Gynecol* 2015;**45**:377–393.

Neubronner SA, Indran IR, Chan YH, Thu AWP, EL Y. Effect of body mass index (BMI) on phenotypic features of polycystic ovary syndrome (PCOS) in Singapore women: a prospective cross-sectional study. *BMC Women's Health* 2021;**21**:1–12.

Özay ÖE, Özay AC, Gün İ. Comparison of stromal thickness and doppler findings in polycystic ovary syndrome and healthy women with ultrasonographic evidence of polycystic ovaries? A cross-sectional study. *J Obstet Gynaecol* 2022;**42**:1–6.

Palomaki GE, Kalra B, Kumar T, Patel AS, Savjani G, Torchen LC, Dunaif A, Morrison A, Lambert-Messerlian GM, Kumar A. Adjusting antimullerian hormone levels for age and body mass index improves detection of polycystic ovary syndrome. *Fertil Steril* 2020;**113**:876–884.e2.

Peña AS, Witchel SF, Hoeger KM, Oberfield SE, Vogiatzi MG, Misso M, Garad R, Dabadghao P, Teede H. Adolescent polycystic ovary syndrome according to the international evidence-based guideline. *BMC Med* 2020;**18**:72.

Peigné M, Catteau-Jonard S, Robin G, Dumont A, Pigny P, Dewailly D. The numbers of 2-5 and 6-9 mm ovarian follicles are inversely correlated in both normal women and in polycystic ovary syndrome patients: what is the missing link? *Hum Reprod* 2018;**33**:706–714.

R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing, 2020. https://www.r-project.org/.

Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005;**58**:982–990.

Rosen MP, Johnstone E, McCulloch CE, Schuh-Huerta SM, Sternfeld B, Reijo-Pera RA, Cedars MI. A characterization of the relationship of ovarian reserve markers with age. *Fertil Steril* 2012;**97**:238–243.

Roth LW, Allshouse AA, Bradshaw-Pierce EL, Lesh J, Chosich J, Kohrt W, Bradford AP, Polotsky AJ, Santoro N. Luteal phase dynamics of follicle-stimulating and luteinizing hormones in obese and normal weight women. *Clin Endocrinol (Oxf)* 2014;**81**:418–425.

Rstudio Team. Rstudio: Integrated Development Environment for R. Boston: Rstudio, PBC, 2020. http://www.rstudio.com/.

Stein IF, Leventhal ML. Amenorrhea associated with bilateral polycystic ovaries. *Am J Obstet Gynecol* 1935;**46**:181–191.

Teede HJ, Misso ML, Costello MF, Dokras A, Laven J, Moran L, Piltonen T, Norman RJ, Andersen M, Azziz R *et al.*; International PCOS Network. Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome. *Hum Reprod* 2018a;**33**:1602–1618.

Teede HJ, Misso ML, Costello MF, Dokras A, Laven J, Moran L, Piltonen T, Norman RJ, Andersen M, Azziz R *et al.* Technical Report for: International Evidence-based Guideline for the Assessment and Management of Polycystic Ovary Syndrome 2018. Monash University, 2018b. https://www.monash.edu/medicine/mchri/pcos/guideline.

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* 2004;**81**:19–25.

Vanden Brink H, Jarrett BY, Pereira N, Spandorfer SD, Hoeger KM, Lujan ME. Diagnostic performance of ovarian morphology on ultrasonography across anovulatory conditions—impact of body mass index. *Diagnostics* 2023;**13**:374.

Vanden Brink H, Pisch AJ, Lujan ME. A comparison of two- and three-dimensional ultrasonographic methods for evaluation of ovarian follicle counts and classification of polycystic ovarian morphology. *Fertil Steril* 2021;**115**:761–770.

Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Soft* 2010;**36**:1–48.

Villa P, Rossodivita A, Sagnella F, Moruzzi MC, Mariano N, Lassandro AP, Pontecorvi A, Scambia G, Lanzone A. Ovarian volume and

gluco-insulinaemic markers in the diagnosis of PCOS during adolescence. *Clin Endocrinol (Oxf)* 2013;**78**:285–290. England.

Villarroel C, López P, Merino PM, Iñiguez G, Sir-Petermann T, Codner E. Hirsutism and oligomenorrhea are appropriate screening criteria for polycystic ovary syndrome in adolescents. *Gynecol Endocrinol* 2015;**31**:625–629.

Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Med Res Methodol* 2014;**14**:135.

Wang ET, Kao CN, Shinkai K, Pasch L, Cedars MI, Huddleston HG. Phenotypic comparison of Caucasian and Asian women with polycystic ovary syndrome: a cross-sectional study. *Fertil Steril* 2013;**100**:214–218.

Webber LJ, Stubbs S, Stark J, Trew GH, Margara R, Hardy K, Franks S. Formation and early development of follicles in the polycystic ovary. *Lancet* 2003;**362**:1017–1021.

Welt CK, Arason G, Gudmundsson JA, Adams J, Palsdóttir H, Gudlaugsdóttir G, Ingadóttir G, Crowley WF. Defining constant versus variable phenotypic features of women with polycystic ovary syndrome using different ethnic groups and populations. *J Clin Endocrinol Metab* 2006;**91**:4361–4368.

Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MMG, Sterne JAC, Bossuyt PMM, Group Q-2. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;**155**:529–536.

Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag, 2016. https://ggplot2.tidyverse.org.

Wiser A, Shalom-Paz E, Hyman JH, Sokal-Arnon T, Bantan N, Holzer H, Tulandi T. Age-related normogram for antral follicle count in women with polycystic ovary syndrome. *Reprod Biomed Online* 2013;**27**:414–418.

Wongwananuruk T, Panichyawat N, Indhavivadhana S, Rattanachaiyanont M, Angsuwathana S, Techatraisak K, Pratumvinit B, Sa-Nga-Areekul N. Accuracy of anti-Müllerian hormone and total follicles count to diagnose polycystic ovary syndrome in reproductive women. *Taiwan J Obstet Gynecol* 2018;**57**:499–506.

Yang B, Mallett S, Takwoingi Y, Davenport CF, Hyde CJ, Whiting PF, Deeks JJ, Leeflang MMG, Bossuyt PMM, Brazzelli MG *et al.*; QUADAS-C Group. QUADAS-C: a tool for assessing risk of bias in comparative diagnostic accuracy studies. *Ann Intern Med* 2021;**174**:1592–1599.

Zawadski J, Dunaif A. *Diagnostic criteria for polycystic ovary syndrome: towards a rational approach*. In: Dunaif A, Givens J, Haseltine F, Merriam G (eds), *Polycystic Ovary Syndrome*. Boston: Blackwell Scientific Publication, 1992.