National Institute for Health and Care Excellence

Polycystic ovary syndrome

Adaptation report 1 – Screening, diagnostic and risk assessment and life-stages

NICE guideline [NGXX]

July 2026

Draft for Consultation

Disclaimer

The recommendations in this guideline represent the view of NICE, arrived at after careful consideration of the evidence available. When exercising their judgement, professionals are expected to take this guideline fully into account, alongside the individual needs, preferences and values of their patients or service users. The recommendations in this guideline are not mandatory and the guideline does not override the responsibility of healthcare professionals to make decisions appropriate to the circumstances of the individual patient, in consultation with the patient and/or their carer or guardian.

Local commissioners and/or providers have a responsibility to enable the guideline to be applied when individual health professionals and their patients or service users wish to use it. They should do so in the context of local and national priorities for funding and developing services, and in light of their duties to have due regard to the need to eliminate unlawful discrimination, to advance equality of opportunity and to reduce health inequalities. Nothing in this guideline should be interpreted in a way that would be inconsistent with compliance with those duties.

NICE guidelines cover health and care in England. Decisions on how they apply in other UK countries are made by ministers in the Welsh Government, Scottish Government, and Northern Ireland Executive. All NICE guidance is subject to regular review and may be updated or withdrawn.

Copyright

ISBN: [XXX]

1

Contents

1 Screening, diagnostic and risk assessment and life-stages

The adaptation reports were produced using the reviews from the International Guideline (IG). Any further details can be found in the technical report from the IG, including results of the analyses and full study references.

1.1 Irregular cycles and ovulatory dysfunction

Review question 1.1: In adolescents, at what time point after onset of menarche do irregular cycles indicate ongoing menstrual dysfunction?

1.1.1 Recommendations from the International evidence-based guideline for PCOS*

Consensus recommendation:

1.1.1 Irregular menstrual cycles are defined as:

- Normal in the first-year post menarche as part of the pubertal transition
- 1 to < 3 years post menarche: < 21 or > 45 days
- 3 years post menarche to perimenopause: < 21 or > 35 days or < 8 cycles per year
- 1 year post menarche > 90 days for any one cycle

Primary amenorrhea by age 15 or > 3 years post thelarche (breast development)
When irregular menstrual cycles are present a diagnosis of PCOS should be considered and assessed according to these PCOS Guidelines.

Practice points:

1.1.2 The mean age of menarche may differ across populations.

1.1.3 In adolescents with irregular menstrual cycles, the value and optimal timing of assessment and diagnosis of PCOS should be discussed with the patient and their parent/s or guardian/s, considering diagnostic challenges at this life stage and psychosocial and cultural factors.

1.1.4 For adolescents who have features of PCOS, but do not meet diagnostic criteria, an 'increased risk' could be considered and reassessment advised at or before full reproductive maturity, 8 years post menarche. This includes those with PCOS features before combined oral contraceptive pill (COCP) commencement, those with persisting features and those with significant weight gain in adolescence

1.1.5 Ovulatory dysfunction can still occur with regular cycles and if anovulation needs to be confirmed serum progesterone levels can be measured.

IG clinical evidence

Critical appraisal - ROBIS systematic review checklist

Section	Question	Answer
Study eligibility criteria	Did the review adhere to pre-defined objectives and eligibility criteria?	Yes *(PICO was clear and detailed, search dates clearly listed)*
Study eligibility criteria	Were the eligibility criteria appropriate for the review question?	Yes *(Appropriate for review)*
Study eligibility criteria	Were eligibility criteria unambiguous?	Yes *(Clearly defined eligibility criteria)*
Study eligibility criteria	Were all restrictions in eligibility criteria based on study characteristics appropriate (e.g. date, sample size, study quality, outcomes measured)?	Probably yes *(Searched from 2017 to 2022 but not stated in*

Section	Question	Answer
		protocol, but as it is an update it is reasonable)
Study eligibility criteria	Were any restrictions in eligibility criteria based on sources of information appropriate (e.g. publication status or format, language, availability of data)?	Yes *(English language restrictions were appropriate)*
Study eligibility criteria	Concerns regarding specification of study eligibility criteria	Low *(All areas appear to be well covered with relevant information)*
Identification and selection of studies	Did the search include an appropriate range of databases/electronic sources for published and unpublished reports?	Yes *(5 stated, all appropriate to review question)*
Identification and selection of studies	Were methods additional to database searching used to identify relevant reports?	No information *(No information stated on additional search methods such as manual searching)*
Identification and selection of studies	Were the terms and structure of the search strategy likely to retrieve as many eligible studies as possible?	Probably yes *(Full search strategy listed, appears to have appropriate search terms)*
Identification and selection of studies	Were restrictions based on date, publication format, or language appropriate?	Yes *(the search dates are appropriate, covering the time period between the previous review and this review.*
Identification and selection of studies	Were efforts made to minimise error in selection of studies?	Probably yes *(One reviewer selected and appraised studies, with full text retrieval if decision could not be made)*
Identification and selection of studies	Concerns regarding methods used to identify and/or select studies	Low *(Search terms seem appropriate, with reasonable dates and databases used)*

Section	Question	Answer
Data collection and study appraisal	Were efforts made to minimise error in data collection?	Probably yes *(Only one reviewer, ideally there would be 2, but this may be because no studies were suitable)*
Data collection and study appraisal	Were sufficient study characteristics available for both review authors and readers to be able to interpret the results?	N/A *(No studies available)*
Data collection and study appraisal	Were all relevant study results collected for use in the synthesis?	N/A *(No studies available)*
Data collection and study appraisal	Was risk of bias (or methodological quality) formally assessed using appropriate criteria?	N/A *(No studies available)*
Data collection and study appraisal	Were efforts made to minimise error in risk of bias assessment?	N/A *(No studies available)*
Data collection and study appraisal	Concerns regarding methods used to collect data and appraise studies	N/A *(No studies met the inclusion criteria, as such this section is largely N/A)*
Synthesis and findings	Did the synthesis include all studies that it should?	N/A *(No studies available*)
Synthesis and findings	Were all pre-defined analyses reported or departures explained?	N/A *(No studies available)*
Synthesis and findings	Was the synthesis appropriate given the nature and similarity in the research questions, study designs and outcomes across included studies?	N/A *(No studies available)*
Synthesis and findings	Was between-study variation (heterogeneity) minimal or addressed in the synthesis?	N/A *(No studies available)*
Synthesis and findings	Were the findings robust, e.g. as demonstrated through funnel plot or sensitivity analyses?	N/A *(No studies available)*

Section	Question	Answer
Synthesis and findings	Were biases in primary studies minimal or addressed in the synthesis?	N/A *(No studies available)*
Synthesis and findings	Concerns regarding the synthesis and findings	N/A *(No studies available)*
Judging risk of bias	Concerns regarding specification of study eligibility	Low *(All areas appear to be well covered with relevant information)*
Judging risk of bias	Concerns regarding methods used to identify and/or select studies	Low *(Search terms seem appropriate, with reasonable dates and databases used)*
Judging risk of bias	Concerns regarding methods used to collect data and appraise studies	N/A *(No studies met the inclusion criteria, as such this section is largely N/A)*
Judging risk of bias	Concerns regarding the synthesis and findings	N/A *(No studies available)*
Overall review ratings	Overall risk of bias	N/A *(No studies met the inclusion criteria, as such this section is largely N/A)*
Overall review ratings	Applicability as a source of data	Partially applicable *(search results did not find any includable studies)*

Evidence to recommendations justification: No evidence was found for the systematic review so there were no evidence-based recommendations (EBRs), only consensus-based recommendations (CRs) and practice points. The CRs made suggestions based on the natural history of menstrual cycles and ovulation in those under 18 years to describe irregular menstrual cycles.

IG economic evidence

No health economic evidence was identified in the international guideline (IG) for review question 1.1.

The IG did however note that to effectively implement their recommendations for adolescents with irregular cycles post menarche; education on timing of diagnosis and normal physiologic processes may need to be provided to paediatricians, physicians and gynaecologists.

1.1.2 NICE economic evidence

Included studies

A single health economic search was performed by NICE to identify published economic evaluations of relevance to all review questions in this guideline. See the literature search strategy in Appendix A.

No economic studies were identified which were applicable to this review question (see economic study selection flow chart in Appendix B).

Excluded studies

No economic studies were reviewed at full text and excluded from this review.

Economic model

Review question 1.1 in the IG was not prioritised for original health economic modelling and therefore no original health economic model was conducted for this review question.

1.1.3 NICE recommendations

The relevant recommendations for this section are Rec 1.3.1 to 1.3.3.

1.1.4 The committee's discussion and interpretation of the evidence

Clinical

The committee generally agreed with CR 1.1.1 in the International Guideline regarding the definition of irregular menstrual cycles and adapted it to the standard NICE recommendation style. The recommendation was broadened to include absent menstrual cycles, as this terminology is used in a UK context.

This section was titled 'when to suspect PCOS' as the committee wanted to ensure that diagnosis is not delayed and that PCOS would be considered early in the pathway. Two new recommendations were added, one for suspecting PCOS if there are signs or symptoms of hyperandrogenism (even if regular menstrual cycles) and when an incidental finding of polycystic ovary morphology (PCOM) is found during other investigations. The committee felt it would be helpful to add this information to highlight that PCOS can still be suspected with regular menstrual cycles.

Health economic

No health economic evidence was identified in the IG for review question 1.1 on irregular cycles and ovulatory disfunction. In addition, no health economic evidence was identified for this review question in the health economic literature search conducted by NICE.

The committee discussed the recommendations in the IG and noted that the IG's recommendations are in line with current NHS practice. As these recommendations were primarily concerned with providing information on the signs and symptoms of irregular menstrual cycles, and how this could be indicative of PCOS, the contextualised recommendations on irregular cycles and menstrual dysfunction do not have any associated resource implications.

1.2 Biochemical hyperandrogenism

Review question Q1.2: In women with suspected PCOS, what is the most effective measure to diagnose PCOS related hyperandrogenism (biochemical)?

1.2.1 Recommendations from the International evidence-based guideline for PCOS*

Evidence-based recommendations:

1.2.1 Healthcare professionals should use total and free testosterone to assess biochemical hyperandrogenism in the diagnosis of PCOS; free testosterone can be estimated by the calculated free androgen index.

1.2.2 If testosterone or free testosterone is not elevated, healthcare professionals could consider measuring androstenedione and dehydroepiandrosterone sulfate (DHEAS), noting their poorer specificity and greater age associated decrease in DHEAS.

1.2.3 Laboratories should use validated, highly accurate tandem mass spectrometry (LC-MS/MS) assays for measuring total testosterone and if needed, for androstenedione and DHEAS. Free testosterone should be assessed by calculation, equilibrium dialysis or ammonium sulfate precipitation.

1.2.4 Laboratories should use LC-MS/MS assays over direct immunoassays (e.g. radiometric, enzyme-linked, etc.) for assessing total or free testosterone, which have limited accuracy and demonstrate poor sensitivity and precision for diagnosing hyperandrogenism in PCOS.

Practice points:

1.2.5 For the detection of hyperandrogenism in PCOS, the assessment of biochemical hyperandrogenism is of greatest value in patients with minimal or no clinical signs of hyperandrogenism (i.e. hirsutism).

1.2.6 It is very difficult to reliably assess for biochemical hyperandrogenism in women on the combined oral contraceptive pill (COCP) as the pill increases sex hormone-binding globulin and reduces gonadotrophin-dependent androgen production. If already on the COCP, yet assessment of biochemical androgens is imperative, the pill should be withdrawn for a minimum of three months, and contraception should be managed otherwise during this time.

1.2.7 Repeated androgen measures for the ongoing assessment of PCOS in adults have a limited role.

1.2.8 In most adolescents, androgen levels reach adult ranges at 12-15 years of age.

1.2.9 If androgen levels are markedly above laboratory reference ranges, causes of hyperandrogenaemia other than PCOS, including ovarian and adrenal neoplastic growths, congenital adrenal hyperplasia, Cushing's syndrome, ovarian hyperthecosis (after menopause), iatrogenic causes, and syndromes of severe insulin resistance, should be considered. However, some androgen-secreting neoplasms are associated with only mild to moderate increases in androgen levels. The clinical history of time of onset and/or rapid progression of symptoms is critical in assessing for an androgen-secreting tumour.

1.2.10 Reference ranges for different methods and laboratories vary widely and are often based on an arbitrary percentile or variances of the mean from a population that has not been fully characterised and is highly likely to include women with PCOS. Normal values should be determined either by the range of values in a well characterised healthy control population or by cluster analysis of general population values

1.2.11 Laboratories involved in androgen measurements in females should consider: determining laboratory normal values by either the range of values in a well characterised healthy control population or by cluster analysis of the values of a large general population applying the most accurate methods where available using

extraction/chromatography immunoassays as an alternative to mass spectrometry only where adequate expertise is available future improvements may arise from measurement of 11-oxygenated androgens, and from establishing cut-off levels or thresholds based on large-scale validation in populations of different ages and ethnicities.

IG clinical evidence

Critical appraisal - ROBIS systematic review checklist

Section	Question	Answer
Study eligibility criteria	Did the review adhere to pre-defined objectives and eligibility criteria?	Yes *(PICO was clear and detailed, search dates clearly listed)*
Study eligibility criteria	Were the eligibility criteria appropriate for the review question?	Yes *(appropriate for review but more details about reference standard would be useful)*
Study eligibility criteria	Were eligibility criteria unambiguous?	Yes *(PICO is detailed however it includes RCTs and comparative prospective cohort studies, but the outcomes are mainly diagnostic accuracy)*
Study eligibility criteria	Were all restrictions in eligibility criteria based on study characteristics appropriate (e.g. date, sample size, study quality, outcomes measured)?	Yes *(restrictions on date and language are present but not well described. They do not state the date but says that they updated WHO searches)*
Study eligibility criteria	Were any restrictions in eligibility criteria based on sources of information appropriate (e.g. publication status or format, language, availability of data)?	Yes *(most eligibility appropriate for review question, English language limit is appropriate)*
Study eligibility criteria	Concerns regarding specification of study eligibility criteria	Low *(all parameters are appropriate to review)*

Section	Question	Answer
Identification and selection of studies	Did the search include an appropriate range of databases/electronic sources for published and unpublished reports?	Probably yes *(names 5 databases used, all appropriate to review, PsycINFO may be superfluous)*
Identification and selection of studies	Were methods additional to database searching used to identify relevant reports?	Probably yes *(does not detail any other additional methods such as manual search strategies, however the general methods says that they sought relevant papers from their experts)*
Identification and selection of studies	Were the terms and structure of the search strategy likely to retrieve as many eligible studies as possible?	No *(they have added a diagnostic filter (which has been shown to miss relevant studies) onto a search that has already combined (added) terms for tools, diagnosis, hyperandrogenism and specific diagnostic texts. This would be expected to over exclude references).* *The terms for hyperandrogenism do not include subject headings, it is good practice to combine these with free text terms. Other subject headings are missing e.g. testosterone. Subject headings have not been translated for individual databases.* *No information is provided for CINAHL beyond "same search". Subject headings and syntax will be different.* *Pre-selected limits for female and humans have been applied. This is risky as indexing can be inconsistent. We would use tested methods to exclude animal studies.*
Identification and selection of studies	Were restrictions based on date, publication format, or language appropriate?	Probably yes *(pragmatic application of date and language limits as per protocol)*
Identification and selection of studies	Were efforts made to minimise error in selection of studies?	Yes *(2 reviewers selected and appraised studies, alongside evidence team)*
Identification and selection of studies	Concerns regarding methods used to identify and/or select studies	High

Section	Question	Answer
		(Major concerns about application of diagnostic filter on top of the search blocks already combining these terms. Numbers reported in PRISMA diagram seem very low)
Data collection and study appraisal	Were efforts made to minimise error in data collection?	Yes *(2 independent reviewers plus additional input from evidence team, where required)*
Data collection and study appraisal	Were sufficient study characteristics available for both review authors and readers to be able to interpret the results?	Yes *(detailed study characteristics table available, appears appropriate for interpretation of results)*
Data collection and study appraisal	Were all relevant study results collected for use in the synthesis?	Yes *(evidence summary details which and how many studies were included for each test (serum testosterone, serum androstenedione, serum DHEAS, FAI, free testosterone, DHT, serum SHBG, salivary testosterone)*
Data collection and study appraisal	Was risk of bias (or methodological quality) formally assessed using appropriate criteria?	Probably yes *(risk of bias assessment completed for each study, also included is a quality appraisal table but not full study extractions as in other reviews)*
Data collection and study appraisal	Were efforts made to minimise error in risk of bias assessment?	Probably yes *(evidence summary mentions reason for high risk of bias in some studies is due to being single centre studies and advises caution on interpretation of these results)*
Data collection and study appraisal	Concerns regarding methods used to collect data and appraise studies	Low *(methods seem appropriate for review)*
Synthesis and findings	Did the synthesis include all studies that it should?	Unclear *(PRISMA diagram details search results and included studies, 17 were included in SR, 15 in meta-analyses in PRISMA but. 16 studies seem to be included in the meta-analyses/review. Results are separated by test method, as not all studies covered all tests, but where meta-analysis was not possible. The results were provided in a*

Section	Question	Answer
		forest plot stating it could not be meta-analysed)
Synthesis and findings	Were all pre-defined analyses reported or departures explained?	No *(lack of information about analysis plan in the technical report or the full guideline except for subgroup untreated, this was not included/found. It was sub-grouped by immunoassays/LC-MS/MS within the results, but methods of assessment was not a subgroup in the protocol. Case-control studies were included even though they excluded anything that was below a comparative, prospective cohort study in the PICO)*
Synthesis and findings	Was the synthesis appropriate given the nature and similarity in the research questions, study designs and outcomes across included studies?	Yes *(meta-analysis was performed on 15 of the 17 included studies. This seems appropriate given differences in index tests. Various reference standards and measurement of index tests combined, although sub-group analyses were done for measurement methods LC-MC/MS and immunoassay. Various cut-off points combined so SROC may be more appropriate than bivariate model)*
Synthesis and findings	Was between-study variation (heterogeneity) minimal or addressed in the synthesis?	Probably yes *(only one analysis states random effects model used, others don't mention whether random or fixed was used. Two studies were not included because there was only one study per test outcome for salivary testosterone and DHT. GRADE tables were rated down for inconsistency in 2 outcomes, DHEAS and free testosterone and subgroup analysis was done for both)*
Synthesis and findings	Were the findings robust, e.g. as demonstrated through funnel plot or sensitivity analyses?	Yes *(SROC curves are available for the meta-analysis, all curves appear to highlight good sensitivity and specificity. No funnel plot or sensitivity analysis conducted. For some outcomes this is appropriate due to low numbers of included studies, however some outcomes had larger numbers of studies which could have had funnel plots conducted)*

Section	Question	Answer
Synthesis and findings	Were biases in primary studies minimal or addressed in the synthesis?	Probably yes *(details provided in quality appraisal summary table regarding biases not in full extraction. GRADE downgraded for bias)*
Synthesis and findings	Concerns regarding the synthesis and findings	Unclear *(unclear due to mixed findings from the above sections)*
Judging risk of bias	Concerns regarding specification of study eligibility	Low *(all parameters are appropriate to review)*
Judging risk of bias	Concerns regarding methods used to identify and/or select studies	High *(Major concerns about application of diagnostic filter on top of the search blocks already combining these terms. Numbers reported in PRISMA diagram seem very low).*
Judging risk of bias	Concerns regarding methods used to collect data and appraise studies	Low *(methods seem appropriate for review)*
Judging risk of bias	Concerns regarding the synthesis and findings	Unclear *(unclear due to mixed findings from the above sections)*
Overall review ratings	Overall risk of bias	Unclear *(overall the criterion for this section appears to be well met, however there are some areas of uncertainty)*
Overall review ratings	Applicability as a source of data	Fully applicable

Evidence to recommendations justification: The International Guideline used the GRADE framework to describe how they went from the evidence to the recommendations. The overall GRADE rating of the certainty of the evidence for section 1.2 was rated as very low. However, the recommendations are a mix of two strong recommendations for the option, one conditional weak recommendation and one strong recommendation against the option, based on the GRADE framework.

The International Guideline highlighted potential limitations of the evidence found, such as most studies were conducted in single centres, with limited numbers of participants and methodological issues. They highlight that most studies had insufficient evidence for the diagnostic accuracy of the index tests reported in the

included studies, all of which had serious of very serious risk of bias and very low certainty of the evidence. The EBRs are very strong in wording given the evidence was insufficient and of low quality, however the IG uses a different way of assessing strength of recommendations from NICE guidelines, and most of the evidence showed good sensitivity/specificity which could support the strength of recommendation in the IG. Sensitivity shows the ability of a test to correctly identify true positives (people who have the condition), specificity shows the ability of a test to correctly identify true negatives (people who do not have the conditions).

Calculated free testosterone and free androgen index (FAI) had the best results, as both their sensitivity was over 80% and specificity was higher in the calculated free testosterone with 93% (95% CI 80 to 98%), when sub-grouped by measurement method LC-MC/MS had higher specificity but lower sensitivity than immunoassay. FAI specificity was 64% (95% CI 77 to 92%), however there seems to have been an error in reporting these figures in the Monash technical document, as the 95% CI does not include the value of the effect estimate. FAI sensitivity was 80% (95% CI 73 to 86%) in 9 studies, which when sub-grouped by measurement method the sensitivity and specificity was higher when FAI was measured by LC-MC/MS. These findings justify the inclusion of calculated free testosterone and FAI in the first recommendation and the use of tandem-mass spectrometry in the fourth recommendation.

The evidence for serum total testosterone came from 11 studies, most of which were high risk of bias in GRADE. The sensitivity was 72% (95% CI 60 to 81%) ranging from 29% to 93% and specificity 87% (95% CI 75% to 94%), ranging from 60% to 100%. The IG sub-grouped by immunoassay compared to LC-MC/MS, which showed lower sensitivity for LC-MS/MS (65%, 95% CI 41% to 84%), but specificity was higher at 93% (95% CI 73% to 98%). They reported that sensitivity was around 70% and specificity 75-85% for serum total testosterone, androstenedione and DHEA-S, which supports recommendation two.

The analyses combined various cut-off thresholds and various reference standards, which differs from NICE methods, where we typically group or pre-define cut-offs for meta-analyses. However, they state a major research priority would be to have large-scale studies comparing DHEAS, A4 and Testosterone measured by tandem mass spectrometry with robust diagnostic accuracy methodology using a pre-defined cut-off which is prospectively tested and validated in independent cohorts of women with well-phenotyped PCOS and women without clinical evidence of PCOS.

IG economic evidence

No health economic evidence was identified in the International Guideline (IG) for review question 1.2 on biochemical hyperandrogenism.

The IG also noted that tandem mass spectrometry assays are considered the reference standard for assessing biochemical hyperandrogenism, but access and costs were raised as an implementation concern.

1.2.2 NICE economic evidence

Included studies

A single health economic search was performed by NICE to identify published economic evaluations of relevance to all review questions in this guideline. See the literature search strategy in Appendix A .

No economic studies were identified which were applicable to this review question (see economic study selection flow chart in Appendix B).

Excluded studies

No economic studies were reviewed at full text and excluded from this review.

Economic model

No original health economic modelling was conducted for review question 1.2 on biochemical hyperandrogenism as other areas of the guideline were judged to be a higher priority for health economic modelling.

1.2.2 NICE recommendations

The relevant recommendations for this section are Rec 1.4.1, 1.4.2, 1.4.6 to 1.4.9, 1.6.1 and 1.19.1.

1.2.3 The committee's discussion and interpretation of the evidence

Clinical

The IG included four EBRs (1.2.1 to 1.2.4) and seven practice points (1.2.5 to 1.2.11) for this review question.

The committee decided to contextualise two EBRs (1.2.1 and 1.2.2) and two practice points (1.2.7 and 1.2.8). The two contextualised EBRs provide information on the initial tests that should be conducted for assessing biochemical hyperandrogenism, in addition to tests which should be considered if testosterone levels are not elevated but biochemical hyperandrogenism is still suspected.

The committee discussed the ways in which people might present that required the use of biochemical hyperandrogenism testing, compared to presenting with the clinical signs and features of hyperandrogenism. The committee noted that in current clinical practice, testing for biochemical hyperandrogenism would be conducted alongside other blood tests to rule out other causes of symptom presentation. The committee also noted that high androgen levels could also indicate the presence of certain tumours, which would require further investigation. The committee also discussed where the testing for biochemical hyperandrogenism testing should take place, in primary or secondary care. The committee agreed that blood tests were often ordered in primary care and were useful in clinics for several reasons, such as establishing other causes of hyperandrogenism when there was marked elevation in testosterone. The committee also discussed that oral oestrogen medications such as the combined contraceptive pill amongst others can interfere with the biochemical hyperandrogenism testing. Recommendations were adjusted to reflect this, with the addition of advising a change to non-oral oestrogens for a period of 3 months prior to a biochemical test in order to obtain an accurate result.

The committee discussed the IG's recommendations alongside the IG's diagnostic algorithm. In general, the committee acknowledged that several recommendations made in the IG were concerned with providing laboratory instructions for interpreting the results of biochemical tests. The committee concluded that these recommendations could be omitted from the guideline as they sit outside the remit of what would usually be recommended in a NICE guideline.

Health economic

No health economic evidence was included in the IG for evidence review 1.2 on biochemical hyperandrogenism. In addition, no health economic evidence was identified in the health economic literature search conducted by NICE.

The committee noted that the two EBRs they contextualised provide information on the initial tests that should be conducted for assessing biochemical hyperandrogenism, in addition to tests which should be considered if testosterone levels are not elevated but biochemical hyperandrogenism is still suspected. The committee acknowledged that these two recommendations are reflective of current UK practice and therefore no significant resource impact is anticipated.

For the two practice point recommendations that were contextualised, one of these recommendations detailed the age at which androgen levels reach adult ranges and the other recommendation noted that androgen levels should not be routinely retested as part of ongoing assessments for PCOS. The committee noted that these recommendations are reflective of best current practice, however acknowledged that in instances where androgen testing is being conducted on a regular basis as part of ongoing assessment, cost savings could be observed. Overall, the committee concluded that as these recommendations provide advice on monitoring and assessment, and these are likely to be similar to current NHS practice, these recommendations would be unlikely to add an additional financial burden to the NHS.

The committee decided to deviate from the IG's diagnostic algorithm where PCOS can be diagnosed based on clinical hyperandrogenism and irregular menstrual cycles

alone. Although the committee agreed that this is clinically possible, it was noted they would still want to assess for biochemical hyperandrogenism at this stage of the diagnostic pathway, conducting this test alongside the other blood tests that are conducted to rule out other causes of symptom presentation. The committee acknowledged that testing for biochemical hyperandrogenism at this stage of the diagnostic pathway would be more cost-effective than conducting a separate blood test to assess for biochemical hyperandrogenism at a later stage. The committee also discussed that high androgen levels can be indicative of certain tumours, and although these are rare, biochemical tests are the best way to identify raised androgen levels that could be indicative of these tumours. Whereby a high androgen level would require referral to secondary care for further investigation. The committee noted that the diagnostic pathway recommended as part of this guideline is reflective of UK current practice. Therefore the deviation from the IG recommendations will not result in a significant resource impact.

1.3 Clinical hyperandrogenism

Review question 1.3: In women with suspected PCOS, what is the most effective measure to clinically diagnose PCOS related hyperandrogenism?

1.3.1 Recommendations from the International evidence-based guideline for PCOS*

Evidence-based recommendations:

1.3.1 The presence of hirsutism alone should be considered predictive of biochemical hyperandrogenism and PCOS in adults.

1.3.2 Healthcare professionals could recognise that female pattern hair loss and acne in isolation (without hirsutism) are relatively weak predictors of biochemical hyperandrogenism.

Consensus recommendations:

1.3.3 A comprehensive history and physical examination should be completed for symptoms and signs of clinical hyperandrogenism, including acne, female pattern hair loss and hirsutism in adults, and severe acne and hirsutism in adolescents.

1.3.4 Healthcare professionals should be aware of the potential negative psychosocial impact of clinical hyperandrogenism and should consider the reporting of unwanted excess hair growth and/or female pattern hair loss as being important, regardless of apparent clinical severity.

1.3.5 A modified Ferriman Gallwey score (mFG) of 4–6 should be used to detect hirsutism, depending on ethnicity, acknowledging that self-treatment is common and can limit clinical assessment.

1.3.6 Healthcare professionals should consider that the severity of hirsutism may vary by ethnicity, but the prevalence of hirsutism appears similar across ethnicities

Practice points:

1.3.7 Healthcare professionals should:

- be aware that standardised visual scales are preferred when assessing hirsutism, such as the mFG scale in combination with a photographic atlas
- consider the Ludwig or Olsen visual scales for assessing female pattern hair loss note that there are no universally accepted visual instruments for assessing the presence of acne
- recognise that women commonly treat clinical hyperandrogenism cosmetically, diminishing their apparent clinical severity
- appreciate that self-assessment of unwanted excess hair growth, and possibly acne and female pattern hair loss, has a high degree of validity and merits close evaluation, even if overt clinical signs of hyperandrogenism are not readily evident on examination
- be aware that only terminal hairs need to be considered in defining hirsutism, and these can reach > 5 mm if untreated, vary in shape and texture and are generally pigmented
- note that new-onset severe or worsening hyperandrogenism, including hirsutism, requires further investigation to rule out androgen-secreting tumours and ovarian hyperthecosis
- monitor clinical signs of hyperandrogenism, including hirsutism, acne and female pattern hair loss, for improvement or treatment adjustment during therapy.

IG clinical evidence

Critical appraisal - ROBIS systematic review checklist

Section	Question	Answer
Study eligibility criteria	Did the review adhere to pre-defined objectives and eligibility criteria?	Yes *(search details clearly listed, PICO seems appropriate)*
Study eligibility criteria	Were the eligibility criteria appropriate for the review question?	Probably yes *(criteria appear appropriate for review, although the reference standard is not*

Section	Question	Answer
		clearly stated and we would de-prioritise case-control studies)
Study eligibility criteria	Were eligibility criteria unambiguous?	Yes *(eligibility criteria clearly described in PICO)*
Study eligibility criteria	Were all restrictions in eligibility criteria based on study characteristics appropriate (e.g. date, sample size, study quality, outcomes measured)?	Probably yes *(restrictions on date are present but not well described)*
Study eligibility criteria	Were any restrictions in eligibility criteria based on sources of information appropriate (e.g. publication status or format, language, availability of data)?	Yes *(eligibility appears appropriate for review question, English language and human studies limits are appropriate)*
Study eligibility criteria	Concerns regarding specification of study eligibility criteria	Low *(most PICO is appropriate although it would be better for some details to be more clearly stated)*
Identification and selection of studies	Did the search include an appropriate range of databases/electronic sources for published and unpublished reports?	Probably yes *(5 databases named, all are appropriate for review except Psychinfo may be superfluous)*
Identification and selection of studies	Were methods additional to database searching used to identify relevant reports?	Probably yes *(does not detail any additional searches/manual search strategies however the general methods says that they sought relevant papers from their experts)*
Identification and selection of studies	Were the terms and structure of the search strategy likely to retrieve as many eligible studies as possible?	No *(subject headings are not translated for specific databases and in some cases missing.* *The diagnosis filter (which has been shown to exclude relevant studies) is added onto specific lines for diagnosis terms, leading to likely over exclusion.*

Section	Question	Answer
		Use of automated limits to exclude animal studies rather than tested other methods)
Identification and selection of studies	Were restrictions based on date, publication format, or language appropriate?	Probably no *(They have applied the date limits separately with and without the diagnostic filter, but the presentation of the search strategy does not show these, or the separate numbers reported if they were sifted as two sets)*
Identification and selection of studies	Were efforts made to minimise error in selection of studies?	Yes *(2 reviewers for title and abstract, full text used if decision could not be made)*
Identification and selection of studies	Concerns regarding methods used to identify and/or select studies	High *(use of diagnostic filters on top of diagnostic terms and lack of accurate subject headings for individual databases)*
Data collection and study appraisal	Were efforts made to minimise error in data collection?	Probably yes *(only one reviewer for selection and appraisal of studies, two is preferable)*
Data collection and study appraisal	Were sufficient study characteristics available for both review authors and readers to be able to interpret the results?	Probably yes *(characteristics table available, seems appropriate for result interpretation)*
Data collection and study appraisal	Were all relevant study results collected for use in the synthesis?	Probably yes *(number of studies included and reasons for exclusion detailed in evidence review, no obvious exclusions noted)*
Data collection and study appraisal	Was risk of bias (or methodological quality) formally assessed using appropriate criteria?	Probably yes *(risk of bias assessment noted for each study and assessed using structured extraction form with some questions from QUADAS-2)*
Data collection and study appraisal	Were efforts made to minimise error in risk of bias assessment?	Probably yes *(evidence summary briefly mentions reasons for risk of bias where relevant)*
Data collection and study appraisal	Concerns regarding methods used to collect data and appraise studies	Low *(methods seem appropriate for reviews)*

Section	Question	Answer
Synthesis and findings	Did the synthesis include all studies that it should?	Yes *(PRISMA diagram details search results and included studies included 4 original studies, 1 systematic review and 3 additional studies from the SR, total included in GRADE profiles - 7)*
Synthesis and findings	Were all pre-defined analyses reported or departures explained?	No information *(Subgroup in protocol was untreated but not enough studies to subgroup where heterogeneity)*
Synthesis and findings	Was the synthesis appropriate given the nature and similarity in the research questions, study designs and outcomes across included studies?	Probably yes *(They used descriptive analysis because meta-analysis was not possible and explained why. They state very low quality due to imprecision, due to small sample sizes or single studies. Case-control studies were also included which was not mentioned in the PICO)*
Synthesis and findings	Was between-study variation (heterogeneity) minimal or addressed in the synthesis?	Yes *(they did not meta-analyse because of significant heterogeneity of studies and downgraded for very serious inconsistency in GRADE).*
Synthesis and findings	Were the findings robust, e.g. as demonstrated through funnel plot or sensitivity analyses?	Probably yes *(there were not enough studies to do funnel plots and no meta-analyses conducted)*
Synthesis and findings	Were biases in primary studies minimal or addressed in the synthesis?	Probably yes *(GRADE table presented listing risk of bias, no further discussion on attempts to minimise risk of bias)*
Synthesis and findings	Concerns regarding the synthesis and findings	Low *(most were probably yes because they could not meta-analyse so used descriptive analysis)*
Judging risk of bias	Concerns regarding specification of study eligibility	Low *(most PICO is appropriate although it would be better for some details to be more clearly stated)*
Judging risk of bias	Concerns regarding methods used to identify and/or select studies	High *(use of diagnostic filters on top of diagnostic terms and lack of accurate*

Section	Question	Answer
		subject headings for individual databases)
Judging risk of bias	Concerns regarding methods used to collect data and appraise studies	Low *(methods seem appropriate for reviews)*
Judging risk of bias	Concerns regarding the synthesis and findings	Low *(most were probably yes because they could not meta-analyse so used descriptive analysis)*
Overall review ratings	Overall risk of bias	Low *(the eligibility criteria, methods to collect and appraise data are appropriate, although diagnostic search filters may be overused).*
Overall review ratings	Applicability as a source of data	Fully applicable

Evidence to recommendations justification: Seven studies were included in the updated systematic review in the International Guideline. Each of the 11 comparisons assessed in this review question had a GRADE score of very low for critical outcomes. The authors describe how measuring both hirsutism and acne can be difficult due to a lack of established reference ranges, they also highlight that ethnicity can affect hair growth. At the time of the review, The PCOS Phenotypes in Unselected Populations (P-PUP) study (Demis Bizuneh 2025) was in progress, to assess the cut-off values for different racial and ethnic groups using the modified Ferriman and Gallwey (mFG) scoring system, which the authors stated could add evidence to this area. This has since been published and included 9979 women from the general population, aged 18-45 who were assessed for PCOS features. Studies included women from 11 ethnicities and eight countries. The PUPP study found significant heterogeneity in both the definitions used for PCOS and the data collection methods across their included studies, which may have affected the estimate of PCOS prevalence. The PUPP study concludes that there is a clear lack of standardised reporting, data collection and cut offs for PCOS diagnostic features.,

which supports the high priority research areas outlined in the international PCOS guideline.

Both EBRs are a conditional (weak) recommendation for the option in the IG; the CRs are a mix of conditional (weak) and strong recommendation for the option in the IG. The EBR suggests that “hirsutism alone should be considered predictive of biochemical hyperandrogenism and PCOS” however the evidence found is less clear. Three studies compared the specificity and sensitivity of mFG to total and free testosterone, and two studies compared mFG to DHEAS. Modified Ferriman-Gallwey (mFG) >8 was compared to various reference standards (total testosterone, free testosterone, DHEAS, androstenedione), however meta-analysis was not possible due to significant heterogeneity. When the study results were reported individually, they did not show adequate paired sensitivity and specificity. None of the studies used LC-MC/MC as a reference standard for measuring total testosterone. There were also 3 different measurement reference standards of total testosterone (ELISA, ECL and CLIA). The studies were conducted in different countries and had small sample sizes (which is accounted for in the low GRADE rating. Likewise, for mFG>8 vs Free testosterone and mFG>8 vs DHEAS, mFG>8 vs Androstenedione there is not adequate paired sensitivity and specificity and a lot of difference in sensitivity/specificity existed between the studies.

Studies including Ferriman Gallwey>9 compared to Total Testosterone, FAI and DHT showed good sensitivity and specificity; however, these results were from one study (Kumar 2022). One study (Yang 2020) used LC-MC/MC as the reference for total testosterone but included only area under the curve (AUC) data, which showed various results of mFG>5 or sFG>3 when compared to different reference standards. When LC-MC/MC was used to measure the reference standards the AUC for mFG or sFG was notably higher. All studies had a moderate risk of bias and the overall GRADE rating for all outcomes was very low. All other comparisons were single studies at high or moderate risk of bias. Evidence for all outcomes was very low

quality mainly due to imprecision (single or small study design), risk of bias and inconsistency.

The second EBR suggests that “female pattern hair loss and acne in isolation are relative weak predictors of biochemical hyperandrogenism”. The IG report that there is a lack of reference ranges for both acne and female pattern hair loss (FPHL) which makes comparison between studies difficult, particularly as FPHL also presents differently depending on ethnicity. One study found alopecia compared to free testosterone or total testosterone and acne compared to free testosterone were not high for paired sensitivity and specificity. The IG concluded that the predictive value of acne alone is low, and FLPH with hirsutism is moderate, for clinical hyperandrogenism. This uncertainty in the evidence does not seem to correlate with the strength of recommendation.

The evidence overall is not that clear or strong: there are few studies per tool or per-cut-off; there are small sample sizes; meta-analysis could not be conducted; and all results had a very low GRADE rating. The IG review looked at the exact cut-off for defining what is ‘abnormal’ for the various signs and symptoms however there is not enough evidence within the review to conclude. Cut-off values differ by ethnicity, so it is likely to be dependent on the population. The IG states that the evidence is limited for cut-offs, but that this may be answered by the upcoming PUPP study (an International IPD meta-analysis). The phenotype in unselected populations study: ethnic variation in population based normative cut-offs for defining hirsutism (PUPP) study has since published, reporting that cut-offs for PCOS diagnostic features varied which may impact on diagnostic accuracy and consistency. The PUPP study highlighted that cut offs for ovarian volume and FNPO can vary by both ultrasound technology used and approach to interpreting those results. The PUPP study states that this supports the high priority research areas identified by the international guideline and emphasises the need for a standardised approach to diagnosis.

Implementation considerations relate mainly to enhanced education of both healthcare workers and the general population to help identify an individual with

clinical signs of hirsutism as part of routine health assessments. The frequent self-treatment of hirsutism and the impact this will have on the assessment of clinical hyperandrogenism was also noted by the IG.

IG economic evidence

No health economic evidence was identified in the International Guideline (IG) for review question 1.3 on clinical hyperandrogenism.

The IG noted that the cost of implementing their recommendations will vary depending on the extent of healthcare practitioner and public education – noting that if healthcare practitioners are already aware of the signs and symptoms of clinical hyperandrogenism, costs will likely be minimal.

The IG noted the high certainty regarding the low costs of using visual scales for hirsutism, acne and female pattern hair loss (FPHL). However, they did note that if these visual scales are not already employed in current practice, the cost of widespread education regarding their use is unclear.

The IG also stated that the low cost of implementing visual scales for the assessment of hirsutism, acne and FPHL strongly suggests that the option of using signs of clinical hyperandrogenism instead of (more costly) biochemical measures of hyperandrogenaemia is cost-effective and would therefore have a positive effect on care for the underserved, under-resourced population. However, no formal economic analysis or additional qualitative discussion was provided to support their claims of cost-effectiveness.

1.3.2 NICE economic evidence

Included studies

A single health economic search was performed by NICE to identify published economic evaluations of relevance to all review questions in this guideline. See the literature search strategy in Appendix A .

No economic studies were identified which were applicable to this review question (see economic study selection flow chart in Appendix B).

Excluded studies

No economic studies were reviewed at full text and excluded from this review.

Economic model

No original health economic modelling was conducted for review question 1.3 on clinical hyperandrogenism. This review question was concerned with ascertaining the most effective measure to clinically diagnose PCOS related hyperandrogenism. The committee acknowledged that clinical assessments undertaken to assess for clinical hyperandrogenism are likely to be very similar in cost and therefore cost-effectiveness can be deduced from the clinical evidence alone.

1.3.3 NICE recommendations

The relevant recommendations for this section are Rec 1.4.1, 1.4.4 and 1.4.5.

1.3.4 The committee's discussion and interpretation of the evidence

Clinical

The committee decided not to include EBR 1.3.1 from the international guideline, as they did not feel there was enough evidence that hirsutism alone was predictive of biochemical hyperandrogenism or PCOS in women. The terminology of "terminal hairs" was discussed, primary care clinicians felt that this terminology was rarely used in the UK and as such was not considered helpful for women with hirsutism.

The committee did not contextualise the CRs from the IG regarding the use of the modified Ferriman Gallway (mFG) score as this is rarely used in clinical practice in the UK and is generally limited to research settings only. The committee felt it was important to emphasise that a woman's lived experience with hirsutism was a better approach than standardised measurements such as the mFG. However, the committee did not feel there was evidence to make a "do not use" recommendation

for the mFG score. Committee members also highlighted that different clinicians would rate hirsutism differently depending on their experience and interpretation of the mFG scoring system which could add to a woman's distress if it was not felt they met the stated score.

The committee did not include EBR 1.3.2, instead highlighting that anyone with PCOS should be assessed for hair loss to see if it follows an androgen-dependant pattern, and acne. The committee felt that GPs can deal with any issues related to acne in women with PCOS or suspected PCOS as it is an issue that presents often. They also referenced NICE guideline NG198 acne vulgaris: management as a useful tool for this area. It was also highlighted that there are established scales for acne that are more well defined compared to what is available for scalp hair loss. A cross reference to NICE guideline NG198 was added to the section on acne, NICE recommendation 1.11.16.

The committee felt it was important to add a description of symptoms that would require an urgent referral to endocrinology to rule out certain causes, however as this was adapted from IG recommendation 1.7.4 it is covered in section 1.7 of this report.

The committee chose to adapt two CRs (1.3.3 and 1.3.4) and two practice points (1.1.1 and 1.3.7) as they felt they had an important impact on the mental health of women in this area.

Health economic

No health economic evidence was included in the IG for evidence review 1.3 on clinical hyperandrogenism. In addition, no health economic evidence was identified in the health economic literature search conducted by NICE.

The IG noted the low costs of using visual scales for the assessment of hirsutism, acne and female pattern hair loss (FPHL) stating that this strongly suggests assessing the signs of clinical hyperandrogenism – instead of biochemical measures – is cost-effective and will therefore have a positive effect on care for the

underserved, under-resourced population. The committee discussed this statement, acknowledging that no formal economic analysis or additional qualitative discussion was provided to support the IG's claims of cost-effectiveness.

The committee also highlighted the IG's reference to the positive effects of assessing clinical hyperandrogenism for an under-resourced population. Within an NHS context, however, the committee concluded that testing for biochemical hyperandrogenism, alongside other blood tests to exclude other causes, is not only cost-effective, but also allows healthcare professionals to rule out other rare causes of high testosterone, for example certain tumours in either the ovaries or adrenal gland or congenital adrenal hyperplasia. For more information on the committee's discussion of biochemical hyperandrogenism and their rationale for cost-effectiveness, please see the committee discussion of the evidence in the Biochemical hyperandrogenism section.

The committee acknowledged that all the contextualised recommendations were either reflective of current practice, for example the IG's recommendation 1.3.3 which was concerned with assessing for signs and symptoms of clinical hyperandrogenism, or were concerned with providing information which could be useful to healthcare professionals and people with PCOS to improve care. The committee therefore concluded that no significant resource impact is anticipated as a result of the recommendations made for this review question.

1.4 Ultrasound and polycystic ovarian morphology

Review question 1.4: What is the most effective ultrasound criteria to diagnose PCOS?

1.4.1 Recommendations from the International evidence-based guideline for PCOS*

Evidence-based recommendations:

1.4.1 Follicle number per ovary (FNPO) should be considered the most effective ultrasound marker to detect polycystic ovarian morphology (PCOM) in adults.

1.4.2 Follicle number per ovary (FNPO), follicle number per cross-section (FNPS) and ovarian volume (OV) should be considered accurate ultrasound markers for PCOM in adults.

Consensus recommendations:

1.4.3 PCOM criteria should be based on follicle excess (FNPO, FNPS) and/or ovarian enlargement (OV).

1.4.4. Follicle number per ovary (FNPO) ≥ 20 in at least one ovary should be considered the threshold for PCOM in adults.

1.4.5 Ovarian volume (OV) ≥ 10 ml or follicle number per section (FNPS) ≥ 10 in at least one ovary in adults should be considered the threshold for PCOM if using older technology or image quality is insufficient to allow for an accurate assessment of follicle counts throughout the entire ovary.

Practice points:

1.4.6 There are no definitive criteria to define polycystic ovary morphology (PCOM) on ultrasound in adolescents, hence it is not recommended in adolescents.

1.4.7 When an ultrasound is indicated, if acceptable to the individual, the transvaginal approach is the most accurate for the diagnosis of PCOM.

1.4.8 Transabdominal ultrasound should primarily report ovarian volume (OV) with a threshold of ≥ 10 ml or follicle number per section (FNPS) ≥ 10 in either ovary in adults given the difficulty of assessing follicle counts throughout the entire ovary with this approach.

1.4.9 In patients with irregular menstrual cycles and hyperandrogenism, an ovarian ultrasound is not necessary for PCOS diagnosis.

1.4.10 Thresholds for PCOM should be revised regularly with advancing ultrasound technology, and age-specific cut-off values for PCOM should be defined.

1.4.11 There is a need for training in careful and meticulous follicle counting per ovary and clear standardised protocols are recommended for PCOM reporting on ultrasound including at a minimum:

- last menstrual period (or stage of cycle)
- transducer bandwidth frequency
- approach/route assessed
- total number of 2–9 mm follicles per ovary
- measurements in three dimensions (in cm) or volume of each ovary
- other ovarian features and/or pathology including ovarian cysts, corpus lutea, dominant follicles (≥ 10 mm) (which should not be included in ovarian volume calculations)
- reliance on the contralateral ovary FNPO for diagnosis of PCOM, where a dominant follicle is noted
- uterine features and/or pathology including endometrial thickness and pattern.

IG clinical evidence

Critical appraisal - ROBIS systematic review checklist

Section	Question	Answer
Study eligibility criteria	Did the review adhere to pre-defined objectives and eligibility criteria?	Yes *(PICO suitably detailed)*
Study eligibility criteria	Were the eligibility criteria appropriate for the review question?	Yes *(seem appropriate for review)*
Study eligibility criteria	Were eligibility criteria unambiguous?	Probably yes *(eligibility criteria clearly described in PICO, however RCT and other types of study design are included, whereas these would be different types of review)*
Study eligibility criteria	Were all restrictions in eligibility criteria based on study characteristics appropriate (e.g. date, sample size, study quality, outcomes measured)?	Unclear *(restrictions on date present but not well described, the search seems to be limited for studies published 1990 to present but no explanation why)*
Study eligibility criteria	Were any restrictions in eligibility criteria based on sources of information appropriate (e.g. publication status or format, language, availability of data)?	Probably yes *(Limited to English language is appropriate. Study types selected are appropriate for review question, however NICE guideline reviews would de-prioritise two-gate case control studies).*
Study eligibility criteria	Concerns regarding specification of study eligibility criteria	Low *(Most results were yes or probably yes, PICO was clearly defined, however no mention of why search date criteria went back to 1990 when most other reviews in the IG were much shorter time frames.)*
Identification and selection of studies	Did the search include an appropriate range of databases/electronic sources	Yes *(5 named databases appropriate to review)*

Section	Question	Answer
	for published and unpublished reports?	
Identification and selection of studies	Were methods additional to database searching used to identify relevant reports?	Probably yes *(does not detail any additional searches/manual search strategies however the general methods says that they sought relevant papers from their experts)*
Identification and selection of studies	Were the terms and structure of the search strategy likely to retrieve as many eligible studies as possible?	No *(The blocks for ovarian morphology and measurements would seem to narrow the search down too much and there is a risk of excluding relevant studies. A structure of PCOS and ultrasound and diagnosis terms would seem more appropriate. Use of DTA filters have been shown to miss relevant studies.* *The Medline strategy only is shown. There is no information on the translations beyond "similar terms appropriately translated".* *Pre-selected limits of female, humans and journal articles are used. This is unusual and there is a risk of inconsistency in indexing)*
Identification and selection of studies	Were restrictions based on date, publication format, or language appropriate?	Probably yes *(Follows protocol but unusual to use the preselected journal article limit)*
Identification and selection of studies	Were efforts made to minimise error in selection of studies?	Yes *(articles reviewed at title and abstract by three reviewers in consultation with evidence team, full text was used where decisions could not be made)*
Identification and selection of studies	Concerns regarding methods used to identify and/or select studies	High *(The structure of the search seems too narrow. There is no information given about*

Section	Question	Answer
		the databases searched other than Medline)
Data collection and study appraisal	Were efforts made to minimise error in data collection?	Yes *(3 independent reviewers plus additional input from the evidence team and structured extraction form)*
Data collection and study appraisal	Were sufficient study characteristics available for both review authors and readers to be able to interpret the results?	Yes *(Yes study characteristics are provided in an evidence summary table and in a full extraction form)*
Data collection and study appraisal	Were all relevant study results collected for use in the synthesis?	Yes *(PRISMA diagram states 27 studies included in SR, 23 in meta-analysis and grade profiles available for 27 studies. 2x2 data provided for sensitivity, specificity and AUC)*
Data collection and study appraisal	Was risk of bias (or methodological quality) formally assessed using appropriate criteria?	Yes *(risk of bias completed for each outcome using GRADE tables and a structured form but no assessment for each domain of bias just an overall rating)*
Data collection and study appraisal	Were efforts made to minimise error in risk of bias assessment?	Yes *(three reviewers plus evidence team for study selection and appraisal)*
Data collection and study appraisal	Concerns regarding methods used to collect data and appraise studies	Low *(All scoring was "Yes", no issues noted for this section)*
Synthesis and findings	Did the synthesis include all studies that it should?	Yes *(PRISMA diagram showed all 27 included studies had GRADE profiles, 3 studies were not suitable for inclusion in meta-analysis)*
Synthesis and findings	Were all pre-defined analyses reported, or departures explained?	Yes

Section	Question	Answer
		2x2 data, sensitivity, specificity and AUC provided as stated in the PICOs.
Synthesis and findings	Was the synthesis appropriate given the nature and similarity in the research questions, study designs and outcomes across included studies?	*Unclear* *(They meta-analysed and conducted GRADE on studies which had various optimal cut-offs or age-related cut-offs, or phenotype specific cut-offs but did not report a summary statistic. They did report the SROC which is probably appropriate given the different cut-off points, and if used HSROC model. However, there were differing reference standards (NIH and Rotterdam) and review 1.6 says "it has been shown that diagnosing according to Rotterdam criteria - endorsed by the 2018 International Guideline for PCOS. leads to a 1.5 times prevalence over NIH criteria." Different study designs were meta-analysed, but no sensitivity analysis conducted for this)*
Synthesis and findings	Was between-study variation (heterogeneity) minimal or addressed in the synthesis?	Probably yes *(GRADE assesses there to be no inconsistency. Random effects analysis is used for all sensitivity and specificity analyses although possibly default as DTA review and may assume heterogeneity inherent)*
Synthesis and findings	Were the findings robust, e.g. as demonstrated through funnel plot or sensitivity analyses?	Probably yes *(no funnel plot reported but was downgraded in GRADE for funnel plot asymmetry so must have been conducted)*
Synthesis and findings	Were biases in primary studies minimal or addressed in the synthesis?	No information *(GRADE table present, all studies serious or very serious risk of bias, limited text as to how this was considered, no sensitivity analyses for different study types)*
Synthesis and findings	Concerns regarding the synthesis and findings	Unclear *(unclear as to whether bias was addressed from different study designs as no sensitivity analysis. No summary statistic*

Section	Question	Answer
		but probably appropriate given the model type used, because of multiple cut-offs, but more details would be useful)
Judging risk of bias	Concerns regarding specification of study eligibility	Low *(PICO was clearly defined, however no mention of why search date criteria went back to 1990 when most other reviews in the IG were much shorter time frames)*
Judging risk of bias	Concerns regarding methods used to identify and/or select studies	High *(The structure of the search seems too narrow. There is no information given about the databases searched other than Medline)*
Judging risk of bias	Concerns regarding methods used to collect data and appraise studies	Low *(All scoring was "Yes", no issues noted for this section)*
Judging risk of bias	Concerns regarding the synthesis and findings	Unclear *(unclear as to whether bias was addressed from different study designs as no sensitivity analysis. No summary statistic but probably appropriate given the model type used, because of multiple cut-offs, but more details would be useful)*
Overall review ratings	Overall risk of bias	Unclear *(study eligibility and data collection are acceptable, although analysis needed more details and there were issues with the searches)*
Overall review ratings	Applicability as a source of data	Fully applicable

Evidence to recommendations justification: Both EBRs and all three CRs are highlighted as a strong recommendation for the option in the IG. The IG highlights that all criteria for PCOM to date are limited as it does not consider age, life status, and a lack of standardised ultrasound measurements. The GRADE certainty for the

comparison was low. Fifteen studies were included in the meta-analysis for FNPO diagnostic accuracy in women with or without PCOS, with 11 studies providing an optimal diagnostic threshold for FNPO. Studies varied on their diagnostic criteria for PCOS, with 9 using the NIH criteria, 4 using Rotterdam criteria and one using AE-PCOS criteria which could account for variation in results. One study had a low risk of bias, 10 were moderate and 4 had high risk of bias.

For the outcome of follicle number per cross sectional image (FNPS), 4 studies including women with and without PCOS were included in the meta-analysis, with 3 studies providing an optimal diagnostic threshold. These 4 studies all used the NIH criteria to diagnose PCOS.

For the outcome of additional ovarian morphology features, 7 studies proposed features not currently covered by the IG. Meta-analysis was not conducted due to moderate/high risk of bias, low quality of evidence, observational study design and lack of comparisons with previously established thresholds. As such a consensus recommendation was made which appears appropriate. For the practice points, the majority relate to training and increasing patient willingness to attend ultrasound scans and ensuring they are given a choice of method where appropriate, which should not present additional implementation issues.

FNPO had a variety of studies included in the forest plots, but no summary statistic, but a SROC graph was provided. This is possibly due to the variety of cut-off points included within the analysis and the various types of ultrasound measurement are used (TVUS 2D real-time, 3D-offline, 2D-offline, TAUS, TVUS/TRUS (2D real-time). Most of the studies provided an optimal cut-off, two had an age-related cut-off, one proposed 2D and 3D thresholds and one had a phenotype-specific cut-off. There was also a variety of ways to diagnose people with PCOS, most using NIH, then Rotterdam and one used AE-PCOS criteria. FNPO was proposed as the most accurate diagnostic accuracy marker for PCOS in adult women. Sensitivity ranged from 70% to 95% and specificity ranged from 75% to 100%. Area under the curve was consistently high with all but one result over 90%.

FNPS had fewer studies pooled but sensitivity ranged from 57% to 100% and specificity 60% to 100%. One study (Kim 2017) made up 6 parts of the forest plot as they included different age bands with differing cut-off points, 3 other studies were included in the meta-analysis. Method of measurement also varied from 2D Real time, 2D offline, 3D offline. Area under the curve was over 80% for all studies. Peripheral follicle distribution pattern had only one study with very low sensitivity of 29% and specificity of 98%.

Ovarian volume included a lot of studies which ranged in sensitivity and specificity from 50% to 100% and AUC ranging from 61% to 96%.

An adolescent subgroup (3 studies) had sensitivity ranging from 74 to 90% and specificity ranging from 79 to 92% and for NIH PCOS diagnosis only. For NIH PCOS diagnosis and TAUS only two studies had sensitivity ranging from 77 to 90% and specificity from 79 to 94%.

The outcomes of ovarian area, ovarian to uterine index, stromal area, stromal to ovarian area, stromal thickness, stromal index, and stromal strain ratio all had one or two studies showing good sensitivity and some good specificity with some low values.

Therefore, the EBR ‘follicle number per ovary (FNPO) should be considered the most effective ultrasound marker to detect polycystic ovarian morphology (PCOM) in adults’ is supported by the evidence. ‘Follicle number per ovary (FNPO), follicle number per cross-section (FNPS) and ovarian volume (OV) should be considered accurate ultrasound markers for PCOM in adults’ has less clear sensitivity and specificity range. The area under the curve results were supportive although only based on 4 studies. Ovarian volume had lower sensitivity in some studies but good area under the curve results to support them as markers.

IG economic evidence

No health economic evidence was identified in the International Guideline (IG) for review question 1.4 on ultrasound and polycystic morphology.

1.4.2 NICE economic evidence

Included studies

A single health economic search was performed by NICE to identify published economic evaluations of relevance to all review questions in this guideline. See the literature search strategy in Appendix A.

One economic study was identified which was applicable to this review question (see economic study selection flow chart in Appendix B).

This UK study compared Transvaginal Ultrasound (TVUS) to Anti-mullerian hormone (AMH) in people with suspected PCOS (Garay, 2025). Characteristics of the study are summarised in **Table 1**. Full details of this study are provided in the economic evidence study extraction tables in Appendix C.

An original health economic analysis was also conducted by NICE for this review question.

Excluded studies

No economic studies were reviewed at full text and excluded from this review.

Table 1: Summary characteristics of included study and original health economic analysis

Study details	Study design and type of analysis	Population	Interventions and comparators	Perspective	Primary outcome	Time horizon
Garay 2025 UK	Study design: Decision analytic model (decision tree) for a simulated cohort of women aged 25–45 years Source of effectiveness data: Liu 2021, Gabrielli 2012, Lizneva 2016, HARMONIA (internal analysis by Roche to estimate the proportion of people who do not present with clinical hyperandrogenism but do have biochemical hyperandrogenism), Riestenberg 2022, Costa 2012, Lindström 2013 Sensitivity & specificity of AMH *APHRODITE study* (Dietz de Loos, 2021)	Women with suspected PCOS	**Intervention 1:** Transvaginal Ultrasound (TVUS) – which was requested if after an initial GP appointment only irregular menstrual cycles or hyperandrogenism was confirmed to assess polycystic ovary morphology (PCOM). **Intervention 2:** Anti-Müllerian hormone (AMH) – in this case the Elecsys AMH plus immunoassay using a cut-off of 3.2bg/mL for identifying PCOM. An AMH test was conducted when a person with suspected PCOS presented an initial GP appointment.	NHS/PSS	Cost of a PCOS diagnosis	Lifetime

Study details	Study design and type of analysis	Population	Interventions and comparators	Perspective	Primary outcome	Time horizon
	Sensitivity and specificity of TVUS – meta-analysis based on IG Type of analysis: cost-consequence					
NICE 2026	Study design: Decision analytic model consisting of a one-year decision tree and life-time horizon Markov model Type of analysis: cost utility analysis	Adults with suspected signs and symptoms of PCOS	**Intervention**: Anti-müllerian hormone test **Comparator:** Transvaginal ultrasound scan	NHS/PSS	QALYs (SF-36 values mapped to EQ-5D-3L)	Lifetime

Abbreviations: AMH: anti-müllerian hormone; GP: general practitioner; NHS: national health service; PCOM: polycystic ovarian morphology; PCOS: polycystic ovarian syndrome; PSS: personal social services; QALYs: quality-adjusted life years TVUS: transvaginal ultrasound; UK: United Kingdom.

Summary of economic evidence

See **Table 2** for a summary of the economic evidence. The economic evidence study extraction table can be found in Appendix C and the model write-up for the original health economic analysis can be found in supplement A.

Table 2: Economic evidence summary table: TVUS versus AMH

Study	Applicability and limitations	Incremental cost	Incremental effects	Cost effectiveness	Uncertainty	Economic evidence statement
Garay 2025 (UK)	Partially applicable[1] Potentially serious limitations[2]	**Total costs of PCOS per year (per diagnosis):** TVUS: £1,536 AMH: £1,514 Incremental (2−1): Cost saving of £22 (95% CI: NR; p=NR) **Total costs of PCOS based on those who presented to their GP with signs and**	**True positive** TVUS: 6,584 AMH: 6,721 Incremental (2−1): 136 (95% CI: NR; p=NR) **False positive** TVUS: 344 AMH: 697 Incremental (2−1): 353 (95% CI: NR; p=NR) **True negative** TVUS: 5,478 AMH: 5,125 Incremental (2−1): 353 fewer	£22 cheaper for a diagnosis of PCOS with AMH (compared to TVUS)	Deterministic analysis conducted altering the base case values to an upper and lower limit to assess the total cost differences for TVUS and AMH – AMH was cost saving in all scenarios apart from when the sensitivity of TVUS was increased. The results are presented graphically so it is unclear at exactly what point TVUS becomes the cheaper intervention. A number of scenario analysis were conducted. The one with the greatest impact on results was with TVUS for AFC (base-case)	AMH was cheaper when compared to TVUS

Study	Applicability and limitations	Incremental cost	Incremental effects	Cost effectiveness	Uncertainty	Economic evidence statement
		symptoms of PCOS (per year) TVUS: £19,941,734 AMH: £19,657,705 Incremental (2−1): Cost saving of £284,029 Additional cost breakdown is presented in the Evidence table in Appendix A Cost year: 2022	true negatives with AMH (95% CI: NR; p=NR) **False negative** TVUS: 581 AMH: 444 Incremental (2−1): 136 fewer false negatives with AMH (95% CI: NR; p=NR)		versus with the AMH test (if test is required only after HA): this scenario assumed that the Elecsys AMH Plus immunoassay was required only after the first set of laboratory tests, and only for women requiring a PCOM diagnosis to rule-out or rule-in PCOS. Cost saving of £3.80 per PCOS diagnosis (per year) with AMH (HARMONIA data, cost saving of £7.10)	
NICE 2026	Directly applicable[3] Potentially serious limitations[4]	TVUS is £27 cheaper	TVUS results in additional QALY gain of 0.001	TVUS is dominant (more effective, less costly)	Probability	TVUS is the dominant intervention (less costly and more effective), but the differences in costs and QALYs are very small.

Study	Applicability and limitations	Incremental cost	Incremental effects	Cost effectiveness	Uncertainty	Economic evidence statement
						The NHB at £20,000 per QALY gained is 16.850 for TVUS and 16.847 for AMH.

Abbreviations: AMH: anti-müllerian hormone; GP: general practitioner; HA: hyperandrogenism; NHB: net health benefit; NHS: national health service; NR: not reported; PCOM: polycystic ovarian morphology; PCOS: polycystic ovarian syndrome; PSS: personal social services; QALYs: quality-adjusted life years; TVUS: transvaginal ultrasound; UK: United Kingdom

1. UK study, cost-consequence analysis (no QALYs), no mention of discounting
2. Analysis based on only one of six included studies from the IG, long-term outcomes modelled are cardiovascular risk and type two diabetes, people diagnosed with PCOS assumed to receive an expensive exercise intervention which is not reflective of UK current practice, no mention of PSA, analysis does not include the probability of requiring a TVUS after an AMH test
3. UK setting, cost-utility analysis
4. Costs do not include fertility costs for PCOS, SF-36 utility values from small study (mapped to EQ-5D-3L), several committee assumptions (e.g. the annual probability of obtaining a correct diagnosis)

Economic model

The use of transvaginal ultrasound (TVUS) and AMH for assessing polycystic ovary morphology (PCOM) was prioritised for original health economic work, and a subsequent health economic model was developed for this topic.

A cost-utility analysis was developed to assess the cost effectiveness of AMH compared to TVUS for identifying PCOM when required to ascertain a PCOS diagnosis. The population of the model was women, trans men and non-binary people aged 20. The analysis was undertaken from an NHS and Personal Social Services (PSS) perspective, following the NICE reference case. Although there was an existing health economic study published by Garay 2025, a decision was made for additional health economic work to be conducted as Garay 2025 only incorporated costs in their analysis and did not consider QALYs. The committee noted that recommending AMH would represent a change in UK current practice and therefore decided that supplementary evidence would help the determine optimal resource allocation in the NHS.

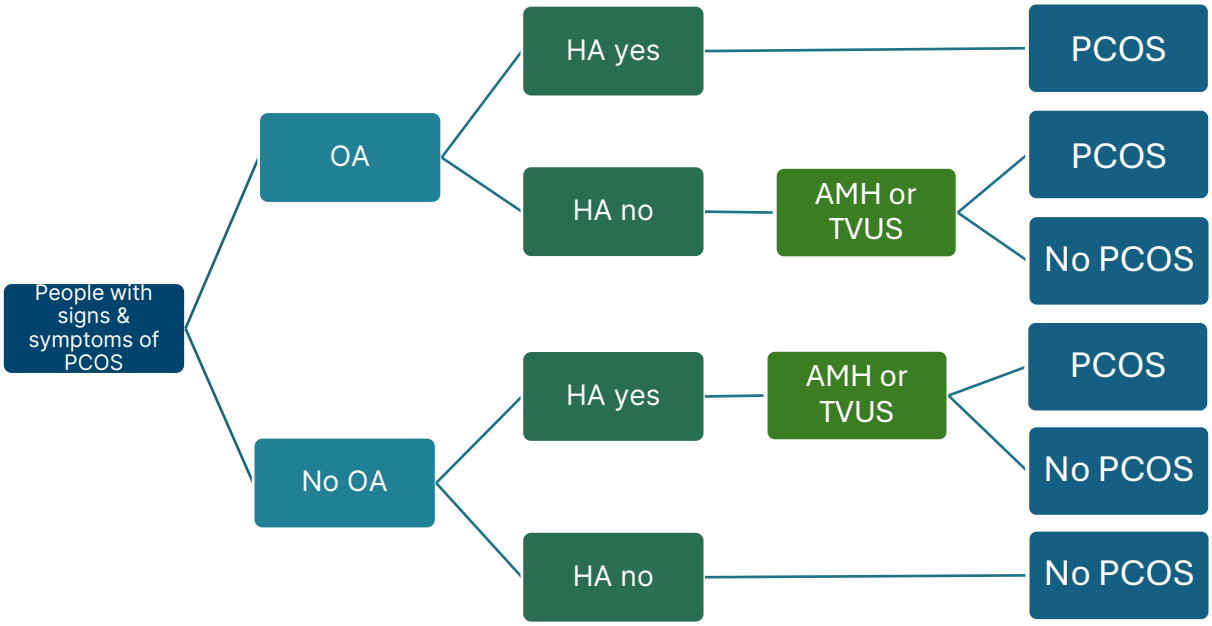
The model consisted of a one-year decision tree and a life-time horizon Markov model. The one-year decision tree was based on the structure employed in the included health economic study for this review question (Garay, 2025). Diagnostic accuracy for AMH and TVUS was obtained from the IG. Long-term outcomes were modelled based on a person either having true positive, false positive, true negative or false negative diagnosis after their evaluation for PCOS. For people who received an incorrect diagnosis (false positive, false negative), an annual probability of obtaining a true diagnosis was applied. The structure of the model can be found in Figure 1 (decision tree) and Figure 2 (long-term Markov model)

Prior to be people being categorised as having either oligomenorrhea or no oligomenorrhea (see Figure 1), people will have received an initial GP appointment, a set of blood tests and a follow-up GP appointment to discuss the results of these

blood tests. At this follow-up GP appointment, if the person is assessed to have oligomenorrhea and hyperandrogenism, no further costs are accrued in the decision tree. This is also true for those who do not have oligomenorrhea and hyperandrogenism.

For those meeting only one criterion of oligomenorrhea or hyperandrogenism, after their second GP appointment, this cohort will go onto receive either an AMH test or a TVUS, incurring the costs associated with these tests. In terms of QALYs, a utility was applied based on whether a correct or incorrect diagnosis was received.

Figure 1: Schematic diagram of the decision tree structure

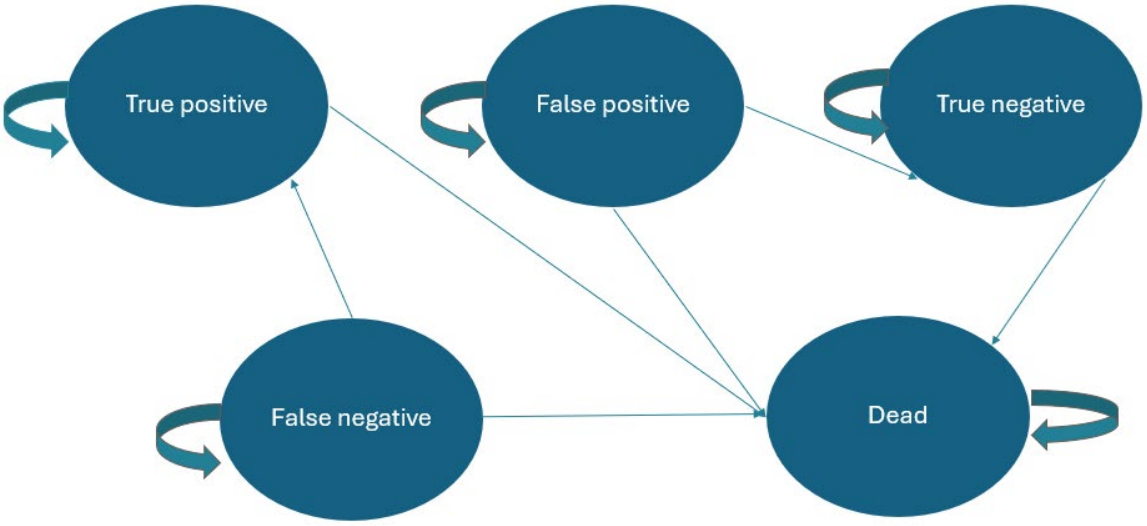


Abbreviations: AMH = Anti-Müllerian Hormone; HA = Hyperandrogenism; OA = Oligomenorrhea; PCOS = Polycystic ovarian syndrome; TVUS = Transvaginal ultrasound

Based on the sensitivity and specificity of AMH and TVUS, at the end of the one-year decision tree people could either have a received a true positive diagnosis, false

positive diagnosis, true negative diagnosis or false negative diagnosis (Figure 2). A health state cost was applied to these respective health states. Additional costs were also associated with obtaining a correct diagnosis. A utility value was also applied to these respective health states.

Figure 2: Schematic diagram of the life-time horizon Markov model



Probabilistic analysis was used as the reference case and various scenario analyses explored uncertainty.

The base-case analysis showed that TVUS was dominant (cheaper and more effective compared to AMH). In the majority of scenario analyses this conclusion remained unchanged.

Overall, the model results indicate that TVUS is the most cost-effective strategy. However, the difference in costs and QALYs were very small, with incremental costs being £27 and incremental QALYs being 0.001. Limitations of this analysis include; assuming the same mortality rate for those with PCOS and those without PCOS, not

capturing any potential additional benefits of TVUS over and above for diagnosing PCOS, various assumptions made regarding utilities and costs. The analysis also did not account for healthcare capacity for either test.

A summary of the guideline model characteristics is provided in **Table 1** the economic model evidence summary is shown in **Table 2**. The full model write up is provided in the supplementary document A.

1.4.3 NICE recommendations

The relevant recommendations for this section are Rec 1.7.5 to 10, 1,19.2

1.4.4 The committee's discussion and interpretation of the evidence

Clinical

The committee combined EBR 1.4.1 and CR 1.4.5 from the IG to make a recommendation regarding the use of follicle number per ovary as the preferred way to detect polycystic ovarian morphology when using a transvaginal ultrasound. However, they also concluded that follicle number per cross section and ovarian volume are more accurate if an abdominal ultrasound is required.

The committee discussed the use of ultrasound for diagnosing PCOM and highlighted that it provides more information to clinicians when treating women with PCOS than just the PCOM diagnosis and can be useful in treatment pathway planning. It was also described as reassuring for patients to have an ultrasound scan. The committee also discussed the potential limitations of AMH, as they noted that the assay used can make a difference. The committee also noted concern that AMH is also used as a marker of fertility and that this might be interpreted as such by service users and could create unnecessary anxiety. The committee also noted that AMH is a spectrum of results that vary greatly by age group, and as such should not be the sole method for diagnosing PCOS or PCOM and felt that AMH was included in the IG for those areas where ultrasound was less common. It was further discussed that when results were reported to clinicians from the laboratory, a reference range would be included to help clinicians interpret results in combination with the persons clinical presentation. As such the committee agreed to have either ultrasound or AMH as an option to allow clinician preference.

The committee agreed that only one method to diagnose was required, either ultrasound or AMH testing, but added that neither are required if a patient can be

diagnosed clinically with PCOS. The committee agreed with the IG that neither AMH nor ultrasound are reliable in people aged 10 to 17.

Health economic

One health economic study was identified that was applicable to this review question. This study was rated as partially applicable with potentially serious limitations.

The committee acknowledged that the recommendations in the IG on AMH and TVUS differed from current practice whereby AMH is not typically utilised in current practice for assessing PCOM. The IG recommendations stipulate that either AMH or TVUS could be used for identifying PCOM.

The included health economic study was presented to the committee, however due to the associated limitations of this study the committee were not confident in contextualising the IG's recommendations based on this health economic study alone. This was due to the potential change in clinical practice. The committee acknowledged that the clinical evidence in the IG illustrated that AMH has a relatively high sensitivity and specificity for identifying PCOM but noted that this was lower compared to TVUS. The committee also noted that AMH is the cheaper test of the two and that there could also be the potential for cost savings with AMH. However, the committee did have concerns with AMH regarding the lack of established cut-off threshold for identifying PCOM as well as noting the additional benefits of TVUS (such as identifying endometrial thickness as well as being able to diagnose other conditions). The committee therefore concluded that supplementary health economics would help aid their decision-making.

The committee discussed the health economic results of the guideline cost utility analysis, noting that TVUS is the dominant strategy (less costly and more effective) over a life-time horizon. The committee did, however, acknowledge that the differences in costs and effects were very small (£25 and 0.001 QALYs). The approach taken in the guideline health economic analysis slightly differed to that of

the one employed in Garay 2025. The guideline health economic analysis employed the same decision-tree structure as that employed by Garay 2025 and utilised the same data for the estimates of phenotype distributions. The data employed for diagnostic accuracy differed slightly. For AMH, Garay 2025 used a single study for diagnostic accuracy data whereas the guideline health economic analysis employed the meta-analysed outcomes from the included studies in the IG. For the diagnostic accuracy of TVUS, studies from the IG were meta-analysed that met our inclusion criteria for the model and aligned to NICE's methodology for meta-analyses. This resulted in a small discrepancy between the studies meta-analysed for NICE's analysis compared to studies meta-analysed by Garay 2025. However, the overall differences in sensitivity and specificity were minimal and these differing values were also tested in scenario analyses.

The primary way in which the two analyses differed was how long-term outcomes were captured and the application of QALYs for the guideline health economic analysis. In Garay 2025, long-term outcomes were captured by assuming a risk of type two diabetes and cardiovascular disease for those with PCOS. This risk was lowered when a correct diagnosis was obtained through the prescription of an exercise intervention. The committee discussed the high cost of this intervention (£363) and noted that this model assumption does not reflect current UK practice or align to the recommendations made as part of this guideline. In the NICE guideline health economic analysis costs were obtained from Berni 2024, which reported an annual cost of PCOS versus controls based on real-world evidence data. These costs were applied conservatively in the long-term model whereby the cost for PCOS was applied to all people in the model apart from those with a true negative diagnosis (where the cost for controls was applied). Utilities were also applied to corresponding health states (true positive, false positive, true negative, false negative). All utilities were assumed to be the same in the long-term Markov model apart from for those people who have a false negative. In the decision tree the same utility was applied for those people with false negative and a true positive to

represent untreated PCOS. Further details on the methodology and assumptions applied in the guideline health economic analysis can be found in supplementary material A.

Overall, the committee acknowledged that the guideline health economic analysis also had potentially serious limitations primarily due to assumptions made within the model due to a lack of available data. A full discussion of these can be found in model write-up in supplementary material A.

However, based on the results of the guideline health economic analysis, and the rest of the available evidence, the committee were confident that adopting the IG's recommendations would not result in a significant resource impact. The committee discussed that the cost of providing AMH was initially cheaper and over a life-time horizon there were small cost savings associated with providing TVUS. Therefore, there are no additional costs in the short-term associated with providing AMH. On the other hand, however, the committee did note that if the guideline increases the number of people looking to obtain a PCOS diagnosis this could result in an increase in costs for the NHS.

In general, the committee discussed the adoption of the IG's recommendations. The committee noted their concerns with AMH cut-off thresholds and the potential for diagnostic accuracy differences for different populations (e.g. age, weight, race). The committee also discussed some potential implementation costs for AMH, such as training for GPs. In addition, concerns were raised with respect to additional sub-fertility investigation costs that may be incurred as a consequence of conducting an AMH test. A new recommendation was therefore added to explain to people with PCOS that AMH levels alone do not predict people's chances of having a baby.

On the other hand, the committee acknowledged the long wait times for TVUS scans and therefore noted that AMH could have a place to play in helping with capacity issues and allow for quicker diagnoses. The committee, however, wanted to

emphasise that when signs and symptoms suggest another condition or a co-existing condition ultrasound will always be the most appropriate test to conduct to assess for PCOM (examples of such include pelvic pain or abnormal uterine bleeding such as fewer than 4 cycles a year). A recommendation was therefore made to reflect this. In instances where signs and symptoms do not suggest another condition or a co-existing condition the committee made a recommendation to offer either ultrasound or AMH.

The committee also wanted to highlight that it is not necessary to use both tests to assess for PCOM and doing so would be waste of NHS resources.

1.5 Anti-mullerian hormone in the diagnosis of PCOS

Review question 1.5: Is AMH effective for diagnosis of PCOS? Is AMH effective for diagnosis of PCOM?

1.5.1 Recommendations from the International evidence-based guideline for PCOS*

Evidence based recommendations:

1.5.1 Serum anti-mullerian hormone (AMH) could be used for defining PCOM in adults.

1.5.2 Serum AMH should only be used in accordance with the diagnostic algorithm, noting that in patients with irregular menstrual cycles and hyperandrogenism, an AMH level is not necessary for PCOS diagnosis.

1.5.3 We recommend that serum AMH should not be used as a single test for the diagnosis of PCOS.

1.5.4 Serum AMH should not yet be used in adolescents.

Practice points:

1.5.5 Either serum AMH or ultrasound may be used to define PCOM; however, both tests should not be performed to limit overdiagnosis.

1.5.6 Laboratories and healthcare professionals need to be aware of factors that influence AMH in the general population including:

- age: Serum AMH generally peaks between the ages of 20-25 years in the general population
- body mass index (BMI): Serum AMH is lower in those with higher BMI in the general population

- hormonal contraception and ovarian surgery: Serum AMH may be suppressed by current or recent COCP use
- menstrual cycle day: Serum AMH may vary across the menstrual cycle.

1.5.7 Laboratories involved in AMH measurements in females should use population and assay specific cut-offs.

IG clinical evidence

Critical appraisal - ROBIS systematic review checklist

Section	Question	Answer
Study eligibility criteria	Did the review adhere to pre-defined objectives and eligibility criteria?	Yes *(PICO has well described criteria)*
Study eligibility criteria	Were the eligibility criteria appropriate for the review question?	Yes *(Two protocols describe the two questions well and are appropriate for review questions)*
Study eligibility criteria	Were eligibility criteria unambiguous?	Yes *(as above, described well in PICO and appears appropriate for review. Reference standard could be clearer as it just states diagnosed PCOS. Diagnosis by NIH, Rotterdam or AES is in the population column)*
Study eligibility criteria	Were all restrictions in eligibility criteria based on study characteristics appropriate (e.g. date, sample size, study quality, outcomes measured)?	Probably yes *(restrictions seem reasonable with appropriate outcomes for a diagnostic accuracy review of sensitivity, specificity, AUC and ROC curves. No date limit. Wide range of study types including case control studies, which we would de-prioritise. RCTs included but this would be for a test-and-treat type review)*
Study eligibility criteria	Were any restrictions in eligibility criteria based on sources of information	Probably yes *(study type eligibility appears appropriate*

Section	Question	Answer
	appropriate (e.g. publication status or format, language, availability of data)?	*for review question, English language appropriate limit)*
Study eligibility criteria	Concerns regarding specification of study eligibility criteria	Low *(criterion appropriate for review)*
Identification and selection of studies	Did the search include an appropriate range of databases/electronic sources for published and unpublished reports?	Yes *(names 6 appropriate databases that have been used but PsycINFO seems superfluous)*
Identification and selection of studies	Were methods additional to database searching used to identify relevant reports?	Probably yes *(Other sources shown in PRISMA chart but does not detail any additional searches/manual search strategies however the general methods says that they sought relevant papers from their experts)*
Identification and selection of studies	Were the terms and structure of the search strategy likely to retrieve as many eligible studies as possible?	Yes *(search terms in full search strategy appear appropriate, subject headings translated between databases, no DTA filters used, combinations accurate)*
Identification and selection of studies	Were restrictions based on date, publication format, or language appropriate?	Probably yes *(Protocol says no date limits but there is reference to this being an updated review. Only described in full search strategy but appears appropriate, includes dates from previous review to date of this review)*
Identification and selection of studies	Were efforts made to minimise error in selection of studies?	Yes *(2 reviewers and evidence team completed study selection and appraisal using protocol. Full text retrieval used where decisions could not be made)*
Identification and selection of studies	Concerns regarding methods used to identify and/or select studies	Low *(Some details missing from search strategies but generally well conducted).*
Data collection and study appraisal	Were efforts made to minimise error in data collection?	Yes *(2 reviewers plus evidence team)*

Section	Question	Answer
Data collection and study appraisal	Were sufficient study characteristics available for both review authors and readers to be able to interpret the results?	Yes *(study characteristics table and data extraction table, which provided 2x2 data was available with sufficient information to aid interpretation of results)*
Data collection and study appraisal	Were all relevant study results collected for use in the synthesis?	Unclear *(Review text states that 81 studies were included (51 from new search and 28 from previous guideline), however the PRISMA diagram shows 53 were included in the systematic review and meta-analyses, and 51 in the GRADE tables)*
Data collection and study appraisal	Was risk of bias (or methodological quality) formally assessed using appropriate criteria?	Unclear *(risk of bias assessment has been completed for the 2 outcomes, however no detailed information about how this was undertaken is given. Risk of bias was assessed using a structured extraction form using QUADAS-2, but they do not downgrade by domains and gives only an overall assessment. Some of the extractions have a lot of 'no' for important questions but are still given moderate rating in the study extraction)*
Data collection and study appraisal	Were efforts made to minimise error in risk of bias assessment?	Yes *(2 reviewers for study inclusion and appraisal with evidence team for decision making if needed, can assume this was the same team used for completing risk of bias assessments)*
Data collection and study appraisal	Concerns regarding methods used to collect data and appraise studies	*Unclear* *(some inconsistencies with PRISMA and studies included in the SR and risk of bias seems to not downgrade enough)*
Synthesis and findings	Did the synthesis include all studies that it should?	Probably yes *(Study numbers is confusing in this review, PRISMA diagram states that 53 studies were included in the SA & MA, 51 in the GRADE tables, however the initial text stated 49 new studies were included.)*
Synthesis and findings	Were all pre-defined analyses reported or departures explained?	No information *(unable to determine due to a lack of*

Section	Question	Answer
		information about how they planned to do the analysis)
Synthesis and findings	Was the synthesis appropriate given the nature and similarity in the research questions, study designs and outcomes across included studies?	Probably yes *(Meta-analysis seems appropriate given the outcomes assessed. Hierarchical random effects models combined the estimates of sensitivity and specificity using a bivariate model. Case and cohort studies were combined as was development and validation studies, where we would de-prioritise case-control studies)*
Synthesis and findings	Was between-study variation (heterogeneity) minimal or addressed in the synthesis?	Probably yes *(There was a lot of heterogeneity, but a random effects model was used, and inconsistency was downgraded in GRADE. Subgroup analysis was conducted for 3 different assays, which did not resolve the heterogeneity. For the outcome of AMH as a diagnostic marker for PCOS there were serious inconsistencies for both adults and adolescents. For the outcome of AMH as a diagnostic marker for PCOM no serious inconsistency was noted. I^2 is also high on all but 1 forest plot)*
Synthesis and findings	Were the findings robust, e.g. as demonstrated through funnel plot or sensitivity analyses?	Yes *(sensitivity analysis has been completed for risk of bias. They stated that excluding studies with high risk of bias did not change the results. Funnel plots have been conducted and is documented in tables)*
Synthesis and findings	Were biases in primary studies minimal or addressed in the synthesis?	Unclear *(have used internal and external validation methods and used GRADE, however very little discussion. GRADE tables included but footnote says risk of bias is downgraded once when most of the evidence is at moderate or high risk of bias or downgraded twice as majority of evidence is at very high risk of bias, but there is also downgrading for imprecision and consistency, and overall GRADE is still moderate. No certainty or importance in row for adolescents' quality assessment)*

Section	Question	Answer
Synthesis and findings	Concerns regarding the synthesis and findings	Low *(GRADE ratings are unclear with regards downgrading, there is a lot of heterogeneity and case-control, and cohort studies are combined as are development and validation studies, however they do sensitivity analyses for risk of bias)*
Judging risk of bias	Concerns regarding specification of study eligibility	Low *(criterion appropriate for review)*
Judging risk of bias	Concerns regarding methods used to identify and/or select studies	Low *(A few things are unclear but generally well conducted).*
Judging risk of bias	Concerns regarding methods used to collect data and appraise studies	*Unclear* *(some inconsistencies with PRISMA and studies included in the SR and risk of bias seems to not downgrade enough)*
Judging risk of bias	Concerns regarding the synthesis and findings	Low *(GRADE ratings are unclear with regards downgrading, there is a lot of heterogeneity and case-control, and cohort studies are combined as are development and validation studies, however they do sensitivity analyses for risk of bias)*
Overall review ratings	Overall risk of bias	Low *(Some areas of study identification, inconsistencies in risk of bias downgrading are unclear)*
Overall review ratings	Applicability as a source of data	Fully applicable

Evidence to recommendations justification: The IG states that sensitivity and specificity are generally very good for adults but not for adolescents. The results of 64 studies were meta-analysed with a bivariate model and the pooled sensitivity for adults was 79% (95% CI 76% to 82%) and specificity was 87% (95% CI 84% to 89%)

for AMH as a substitute for PCOS diagnosis. For adolescents the sensitivity was 66% (95% CI 58% to 73%) and specificity 78% (95% CI 71% to 83%). However, there was a lot of heterogeneity in the results for adults with I^2 of 82% for sensitivity and 91% for specificity. Meta-analyses of children also showed high I^2 of 74% for sensitivity. The heterogeneity was not resolved with sensitivity or subgroup analyses except when different assays were sub-grouped. They found significant differences between automated, Elecsys and ELISA assays, with ELISA having higher sensitivity and specificity.

The heterogeneity was considered to be due to several issues with the study designs, such as the PCOS criterion used for diagnosing patients, and critical issues regarding patient population. Rotterdam 2003 criterion was used as the reference standard for nearly all adult studies, however the IG states it is not always an accurate measurement for ovarian morphology, as it relies on sonographer skills and equipment capabilities. There were different thresholds cut-offs for the studies, which the IG says was influenced by different AMH assays, differences in age, BMI and whether PCOS or PCOM was excluded in the control population, which were meta-analysed. Case control and cohort studies were both included in the analyses, of moderate or high risk of bias. Imprecision and inconsistency were also reported but the overall risk of bias in GRADE was moderate certainty, whereas in NICE guidance it would likely be graded very low certainty.

The outcome of AMH as a diagnostic marker for both PCOS and PCOM had a moderate GRADE outcome. AMH as a diagnostic marker for PCOM was meta-analysed in 6 studies to have sensitivity of 80% (72% to 86%) and specificity of 84% (79% to 88%), with no serious risk of bias, inconsistency, indirectness or imprecision and this was graded moderate certainty. However, this analysis included both the development and validation cohorts from the Dietz de Loos 2021 study, whereas NICE guidelines usually prioritise the validation cohort results. The EBRs for this section highlight that AMH can be a useful tool in diagnoses however they express an appropriate level of caution given the uncertainty of the included studies.

The practice points are unlikely to add additional demand compared to usual NHS process; however, they seem to add a level of ambiguity to the EBR by stating that "either serum AMH or ultrasound may be used to define PCOM". This is likely to require greater clarity due to the lack of conclusive evidence for serum AMH at present.

IG economic evidence

No health economic evidence was identified in the IG for review question 1.5 on anti-mullerian hormone for the diagnosis of PCOS.

1.5.2 NICE economic evidence

Included studies

Please see the NICE economic evidence section in section 1.4.2 of this report for details on the included study applicable to this review question. This study compared AMH to TVUS.

Excluded studies

No economic studies were reviewed at full text and excluded from this review.

Economic model

As noted in section 1.4.2 above, an original health economic model was developed for transvaginal ultrasound and anti-mullerian hormone for diagnosing polycystic ovary morphology. An overview of this model can be found in section 1.4.2 of this report.

The full model write up is provided in the supplementary document A. A summary of the guideline model characteristics and the economic model evidence summary is provided in Table 3 and Table 4 respectively. These tables can be found in section 1.4.2 of this report.

1.5.3 NICE recommendations

The relevant recommendations for this section are Rec 1.7.4, 1.7.5 and 1.7.10.

1.5.4 The committee's discussion and interpretation of the evidence

Clinical

The committee discussed that AMH is sometimes ordered as part of a group of tests and can add value in certain circumstances, particularly in fertility care settings. The committee explained how AMH levels are often higher in women with PCOS and has in some health settings been proposed as an alternative marker of PCOS. However, the committee also discussed that reference ranges can be difficult to interpret as they vary by age group, with primary care colleagues highlighting that this test is not standard practice in all areas, and as such there would be a staff education aspect to consider. Furthermore, other factors can influence the results such as those who have been on a combined oral contraceptive pill (COCP) long-term, age and variation in BMI. The committee agreed that AMH should not be encouraged as a marker of fertility and should be used only for assessing ovarian reserve in the appropriate secondary care setting. However, the committee agreed to include AMH as an option in NICE recommendations as some practitioners may prefer to use this test. The committee felt that AMH should be contextualised with recommendations on ultrasound scans to detect PCOM as there would be overlap between recommendations. The committee noted concerns that indicating a preference for ultrasound over AMH may place additional pressures on ultrasound services, where capacity is already limited, however they felt the evidence supported the contextualisation of this recommendation. The committee acknowledged that ultrasound is currently used most often, however there will be patients that have not been offered ultrasound or AMH, and as such the recommendation would encourage more clinicians to refer and more patients to ask for ultrasound. The committee contextualised IG recommendation 1.5.4 which states serum AMH should not be used in adolescents. This is due to it being less reliable, as serum AMH has poor

specificity in this age group. They agreed with the practice point that both AMH and ultrasound should not be used to check for PCOM as this would be a waste of resources.

Economic

The committee's disscussion of the evidence relating to both AMH and TVUS can be found in section 1.4.4 of this report.

1.6 Ethnic variation

Review question 1.6: In women with suspected PCOS, is there evidence of ethnic and geographic variations in prevalence?

1.6.1 Recommendations from the International evidence-based guideline for PCOS*

Evidence-based recommendations:

1.6.1 Healthcare professionals should be aware of the high prevalence of PCOS in all ethnicities and across world regions, ranging from 10 to 13% globally using the Rotterdam criteria.

1.6.2 Healthcare professionals should be aware that PCOS prevalence is similar across world regions and ethnicities but may be higher in Southeast Asian and Eastern Mediterranean regions.

Practice point:

1.6.3 Healthcare professionals should be aware that the presentation of PCOS may vary across ethnic groups.

IG clinical evidence

Critical appraisal - ROBIS systematic review checklist

Section	Question	Answer
Study eligibility criteria	Did the review adhere to pre-defined objectives and eligibility criteria?	Yes *(appears to meet well described criteria)*
Study eligibility criteria	Were the eligibility criteria appropriate for the review question?	Yes *(PICO describes eligibility criteria well and is appropriate for review question)*

Section	Question	Answer
Study eligibility criteria	Were eligibility criteria unambiguous?	Yes *(yes well described in PICO)*
Study eligibility criteria	Were all restrictions in eligibility criteria based on study characteristics appropriate (e.g. date, sample size, study quality, outcomes measured)?	Yes *(restrictions seem appropriate however are not well documented)*
Study eligibility criteria	Were any restrictions in eligibility criteria based on sources of information appropriate (e.g. publication status or format, language, availability of data)?	Yes *(study type is appropriate for review question)*
Study eligibility criteria	Concerns regarding specification of study eligibility criteria	Low *(eligibility criteria is appropriate for review)*
Identification and selection of studies	Did the search include an appropriate range of databases/electronic sources for published and unpublished reports?	Yes *(5 named databases are appropriate for this review except Psychinfo which is superfluous)*
Identification and selection of studies	Were methods additional to database searching used to identify relevant reports?	Probably yes *(mention of additional sources in PRISMA chart but does not detail any additional searches/manual search strategies however the general methods says that they sought relevant papers from their experts)*
Identification and selection of studies	Were the terms and structure of the search strategy likely to retrieve as many eligible studies as possible?	No *(Subject headings not translated from Medline to other databases.* *Block for ethnicity is missing relevant subject headings; free text terms are not applied consistently.* *There is a redundant OR line, but this should not exclude studies.* *Animal studies excluded with pre-selected limits rather than a tested method)*

Section	Question	Answer
Identification and selection of studies	Were restrictions based on date, publication format, or language appropriate?	Probably yes *(not well described other than in search strategy but appears relevant. Restrictions follow protocol)*
Identification and selection of studies	Were efforts made to minimise error in selection of studies?	Probably yes *(states 1 reviewer did each stage of study selection, screening on title and abstract and study appraisal but was not clear if it was the same reviewer for all 3 stages)*
Identification and selection of studies	Concerns regarding methods used to identify and/or select studies	High *(Subject headings not translated, ethnicity block poorly designed)*
Data collection and study appraisal	Were efforts made to minimise error in data collection?	Probably yes *(1 reviewer plus evidence team where required)*
Data collection and study appraisal	Were sufficient study characteristics available for both review authors and readers to be able to interpret the results?	Yes *(detailed study characteristics table available, Hoy prevalence checklist used)*
Data collection and study appraisal	Were all relevant study results collected for use in the synthesis?	Probably yes *(49 papers found, however 2 reported on results of 2 other studies, leading to 47 papers (45 studies). PRISMA diagram states 38 were included in meta-analysis and GRADE tables, 2 papers reported on the same study, 4 studies used ICD codes to diagnose PCOS and/or studies reporting on incidence or age-adjusted prevalence rates were excluded from the meta-analysis. One study unaccounted for)*
Data collection and study appraisal	Was risk of bias (or methodological quality) formally assessed using appropriate criteria?	Yes *(risk of bias completed for each study, JBI criteria used)*
Data collection and	Were efforts made to minimise error in risk of bias assessment?	Probably yes *(potentially if different reviewers did separate stages but difficult to determine)*

Section	Question	Answer
study appraisal		
Data collection and study appraisal	Concerns regarding methods used to collect data and appraise studies	Low *(Appropriate checklists used. Lacking in some detail about studies not included in meta-analysis and GRADE).*
Synthesis and findings	Did the synthesis include all studies that it should?	Probably yes *(All results appear present)*
Synthesis and findings	Were all pre-defined analyses reported or departures explained?	No information *(no information was given regarding the planned analysis of the results)*
Synthesis and findings	Was the synthesis appropriate given the nature and similarity in the research questions, study designs and outcomes across included studies?	Yes *(meta-analysis completed for 38 studies, sensitivity analysis also presented)*
Synthesis and findings	Was between-study variation (heterogeneity) minimal or addressed in the synthesis?	Yes *(forest plots available for prevalence grouped by ethnicity as well as a breakdown by individual region)*
Synthesis and findings	Were the findings robust, e.g. as demonstrated through funnel plot or sensitivity analyses?	Yes *(no funnel plot but sensitivity analysis present)*
Synthesis and findings	Were biases in primary studies minimal or addressed in the synthesis?	Probably yes *(likely yes but limited discussion on how conclusions have been reached)*
Synthesis and findings	Concerns regarding the synthesis and findings	Low *(limited details on how conclusions reached but synthesis seems appropriate)*
Judging risk of bias	Concerns regarding specification of study eligibility	Low *(eligibility criteria is appropriate for review)*
Judging risk of bias	Concerns regarding methods used to identify and/or select studies	High *(Subject headings not translated, ethnicity block poorly designed)*

Section	Question	Answer
Judging risk of bias	Concerns regarding methods used to collect data and appraise studies	Low *(Appropriate checklists used. Lacking in some detail about studies not included in meta-analysis and GRADE).*
Judging risk of bias	Concerns regarding the synthesis and findings	Low *(limited details on how conclusions reached but synthesis seems appropriate)*
Overall review ratings	Overall risk of bias	Low *(eligibility criteria and appropriate checklists used although lack of details of studies not included and conclusions reached)*
Overall review ratings	Applicability as a source of data	Fully applicable

Evidence to recommendations justification: Forty-five cross-sectional and cohort studies across 4 outcomes were included in this new systematic review. Studies varied in their use of PCOS diagnostic criteria: Rotterdam (2003), NIH (1990) or the Androgen Excess & PCOS society (2006) criteria. The reviewers found that the Rotterdam criteria resulted in a 1.5 times higher prevalence of PCOS diagnosis when compared to the NIH criteria. Four outcomes were included comparing the prevalence of PCOS in adult women from different ethnic backgrounds comparing all 3 diagnostic criteria, NIH alone, Rotterdam alone, and NIH plus Rotterdam for adolescent women. All outcomes had a very low GRADE rating except for prevalence using the Rotterdam criteria alone, which was low. Both EBRs are listed as strong recommendations for the option in the IG. There is also a practice point regarding awareness for health professionals regarding the presentation of PCOS across differing ethnic groups. They highlight that the prevalence of PCOS is similar in adolescent women of different ethnicities in the subgroup considerations. The IG reports that the prevalence of PCOS differs according to race, the highest prevalence of PCOS was seen in Black women (6.1% - NIH, 16% Rotterdam, 12.6% AE-PCOS)

and Middle Eastern women (6.1%- NIH), followed by Caucasian women (5.5% - NIH) and the lowest seen in Chinese women (5.6% - Rotterdam).

IG economic evidence

No health economic evidence was identified in the IG for review question 1.6 on ethnic variation.

1.6.2 NICE economic evidence

Included studies

A single health economic search was performed by NICE to identify published economic evaluations of relevance to all review questions in this guideline. See the literature search strategy in Appendix A.

No economic studies were identified which were applicable to this review question (see economic study selection flow chart in Appendix B).

Excluded studies

No economic studies were reviewed at full text and excluded from this review.

Economic model

No original health economic modelling was conducted for review question 1.6.

1.6.3 NICE recommendations

Relevant recommendation for this section is Rec 1.3.4.

1.6.4 The committee's discussion and interpretation of the evidence

Clinical

The committee decided to contextualise recommendations 1.6.1 and 1.6.2 from the IG to raise awareness that the overall prevalence of PCOS is high but may be higher in women of minority ethnic backgrounds. This was contextualised into

recommendation 1.3.4 of the NICE guideline as the committee felt it was more relevant to the section on when to suspect PCOS. It was revised not only to raise awareness among healthcare professionals, but to be actively considered when assessing symptoms or signs of PCOS. The committee also highlighted a study (Berni 2025) which supported the evidence from the IG that the prevalence of PCOS is higher in women from minority ethnic backgrounds.

Health economic

No health economic evidence was included in the IG for review question 1.6 on ethnic variation. In addition, no health economic evidence was identified in NICE's health economic literature search for this review question.

As the contextualised recommendations are concerned with prevalence of PCOS, and therefore informative for healthcare practitioners and other readers of NICE guidance, no resource implications are anticipated for the contextualised recommendations associated with this review on ethnic variation.

1.7 Menopause life stage

Review question 1.7: What is the post-menopausal phenotype of PCOS and how elevated should androgens be to indicate PCOS?

1.7.1 Recommendations from the International evidence-based guideline for PCOS*

Consensus recommendations:

1.7.1 A diagnosis of PCOS could be considered as enduring/lifelong.

1.7.2 Healthcare professionals could consider that both clinical and biochemical hyperandrogenism persist in the post menopause for women with PCOS.

1.7.3 PCOS diagnosis could be considered post menopause, if there is a past diagnosis, or a long-term history of oligo-amenorrhoea with hyperandrogenism and/or PCOM, during the earlier reproductive years (age 20-40).

1.7.4 Further investigations should be considered to rule out androgen-secreting tumours and ovarian hyperthecosis in postmenopausal women presenting with new-onset, severe or worsening hyperandrogenism including hirsutism.

IG clinical evidence

Critical appraisal - ROBIS systematic review checklist

Section	Question	Answer
Study eligibility criteria	Did the review adhere to pre-defined objectives and eligibility criteria?	N/A *(narrative review)*
Study eligibility criteria	Were the eligibility criteria appropriate for the review question?	N/A *(narrative review)*

Section	Question	Answer
Study eligibility criteria	Were eligibility criteria unambiguous?	N/A *(narrative review)*
Study eligibility criteria	Were all restrictions in eligibility criteria based on study characteristics appropriate (e.g. date, sample size, study quality, outcomes measured)?	N/A *(narrative review)*
Study eligibility criteria	Were any restrictions in eligibility criteria based on sources of information appropriate (e.g. publication status or format, language, availability of data)?	N/A *(narrative review)*
Study eligibility criteria	Concerns regarding specification of study eligibility criteria	N/A *(review was a level 4 narrative review, as such no study eligibility)*
Identification and selection of studies	Did the search include an appropriate range of databases/electronic sources for published and unpublished reports?	N/A *(narrative review)*
Identification and selection of studies	Were methods additional to database searching used to identify relevant reports?	N/A *(narrative review)*
Identification and selection of studies	Were the terms and structure of the search strategy likely to retrieve as many eligible studies as possible?	N/A *(narrative review)*
Identification and selection of studies	Were restrictions based on date, publication format, or language appropriate?	N/A *(narrative review)*
Identification and selection of studies	Were efforts made to minimise error in selection of studies?	N/A *(narrative review)*
Identification and selection of studies	Concerns regarding methods used to identify and/or select studies	N/A *(review was a level 4 narrative review, as such no search strategy or screening was completed)*
Data collection and	Were efforts made to minimise error in data collection?	N/A *(narrative review)*

Section	Question	Answer
study appraisal		
Data collection and study appraisal	Were sufficient study characteristics available for both review authors and readers to be able to interpret the results?	N/A *(narrative review)*
Data collection and study appraisal	Were all relevant study results collected for use in the synthesis?	N/A *(narrative review)*
Data collection and study appraisal	Was risk of bias (or methodological quality) formally assessed using appropriate criteria?	N/A *(narrative review)*
Data collection and study appraisal	Were efforts made to minimise error in risk of bias assessment?	N/A *(narrative review)*
Data collection and study appraisal	Concerns regarding methods used to collect data and appraise studies	N/A *(narrative review - no risk of bias assessment)*
Synthesis and findings	Did the synthesis include all studies that it should?	N/A *(narrative review)*
Synthesis and findings	Were all pre-defined analyses reported or departures explained?	N/A *(narrative review)*
Synthesis and findings	Was the synthesis appropriate given the nature and similarity in the research questions, study designs and outcomes across included studies?	N/A *(narrative review)*
Synthesis and findings	Was between-study variation (heterogeneity) minimal or addressed in the synthesis?	N/A *(narrative review)*
Synthesis and findings	Were the findings robust, e.g. as demonstrated through funnel plot or sensitivity analyses?	N/A *(narrative review)*

Section	Question	Answer
Synthesis and findings	Were biases in primary studies minimal or addressed in the synthesis?	N/A *(narrative review)*
Synthesis and findings	Concerns regarding the synthesis and findings	N/A *(narrative review- no studies to assess)*
Judging risk of bias	Concerns regarding specification of study eligibility	N/A *(review was a level 4 narrative review, and so did not have study eligibility)*
Judging risk of bias	Concerns regarding methods used to identify and/or select studies	N/A *(review was a level 4 narrative review, as such no search strategy or screening was completed)*
Judging risk of bias	Concerns regarding methods used to collect data and appraise studies	N/A *(narrative review- no study collecting or appraising)*
Judging risk of bias	Concerns regarding the synthesis and findings	N/A *(narrative review- no studies to assess)*
Overall review ratings	Overall risk of bias	N/A *(narrative review leading to consensus recommendations, stating little data on changes to PCOS diagnosis during the menopausal transition and post menopause. All consensus recommendations in this section have been given a GRADE direction and strength of recommendation rating as conditional (weak) recommendation for the option. They also state that implementation is likely to be acceptable to patients and clinicians as it is pragmatic)*
Overall review ratings	Applicability as a source of data	Partially applicable

Evidence to recommendations justification: This was a narrative review and as such no evidence was found for the systematic review, there were no EBRs, only CRs. The IG stated that narrative reviews were undertaken when a review question was not well suited for a systematic review, or if a systematic review found no evidence. The consensus recommendation suggested that there is little data on changes to PCOS diagnostic criteria during the menopause and for post-menopause,

however they note some indirect evidence on the impact of aging in pre-menopausal women.

IG economic evidence

No health economic evidence was identified in the IG for review question 1.7 on menopause life stage.

The IG noted that the recommendations made on menopause and life stage are pragmatic and therefore likely to be acceptable to patients and clinicians.

1.7.2 NICE Economic evidence

Included studies

A single health economic search was performed by NICE to identify published economic evaluations of relevance to all review questions in this guideline. See the literature search strategy in Appendix A.

No economic studies were identified which were applicable to this review question (see economic study selection flow chart in Appendix B).

Excluded studies

No economic studies were reviewed at full text and excluded from this review.

Economic model

No original health economic modelling was conducted for review question 1.7

1.7.3 NICE recommendations

The relevant recommendations for this section are Rec 1.4.3 and 1.4.12.

1.7.4 The committee's discussion and interpretation of the evidence

Clinical

The committee contextualised three of the four consensus recommendations (1.7.2 to 1.7.4) from the IG. This gives information to the "when to suspect PCOS" section by adding that PCOS can be diagnosed after menopause based on patient history between the ages of 20 and 40, or due to PCOM. The committee also highlighted in this section that androgen secreting tumours and ovarian hyperthecosis should be ruled out if post-menopausal women have worsening hyperandrogenism. The committee decided not to contextualise the IG's recommendation 1.7.1, which stated that a diagnosis of PCOS should be considered as lifelong, as they concluded that this information could be inferred from the additional recommendations they decided to contextualise.

Consensus recommendation 1.7.2 was adapted for the section on hyperandrogenism to highlight that both clinical and/or biochemical hyperandrogenism can still be present after menopause.

Health economic

No health economic evidence was included in the IG for review question 1.7 on menopause life stage. In addition, no health economic evidence was identified in the NICE's health economic literature search for this review question.

The committee acknowledged that the recommendations they contextualised on menopause and life stage are reflective of UK current practice and therefore not anticipated to result in a significant resource impact.

1.8 Cardiovascular disease risk

Review question 1.8: Are women with PCOS at increased risk for cardiovascular disease (CVD)?

1.8.1 Recommendations from the International evidence-based guideline for PCOS*

Evidence-based recommendations:

1.8.1 Women with PCOS should be considered at increased risk of cardiovascular disease and potentially of cardiovascular mortality, acknowledging that the overall risk of cardiovascular disease in premenopausal women is low.

1.8.2 All women with PCOS should be assessed for cardiovascular disease risk factors.

Consensus recommendations:

1.8.3 All women with PCOS, regardless of age and BMI, should have a lipid profile (cholesterol, low density lipoprotein cholesterol, high density lipoprotein cholesterol and triglyceride level) at diagnosis. Thereafter, frequency of measurement should be based on the presence of hyperlipidaemia and additional risk factors or global cardiovascular risk.

1.8.4 All women with PCOS should have blood pressure measured annually and when planning pregnancy or seeking fertility treatment, given the high risk of hypertensive disorders in pregnancy and the associated comorbidities.

1.8.5 Funding bodies should recognise that PCOS is highly prevalent with multisystem effects including cardiometabolic disease and should diversify and increase research support accordingly.

1.8.6 Cardiovascular general population guidelines could consider the inclusion of PCOS as a cardiovascular risk factor.

1.8.7 Healthcare professionals, women with PCOS and other stakeholders should all prioritise preventative strategies to reduce cardiovascular risk.

Practice point:

1.8.8 Consideration should be given to the differences in cardiovascular risk factors, and cardiovascular disease, across ethnicities (see 1.6.1) and age, when determining frequency of risk assessment.

IG clinical evidence

Critical appraisal - ROBIS systematic review checklist

Section	Question	Answer
Study eligibility criteria	Did the review adhere to pre-defined objectives and eligibility criteria?	Yes *(criteria is well described and appropriate)*
Study eligibility criteria	Were the eligibility criteria appropriate for the review question?	Yes *(eligibility well described by PICO and appropriate to review question)*
Study eligibility criteria	Were eligibility criteria unambiguous?	Yes *(PICO is clear and appropriately detailed)*
Study eligibility criteria	Were all restrictions in eligibility criteria based on study characteristics appropriate (e.g. date, sample size, study quality, outcomes measured)?	Probably yes *(restrictions not well discussed but appear appropriate, such as date limit 2017 to 2022 because it was an update of a previous systematic review.)*
Study eligibility criteria	Were any restrictions in eligibility criteria based on sources of information appropriate (e.g. publication	Yes *(study type eligibility appropriate for review question, such as English language and human studies are appropriate limits)*

Section	Question	Answer
	status or format, language, availability of data)?	
Study eligibility criteria	Concerns regarding specification of study eligibility criteria	Low *(All of the criteria were rated yes due to adequate details except for date limit)*
Identification and selection of studies	Did the search include an appropriate range of databases/electronic sources for published and unpublished reports?	Yes *(lists 5 names databases as sources, all are appropriate for review type, except PsycINFO superfluous)*
Identification and selection of studies	Were methods additional to database searching used to identify relevant reports?	Probably yes *(does not detail any additional searches/manual search strategies however the general methods says that they sought relevant papers from their experts)*
Identification and selection of studies	Were the terms and structure of the search strategy likely to retrieve as many eligible studies as possible?	No *(subject terms not translated from Medline to other databases.* *Search seems too narrow, they have combined (with and) blocks for PCOs, terms for tools/charts etc (no subject headings), risk, CVD and a line (sensitiv: or predictive value:).mp. or accurac:.tw. or specificit$.tw. If terms for tools are included it would seem more appropriate to do a combination of PCOS and tools or risk and CV disease with the sensitivity line omitted.* *It is unclear why terms for diabetes have been included.* *Risk concept appears to have been NOTed out of the CINAHL strategy* *Preselected limits for humans and female have been used. This relies on consistent auto indexing.* *appropriate terms included, no inappropriate restrictions noted)*

Section	Question	Answer
Identification and selection of studies	Were restrictions based on date, publication format, or language appropriate?	No *(Date limits have been confusingly applied. No reason apparent from the protocol why different date limits have been used. There is an incomplete NOT line in the OVID strategy meaning we cannot confirm accuracy)*
Identification and selection of studies	Were efforts made to minimise error in selection of studies?	Probably yes *(mentioned 1 reviewer for study selection criteria and 1 reviewer for title and abstract review, however it does not state if these were the same person)*
Identification and selection of studies	Concerns regarding methods used to identify and/or select studies	*High* *(Subject terms not translated. The search seems too narrow. Flawed NOTing of the risk concept from the CINAHL strategy. Date Limits possibly incorrectly applied)*
Data collection and study appraisal	Were efforts made to minimise error in data collection?	Probably yes *(1 reviewer for data extraction and review, whereas 2 would have been better)*
Data collection and study appraisal	Were sufficient study characteristics available for both review authors and readers to be able to interpret the results?	Yes *(study characteristics table available with sufficient information to help interpret results)*
Data collection and study appraisal	Were all relevant study results collected for use in the synthesis?	Probably yes *(text describes 25 studies meeting inclusion criteria, PRISMA lists 16 original papers, 4 systematic reviews and 8 additional papers identified from SR, 20 studies were included in GRADE tables)*
Data collection and study appraisal	Was risk of bias (or methodological quality) formally assessed using appropriate criteria?	Probably yes *(table detailing quality appraisal of studies included, includes selection bias, performance bias, detection bias, attrition bias, report bias and confounding. There is not much detail about controlling for confounders except question if similar at baseline)*

Section	Question	Answer
Data collection and study appraisal	Were efforts made to minimise error in risk of bias assessment?	Probably yes *(only 1 reviewer for each stage of review 2 would have been better to minimise error in ROB assessment)*
Data collection and study appraisal	Concerns regarding methods used to collect data and appraise studies	Low *(most were rated yes or probably yes, with sufficient detail presented on the number of studies included and for quality assessment)*
Synthesis and findings	Did the synthesis include all studies that it should?	Probably yes *(some outliers were not included but this seems reasonable)*
Synthesis and findings	Were all pre-defined analyses reported or departures explained?	No information. *(Subgroups in protocol were adolescents, ethnicity and phenotype, in GRADE the subgroups were defined PCOS (with NIH and Rotterdam analysed separately) and >10 years follow-up)*
Synthesis and findings	Was the synthesis appropriate given the nature and similarity in the research questions, study designs and outcomes across included studies?	Yes *(a meta-analysis/ descriptive analysis table was used, appears to be appropriate for the studies included)*
Synthesis and findings	Was between-study variation (heterogeneity) minimal or addressed in the synthesis?	Probably yes *(forest plots available for composite CVD odds ratio)*
Synthesis and findings	Were the findings robust, e.g. as demonstrated through funnel plot or sensitivity analyses?	Probably yes *(funnel plot available for CVD odds ratio, however results are a little mixed)*
Synthesis and findings	Were biases in primary studies minimal or addressed in the synthesis?	Probably yes *(Could have been further minimised by an additional reviewer. The GRADE ratings seem to start higher than we would rate them given the type of studies included in this review)*
Synthesis and findings	Concerns regarding the synthesis and findings	Probably yes

Section	Question	Answer
		(Synthesis seems appropriate and funnel plots conducted are mixed. GRADE ratings start higher than we would rate them)
Judging risk of bias	Concerns regarding specification of study eligibility	Low *(All of the criteria were rated yes due to adequate details except for date limit)*
Judging risk of bias	Concerns regarding methods used to identify and/or select studies	*High* *(Subject terms not translated. The search seems too narrow. Flawed NOTing of the risk concept from the CINAHL strategy. Date Limits possibly incorrectly applied.)*
Judging risk of bias	Concerns regarding methods used to collect data and appraise studies	Low *(most were rated yes or probably yes, with sufficient detail presented on the number of studies included and for quality assessment)*
Judging risk of bias	Concerns regarding the synthesis and findings	Low *(Synthesis seems appropriate and funnel plots conducted are mixed. GRADE ratings start higher than we would rate them)*
Overall review ratings	Overall risk of bias	Low *(all areas were low except for the search terms and strategy)*
Overall review ratings	Applicability as a source of data	Fully applicable

Evidence to recommendations justification: The IG reports that there is a high prevalence of cardiometabolic risk factors in women with PCOS, which highlights the need for greater CVD screening in this group, however despite research in the area, a specific CVD risk factor tool or instrument is not available. Currently, risk prediction tools such as Framingham Risk Score (FRS) and Pooled Cohort Equations (PCE) are used in the general population and whilst they might be useful, they have not been validated for people with PCOS. More detailed tests also exist that look at

cardiac function however these are more expensive and not routinely used and may increase stress and anxiety in patients due to their more invasive nature. The authors also highlight that although there are CVD risk factors in PCOS patients during the reproductive years, this does not mean the risk will continue to be higher post menopause. They suggest that women with PCOS could have similar CVD risks as the general population during this time. It is also highlighted that the link between PCOS and CVD risk factors is likely to be controversial however it is evidence based. Twelve different outcomes were included, with varying CVD related comparisons. The majority of the 25 included studies were cohort studies, with some cross-sectional studies, a mixture of Rotterdam, NIH and AE-PCOS criteria were used for classification of PCOS patients. Outcomes found CVD risks to be higher in PCOS patients in 9 analyses, and no difference in 4 analyses. The statistically significant results for the composite CVD outcomes and CV mortality did favour the PCOS groups with the odds ratios ranging from 1.01 to 1.68. Results were more profound where MI and stroke were reported individually. There was inconsistency for several outcomes. All were either low or very low certainty.

There are 5 CRs which detail potential approaches to managing CVD risk and monitoring of women with PCOS, and a practice point to highlight the potential impact of varying ethnicities. One CR highlights the need for an annual blood pressure check and when planning pregnancy due to the high risk of hypertensive disorders. Two of the 25 included studies looked at hypertension in women with PCOS. One study concluded that increased weight was the likely cause of the increased risk of hypertension rather than PCOS itself, whilst the other study found an increased risk of hypertension (1.70-fold, 95% CI 1.43 to 2.07) in women with PCOS, but only in those of reproductive age. The IG noted that as the evidence was mixed, a CR was made but did note that this might have implementation considerations. The review also highlights the importance of lifestyle and weight management interventions as preventative approaches. Future research priorities

include comprehensive assessment of CVD events and the validation of CVD risk prediction models in women with PCOS, also considering ethnic variation.

IG economic evidence

No health economic evidence was identified in the IG for review question 1.8 on cardiovascular disease risk associated with PCOS.

The IG noted that there could be a possible increase in costs for screening associated with their recommendations but noted that these costs would likely offset by a reduced incidence of cardiovascular events.

The IG also noted that since CVD risk factors, especially obesity, are more prevalent in people in the lower strata of socio-economic class; screening for cardiovascular disease risk will probably has positive impact on health equity. They acknowledged that increased screening would increase equity, however differential patterns in screening may result in inequity.

1.8.2 NICE economic evidence

Included studies

A single health economic search was performed by NICE to identify published economic evaluations of relevance to all review questions in this guideline. See the literature search strategy in Appendix A.

No economic studies were identified which were applicable to this review question (see economic study selection flow chart in Appendix B).

Excluded studies

No economic studies were reviewed at full text and excluded from this review.

Economic model

No original economic modelling was conducted for review question 1.8 on the cardiovascular disease risk associated with PCOS.

1.8.3 NICE recommendations

The relevant recommendations for this section are Rec 1.2.3 and 1.16.1 to 1.16.4.

1.8.4 The committee's discussion and interpretation of the evidence

Clinical

The IG made two EBRs (1.8.1 and 1.8.2), five CRs (1.8.3 to 1.8.7), and one practice point recommendation (1.8.8) on cardiovascular disease risk.

The committee discussed the recommendations regarding cardiovascular disease risk in women with PCOS. The committee decided to contextualise IG recommendation 1.8.1 to highlight the need for cardiovascular risk assessment when a diagnosis of PCOS was made. One CR (1.8.3) was contextualised by the committee to highlight the importance of testing for cardiovascular disease in women upon diagnosis with PCOS. The committee felt a link to NICE's guideline NG238 on cardiovascular disease would be helpful. They wished to specifically highlight the sections on risk identification and assessment and primary and secondary prevention of CVD for quicker reference when health practitioners are seeing patients. The cross referral to NICE guideline NG238 was incorporated into two new recommendations. A new recommendation was made for those who did not meet the criteria as high risk to receive advice to prevent CVD in the future.

The committee discussed the need for an annual review for patients with PCOS, as IG recommendation 1.8.4 describes the need for annual blood pressure checks due to the risk of hypertension. Committee members felt this was an important opportunity for people with PCOS to discuss their health and a dedicated yearly review would be helpful for this. Lay members highlighted that although anyone on

regular medication should have an annual review, this is often with a pharmacist rather than a doctor and as such would not be able to discuss medical issues. Lay members also described being diagnosed with PCOS and then having limited further contact, other than if wanting to discuss pregnancy options, as such felt having an annual review would give opportunity to discuss their health further. Committee members discussed that having an annual opportunity to discuss any new signs and symptoms, including hypertensive disorders, could lead to earlier diagnosis of comorbidities associated with PCOS, which if detected earlier could lead to prevention measures rather than treatment. An example given was the potential to be identified as having pre-diabetes which allows time for lifestyle changes and interventions, rather than a type 2 diabetes diagnosis which would lead to increased treatment and healthcare visits. The committee also highlighted that the annual review would be in line with the Governments 10-year plan to move from sickness to prevention. The committee also highlighted the impact this could have on GP practices and asked for more detail to be provided in the recommendation as to what the annual review should include and noted that PCOS is a complex condition and as such the annual review would likely need a longer clinical consultation. The committee also suggested links to other long-term conditions that should be considered during the annual review would be helpful. Lay committee members welcomed this change as some felt it has been difficult to get further health checks from their GP at times.

Health economic

No health economic evidence was identified in the IG on cardiovascular risk factors. In addition, no health economic evidence was identified in the health economic literature search conducted by NICE.

The committee discussed the EBR 1.8.2 from the IG which stated that all women with PCOS should be assessed for cardiovascular disease risk factors. The committee acknowledged that they could not be certain of the intended meaning of this

recommendation, however, concluded that offering all women with PCOS clinical tests to assess their cardiovascular risk factors would be significant change in clinical practice. The committee discussed that offering all women who are newly diagnosed with PCOS tests to assess their cardiovascular risk would be highly unlikely to be a cost-effective use of NHS resources as the overall risk of cardiovascular disease for people with PCOS under 40 years of age is low. The committee acknowledged that the recommendations they contextualised, alongside cross-referring to NICE's existing guideline on cardiovascular disease, represented UK current practice whilst also ensuring people over the age of 40 are appropriately screened for cardiovascular disease irrespective of a PCOS diagnosis.

The committee did however discuss annual review in a broader sense. It was acknowledged that in the IG no clinical evidence review was conducted to assess the clinical and cost-effectiveness of providing annual reviews for people with PCOS. However, the committee concluded that it was important to add a new recommendation to the guideline – recommending an annual review for people aged 10 and over with PCOS that includes a review of their signs and symptoms, medicines use and long-term risks of developing health related conditions. It was noted that this recommendation is likely reflective of best clinical practice but would represent a change in practice for most. The committee were made aware of the additional costs to the NHS associated with implementing this recommendation, but the committee were confident that making such a recommendation would be cost effective use of NHS resources. The committee discussed that annual review would result in a number of benefits, both in terms of long-term cost savings and QALY gains.

Firstly, it was noted that an annual review would allow for clinicians to detect endometrial hyperplasia earlier. It was noted that people with PCOS may not always present to their doctor if they are experiencing 4 periods or less a year when this represents a change in their cycle regularity. Although the overall risk of developing endometrial cancer from endometrial hyperplasia is low (around 5%), being able to

effectively treat endometrial hyperplasia significantly reduces the risk of developing cancer.

The committee also acknowledged that annual review may allow for prevention of diabetes, hypertension and CVD disease – predominantly by increasing awareness and providing information on lifestyle interventions to prevent disease. The committee noted that this annual review would allow for testing for these conditions (for example HbA1c) when the healthcare practitioner thought this was required but emphasised that routine testing should not be conducted at every annual review.

Earlier planning around fertility was also discussed as it was noted this is a high area of concern for people with PCOS. The lay members on the committee noted that it is not uncommon practice for people to be told that conception will be difficult due to their PCOS. It was, however, noted by the wider committee that conception is much easier compared to those people with unexplained infertility. The committee discussed that earlier planning should help ease concerns for people with PCOS hoping to conceive at some point in the future, but also earlier planning may increase chances of a successful pregnancy and help people access the appropriate treatment when and if required.

It was also noted than an increased understanding of treatment options will likely support compliance whilst simultaneously improving people's mental health. Weight management and early intervention to prevent weight related health issues was also discussed (such as cancer and arthritis). Assessing for this risk of sleep apnoea was also discussed.

The committee acknowledged that an annual review should be offered to people to help address any concerns they may have and to provide a point of contact. The committee noted that these concerns could include, period regularity, fertility, weight, mood, sleep apnoea, cardiovascular or diabetes risk assessment/management. The committee did, however, note that if for example HbA1 and blood pressure is normal,

or a person's sleep apnoea risk is low, there is no need to see someone annually if they do not wish to be seen this frequently.

The committee acknowledged that their recommendation to offer annual review is validated by the NHS England Health Strategy moving to a preventative approach.

The committee acknowledged that this recommendation could result in a significant resource impact for the NHS, but they noted it was challenging to estimate the true impact as a number of people with PCOS will already have co-existing conditions and therefore may already be having an annual review for these. It is estimated that if 10% of the prevalent population were to have an annual review this would cost over £4million a year, based on a prevalence of approximately 840,000 people with PCOS in England (Bernie et al 2024) and the cost of a 10-minute GP appointment of £48 (PSSRU 2024/2025). It was discussed that the resource impact may be most prevalent for younger people where the probability of co-existing conditions is lower. However, anecdotally the committee discussed that this cohort are more likely to be presenting at the GP with their symptoms for help and support if required due to the increased awareness of PCOS.

1.9 Impaired glucose tolerance and type 2 diabetes risk

Review question 1.9.1: Are women with PCOS at increased risk for impaired glucose tolerance and type 2 diabetes?

Review question 1.9.2: In women with PCOS, what is the most effective tool/method to assess risk of type 2 diabetes mellitus?

1.9.1 Recommendations from the International evidence-based guideline for PCOS*

Impaired glucose tolerance and type 2 diabetes risk

Evidence-based recommendations:

1.9.1 Healthcare professionals and women with PCOS should be aware that, regardless of age and BMI, women with PCOS have an increased risk of impaired fasting glucose, impaired glucose tolerance and type 2 diabetes.

1.9.2 Glycaemic status should be assessed at diagnosis in all adults and adolescents with PCOS.

Consensus recommendations:

1.9.3 Glycaemic status should be reassessed every one to three years, based on additional individual risk factors for diabetes.

1.9.4 Healthcare professionals, women with PCOS and other stakeholders should prioritise preventative strategies to reduce type 2 diabetes risk.

1.9.5 Funding bodies should recognise that PCOS is highly prevalent, has significantly higher risk for diabetes and should be funded accordingly.

1.9.6 Diabetes general population guidelines should consider the inclusion of PCOS as an independent risk factor for diabetes.

Practice points:

1.9.7 Healthcare professionals, adults and adolescents with PCOS and their first-degree relatives, should be aware of the increased risk of diabetes and the need for regular glycaemic assessment.

1.9.8 Women with type 1 and type 2 diabetes have an increased risk of PCOS and screening should be considered in individuals with diabetes.

Glycaemic testing

Evidence based recommendations:

1.9.9 Healthcare professionals and women with PCOS should recommend the 75g oral glucose tolerance test (OGTT) as the most accurate test to assess glycaemic status in PCOS, regardless of BMI.

1.9.10 If an OGTT cannot be performed, fasting plasma glucose and/or glycated haemoglobin (HbA1c) could be considered, noting significantly reduced accuracy.

1.9.11 An OGTT should be considered in all women with PCOS and without pre-existing diabetes, when planning pregnancy or seeking fertility treatment, given the high risk of hyperglycaemia and the associated comorbidities in pregnancy. If not performed preconception, an OGTT could be offered at the first prenatal visit and all women with PCOS should be offered the test at 24-28 weeks gestation.

Practice point:

1.9.12 Insulin resistance is a pathophysiological factor in PCOS, however, clinically available insulin assays are of limited clinical relevance and are not recommended in routine care (refer to 3.1.10).

1 **IG clinical evidence**

2 **Critical appraisal - ROBIS systematic review checklist (1.9.1)**

Section	Question	Answer
Study eligibility criteria	Did the review adhere to pre-defined objectives and eligibility criteria?	Yes *(Criteria well described, studies appear to meet this)*
Study eligibility criteria	Were the eligibility criteria appropriate for the review question?	Yes *(detailed PICO available which is suitable for the review)*
Study eligibility criteria	Were eligibility criteria unambiguous?	Yes *(eligibility criteria clearly described in PICO)*
Study eligibility criteria	Were all restrictions in eligibility criteria based on study characteristics appropriate (e.g. date, sample size, study quality, outcomes measured)?	Yes *(restrictions on date are present but not well described)*
Study eligibility criteria	Were any restrictions in eligibility criteria based on sources of information appropriate (e.g. publication status or format, language, availability of data)?	Yes *(study types selected are appropriate for review question, limited to English language studies is appropriate)*
Study eligibility criteria	Concerns regarding specification of study eligibility criteria	Low *(all criterion rated yes as PICO and limits appropriate and suitably detailed).*
Identification and selection of studies	Did the search include an appropriate range of databases/electronic sources for published and unpublished reports?	Yes *(Appropriate number (5) and range of databases used)*
Identification and selection of studies	Were methods additional to database searching used to identify relevant reports?	Probably yes *(Does not detail any additional searches/manual search strategies however the general methods says that they sought relevant papers from their experts)*
Identification and selection of studies	Were the terms and structure of the search strategy likely to retrieve as many eligible studies as possible?	No *(Subject terms not translated between databases. All lines of*

Section	Question	Answer
		CINAHL search not reported. Structure of search is OK)
Identification and selection of studies	Were restrictions based on date, publication format, or language appropriate?	Probably yes *(Restrictions in line with the protocol)*
Identification and selection of studies	Were efforts made to minimise error in selection of studies?	Probably yes *(Only one reviewer analysed studies for inclusion via title and abstract and used full text where necessary to make decisions. Ideally a second reviewer would have been involved.)*
Identification and selection of studies	Concerns regarding methods used to identify and/or select studies	High *(Subject terms not translated between databases)*
Data collection and study appraisal	Were efforts made to minimise error in data collection?	Probably yes *(Most likely has been done but difficult to ascertain due to only one reviewer)*
Data collection and study appraisal	Were sufficient study characteristics available for both review authors and readers to be able to interpret the results?	Yes *(Study table with study characteristics present allowing interpretation of results)*
Data collection and study appraisal	Were all relevant study results collected for use in the synthesis?	Yes *(18 studies included in review, 12 in MA and GRADE profiles, appears reasonable for review)*
Data collection and study appraisal	Was risk of bias (or methodological quality) formally assessed using appropriate criteria?	Yes *(Risk of bias assessment has been completed for each study and is presented in tables)*
Data collection and study appraisal	Were efforts made to minimise error in risk of bias assessment?	Probably yes *(This could have been minimised further by having an additional reviewer, but no other concerns for this area noted)*
Data collection and study appraisal	Concerns regarding methods used to collect data and appraise studies	Low *(Risk of bias conducted, data collection and appraisal could have been improved with 2 reviewers)*

Section	Question	Answer
Synthesis and findings	Did the synthesis include all studies that it should?	Yes *(Results were included from all 12 studies from the meta-analysis. The 6 remaining studies were included in the review but not considered able to be used for the MA)*
Synthesis and findings	Were all pre-defined analyses reported or departures explained?	No information *(No information was given regarding the planned analysis of the results)*
Synthesis and findings	Was the synthesis appropriate given the nature and similarity in the research questions, study designs and outcomes across included studies?	Yes *(Meta-analysis was used appropriately for this review question, study designs were appropriate for analysis)*
Synthesis and findings	Was between-study variation (heterogeneity) minimal or addressed in the synthesis?	Probably yes *(No serious inconsistencies noted in results tables)*
Synthesis and findings	Were the findings robust, e.g. as demonstrated through funnel plot or sensitivity analyses?	Yes *(Studies are well distributed on funnel plots indicating studies have been appropriately included)*
Synthesis and findings	Were biases in primary studies minimal or addressed in the synthesis?	Probably yes *(No obvious discussion about this but appears reasonable)*
Synthesis and findings	Concerns regarding the synthesis and findings	Low *(Meta-analysis and funnel plots seem appropriate)*
Judging risk of bias	Concerns regarding specification of study eligibility	Low *(All criterion rated yes as PICO and limits appropriate and suitably detailed).*
Judging risk of bias	Concerns regarding methods used to identify and/or select studies	High *(Subject terms not translated between databases)*
Judging risk of bias	Concerns regarding methods used to collect data and appraise studies	Low *(Risk of bias conducted, data collection and appraisal could have been improved with 2 reviewers)*

Section	Question	Answer
Judging risk of bias	Concerns regarding the synthesis and findings	Low *(Meta-analysis and funnel plots seem appropriate)*
Overall review ratings	Overall risk of bias	Low *(Eligibility criteria, risk of bias appraisal and meta-analysis appropriate but issues with search)*
Overall review ratings	Applicability as a source of data	Fully applicable

1

2 **Critical appraisal - ROBIS systematic review checklist (1.9.2)**

Section	Question	Answer
Study eligibility criteria	Did the review adhere to pre-defined objectives and eligibility criteria?	Yes *(Appears to meet well described criteria)*
Study eligibility criteria	Were the eligibility criteria appropriate for the review question?	Yes *(PICO describes it well and is appropriate for review question)*
Study eligibility criteria	Were eligibility criteria unambiguous?	Yes *(PICO describes it well and is appropriate for review question)*
Study eligibility criteria	Were all restrictions in eligibility criteria based on study characteristics appropriate (e.g. date, sample size, study quality, outcomes measured)?	Yes *(Restrictions seem reasonable just not well reported)*
Study eligibility criteria	Were any restrictions in eligibility criteria based on sources of information appropriate (e.g. publication status or format, language, availability of data)?	Probably yes *(Study type eligibility appropriate for review question)*
Study eligibility criteria	Concerns regarding specification of study eligibility criteria	Low *(PICO was sufficiently detailed)*

Section	Question	Answer
Identification and selection of studies	Did the search include an appropriate range of databases/electronic sources for published and unpublished reports?	Yes *(4 databases named, all appropriate for review question except PsycINFO superfluous)*
Identification and selection of studies	Were methods additional to database searching used to identify relevant reports?	Yes *(4 studies found by manual searching and the general methods says that they sought relevant papers from their experts)*
Identification and selection of studies	Were the terms and structure of the search strategy likely to retrieve as many eligible studies as possible?	No *(No subject headings used in tools block for Medline Embase PsycINFO or EBM. Subject terms not translated between databases. DTA filter used. This is not advised.)*
Identification and selection of studies	Were restrictions based on date, publication format, or language appropriate?	Probably no *(Nothing in protocol about date limits but it is an update.* *Unclear how date limits seem to have been applied. its possible references were screened as 2 sets with or without the DTA filter, but this is not explained.)*
Identification and selection of studies	Were efforts made to minimise error in selection of studies?	Yes *(2 independent reviewers plus evidence team)*
Identification and selection of studies	Concerns regarding methods used to identify and/or select studies	High *(Subject terms missing and not translated. Use of DTA filter. Unclear use of limits)*
Data collection and study appraisal	Were efforts made to minimise error in data collection?	Yes *(2 reviewers plus evidence team)*

Section	Question	Answer
Data collection and study appraisal	Were sufficient study characteristics available for both review authors and readers to be able to interpret the results?	Yes *(Study characteristics table available with sufficient information to interpret results)*
Data collection and study appraisal	Were all relevant study results collected for use in the synthesis?	Yes *(Sensitivity and specificity were extracted from all studies and were appropriate for the analysis)*
Data collection and study appraisal	Was risk of bias (or methodological quality) formally assessed using appropriate criteria?	Yes *(Risk of bias tool/checklist seems appropriate)*
Data collection and study appraisal	Were efforts made to minimise error in risk of bias assessment?	Yes *(2 assessors reviewed studies alongside evidence team)*
Data collection and study appraisal	Concerns regarding methods used to collect data and appraise studies	Low *(All criterions were rated yes due to sufficient details of studies and for assessing risk of bias)*
Synthesis and findings	Did the synthesis include all studies that it should?	Yes *(All 6 studies have results accounted for)*
Synthesis and findings	Were all pre-defined analyses reported or departures explained?	No information *(No information given on how they planned to conduct study)*
Synthesis and findings	Was the synthesis appropriate given the nature and similarity in the research questions, study designs and outcomes across included studies?	Yes *(No meta-analysis as looking at the study characteristics, it seems that there were differences in index tests, outcomes and study designs which would preclude a meta-analysis)*
Synthesis and findings	Was between-study variation (heterogeneity) minimal or addressed in the synthesis?	Probably yes *(GRADE summaries must have been based on narrative summaries as no meta-analysis or forest plot,*

Section	Question	Answer
		little discussion to describe how this was undertaken)
Synthesis and findings	Were the findings robust, e.g. as demonstrated through funnel plot or sensitivity analyses?	Probably no *(No funnel plot due to low included study numbers, a discussion about these findings would have been helpful)*
Synthesis and findings	Were biases in primary studies minimal or addressed in the synthesis?	Probably yes *(Difficult to determine how conclusions were reached here due to lack of discussion)*
Synthesis and findings	Concerns regarding the synthesis and findings	Low *(No meta-analysis conducted but this was likely due to differences in index tests, outcomes and study designs so appropriate)*
Judging risk of bias	Concerns regarding specification of study eligibility	Low *(PICO was sufficiently detailed)*
Judging risk of bias	Concerns regarding methods used to identify and/or select studies	High *(Subject terms missing and not translated. Use of DTA filter. Unclear use of limits)*
Judging risk of bias	Concerns regarding methods used to collect data and appraise studies	Low *(All criterions were rated yes due to sufficient details of studies and for assessing risk of bias)*
Judging risk of bias	Concerns regarding the synthesis and findings	Low *(No meta-analysis conducted but this was likely due to differences in index tests, outcomes and study designs so appropriate)*
Overall review ratings	Overall risk of bias	Low

Section	Question	Answer
		(Study eligibility appropriately detailed, sufficient details for studies and risk of bias assessment and appropriate no meta-analysis)
Overall review ratings	Applicability as a source of data	Fully applicable

Evidence to recommendation justification: The review highlights that in general the relative risk of type 2 diabetes (T2D) in PCOS to be approximately 3-fold. All EBRs and CRs in this section were strong recommendations for the option. A significant increased risk of T2D was seen when women with PCOS were compared to non-PCOS women in this review, moderate quality evidence. The pooled analysis found a higher risk of T2D, impaired fasting glucose (IFG) and impaired glucose tolerance (IGT) and pre-diabetes in women with PCOS when compared to control groups. These results were maintained when age matching was considered. However, when looking at adolescent women with PCOS, the risks for T2D and IFG were like the general population, but IGT remained higher. Post-menopausal women also had a higher risk of T2D than controls, but no studies were available for IFG, IGT or pre-diabetes for this group. Ethnic diversity in post-menopausal women was not included in this review.

Six studies were also identified comparing tests for diagnosing T2D in women with PCOS, all using the oral glucose tolerance test (OGTT) as a reference standard. The use of fasting plasma glucose (FPG) and HbA1c were investigated but results found that these should not replace the OGTT. The review suggests that although the quality of evidence is low in this area, the results are consistent enough to support the EBR for the use of the OGTT, however it should be noted that this evidence is from adults and might be different for the adolescent population. Research priorities include the need for more diagnostic tests to be compared to OGTT and a need to

simplify the process for settings where regular OGTT may not be feasible. Ethnicity and menopause status are also areas highlighted for greater understanding.

IG economic evidence

No health economic evidence was identified in the IG for review question 1.9 on impaired glucose tolerance and type 2 diabetes risk.

The IG noted that costs of implementing their recommendations will vary according to healthcare settings. The IG also noted that some women and healthcare professionals may find OGTT inconvenient and costly compared to other tests.

1.9.2 NICE economic evidence

Included studies

A single health economic search was performed by NICE to identify published economic evaluations of relevance to all review questions in this guideline. See the literature search strategy in Appendix A.

No economic studies were identified which were applicable to this review question (see economic study selection flow chart in Appendix B).

Excluded studies

No economic studies were reviewed at full text and excluded from this review.

Economic model

No original health economic model was conducted for review question 1.9.

Unit costs

Table 3: Unit costs for OGTT and HbA1c tests

Resource	Unit costs	Source
Oral glucose tolerance test (OGTT)	£399	National Cost Collection 2024/25, Currency code: KB02K – Diabetes with

Resource	Unit costs	Source
		Hyperglycaemic Disorders, with CC Score 0-1, Daycase
HbA1c	£8.00	National Cost Collection 2024/25, PATH04, clinical biochemistry (service code: 999) + PATH08, phlebotomy (service code: 999)

1.9.3 NICE recommendations

The relevant recommendations for this section are Rec 1.17.1, 1.17.2 and 1.17.37.

1.9.4 The committee's discussion and interpretation of the evidence

Clinical

The committee discussed that people with PCOS have a higher risk of developing impaired glucose tolerance. As such IG recommendations 1.9.1 and 1.9.2 were adapted to provide information and the need for assessing glycaemic status following a PCOS diagnosis. The committee discussed the use of the oral glucose tolerance test (OGTT) and its applicability to the PCOS population. Concerns were raised about the most suitable healthcare setting for this test to be performed, as it is rarely performed outside of maternity care. The committee also highlighted that the OGTT should not be used by people who have undergone bariatric surgery or who are currently taking metformin. Concerns were also raised as to which test methods should be used in preference to the OGTT, as many women with PCOS experience insulin resistance which may not be detected using HbA1c. The use of the OGTT in pregnancy is covered by NICE guideline NG3 – diabetes in pregnancy. The committee concluded that both fasting glucose and HbA1c should offer suitable assessment of glycaemic status and felt that the OGTT was too time and labour intensive for little gain over these tests. The committee contextualised EBRs 1.9.10 to highlight that either HbA1c or fasting plasma glucose can be used to assess glycaemic status.

Health economic

No health economic evidence was identified in the IG on impaired glucose tolerance and type 2 diabetes. In addition, no health economic evidence was identified in the health economic literature search conducted by NICE.

For this section of the guideline the committee made recommendations that were reflective of UK current practice. In terms of how to assess for glycaemic status this differed from the IG's recommendations where they recommended that an OGTT be performed in preference to fasting plasma glucose or HbA1c. This deviation from the IG was due to concerns of the cost-effectiveness for OGTT. The committee were presented unit costs, alongside the clinical evidence from the IG and concluded that the marginally improved outcomes for OGTT could not be justified due the large differential in costs (£399 for OGTT and £8 for HbA1c). The committee also noted that an OGTT can be an unpleasant test for those receiving it and it is also more time and resource intensive for both patients and health care professionals.

Overall, as the committee made recommendations reflective of current practice no significant resource impact is anticipated for this review question.

1.10 Obstructive sleep apnoea

Review question 1.10: Are women with PCOS at increased risk for sleep apnea?

1.10.1 Recommendations from the International evidence-based guideline for PCOS*

Evidence-based recommendations:

1.10.1 Healthcare professionals should be aware that women with PCOS have significantly higher prevalence of obstructive sleep apnea (OSA) compared to women without PCOS, independent of BMI.

1.10.2 Women with PCOS should be assessed for symptoms (i.e. snoring in combination with waking unrefreshed from sleep, daytime sleepiness or fatigue) and if present, screen with validated tools or refer for assessment.

Practice points:

1.10.3 Simple obstructive sleep apnea screening questionnaires (such as the Berlin questionnaire, validated in the general population) can assist in identifying obstructive sleep apnea in women with PCOS, noting that diagnosis requires a formal sleep study.

1.10.4 Goals of treatment should target obstructive sleep apnea related symptom burden.

IG clinical evidence

Critical appraisal - ROBIS systematic review checklist

Section	Question	Answer
Study eligibility criteria	Did the review adhere to pre-defined objectives and eligibility criteria?	Yes *(Clearly defined PICO)*
Study eligibility criteria	Were the eligibility criteria appropriate for the review question?	Yes *(Detailed PICO available which is suitable for the review)*
Study eligibility criteria	Were eligibility criteria unambiguous?	Yes *(Yes clearly defined)*
Study eligibility criteria	Were all restrictions in eligibility criteria based on study characteristics appropriate (e.g. date, sample size, study quality, outcomes measured)?	Yes *(Restrictions were suitable based on the PICO requirements. No date limit)*
Study eligibility criteria	Were any restrictions in eligibility criteria based on sources of information appropriate (e.g. publication status or format, language, availability of data)?	Probably yes *(Language restrictions were appropriate)*
Study eligibility criteria	Concerns regarding specification of study eligibility criteria	Low *(All parameters are appropriate to review)*
Identification and selection of studies	Did the search include an appropriate range of databases/electronic sources for published and unpublished reports?	Yes *(5 named databases used, all appropriate to search except PsycINFO superfluous)*
Identification and selection of studies	Were methods additional to database searching used to identify relevant reports?	Probably yes *(Does not detail any additional searches/manual search strategies however the general methods says that they sought relevant papers from their experts)*
Identification and selection of studies	Were the terms and structure of the search strategy likely to retrieve as many eligible studies as possible?	Yes *(Subject headings not translated between databases. Free text terms would have benefitted from more adjacency searching rather than reliance on phrase and combinations.)*
Identification and selection of studies	Were restrictions based on date, publication format, or language appropriate?	Yes *(No restrictions applied as per protocol)*

Section	Question	Answer
Identification and selection of studies	Were efforts made to minimise error in selection of studies?	Yes *(2 reviewers on title and abstract)*
Identification and selection of studies	Concerns regarding methods used to identify and/or select studies	High *(Subject headings not translated).*
Data collection and study appraisal	Were efforts made to minimise error in data collection?	Yes *(2 reviewers for study selection, consultation with study team available if required)*
Data collection and study appraisal	Were sufficient study characteristics available for both review authors and readers to be able to interpret the results?	Yes *(Study characteristics table is sufficiently detailed for this)*
Data collection and study appraisal	Were all relevant study results collected for use in the synthesis?	Probably yes *(All studies appear to have been included)*
Data collection and study appraisal	Was risk of bias (or methodological quality) formally assessed using appropriate criteria?	Yes *(A risk of bias tool was completed for 10 included studies)*
Data collection and study appraisal	Were efforts made to minimise error in risk of bias assessment?	Probably yes *(Lack of information on how confounding factors may have been approached but otherwise ok)*
Data collection and study appraisal	Concerns regarding methods used to collect data and appraise studies	Low *(Methods seem appropriate for review)*
Synthesis and findings	Did the synthesis include all studies that it should?	Yes *(Yes according to PRISMA diagram)*
Synthesis and findings	Were all pre-defined analyses reported or departures explained?	No information *(Not stated)*
Synthesis and findings	Was the synthesis appropriate given the nature and similarity in the research questions, study designs and outcomes across included studies?	Yes *(MA completed for 8 studies)*
Synthesis and findings	Was between-study variation (heterogeneity) minimal or addressed in the synthesis?	Probably yes *(Appears to have been accounted for, and was less than 50% for individual events)*

Section	Question	Answer
Synthesis and findings	Were the findings robust, e.g. as demonstrated through funnel plot or sensitivity analyses?	Yes *(Funnel plots available)*
Synthesis and findings	Were biases in primary studies minimal or addressed in the synthesis?	Probably yes *(Information limited on potential confounders)*
Synthesis and findings	Concerns regarding the synthesis and findings	Unclear *(Potentially lacking in information around confounders and no information for pre-defined analyses)*
Judging risk of bias	Concerns regarding specification of study eligibility	Low *(All parameters are appropriate to review)*
Judging risk of bias	Concerns regarding methods used to identify and/or select studies	High *(Subject headings not translated).*
Judging risk of bias	Concerns regarding methods used to collect data and appraise studies	Low *(Methods seem appropriate for review)*
Judging risk of bias	Concerns regarding the synthesis and findings	Unclear *(Potentially lacking in information around confounders and no information for pre-defined analyses)*
Overall review ratings	Overall risk of bias	Low *(low overall, one section potentially unclear due to some missing information)*
Overall review ratings	Applicability as a source of data	Partially applicable

Evidence to recommendations justification: There is currently no validated sleep apnoea screening tool for women with PCOS. The results from 4 studies found that symptomatic OSA was 17.95 times more likely to be present in women with PCOS compared to controls. This was based on results from 8 peer reviewed publications that had considered OSA to be present if >5 events per hour were seen. The Berlin questionnaire is used in sleep clinic populations; however, it is stated that it performs less well in general populations, presumably including women with PCOS. Three of the studies compared women with PCOS to controls over a healthy weight and found

OSA was 4.25 times more likely to be present in women with PCOS, however no studies compared adolescents and as such the evidence for this age group might be less generalisable. The review highlights that primary care staff are likely to be familiar with screening for sleep apnoea, and as such the implementation of these recommendations is not anticipated to be an issue, however educating women with PCOS to be aware of OSA is suggested. Research priorities include validating existing OSA risk tools for PCOS, understanding the underlying mechanisms of OSA in women with PCOS, studies in adolescents and long-term research for OSA and PCOS.

IG economic evidence

No health economic evidence was identified in the IG for review question 1.10 on obstructive sleep apnoea.

The IG noted that access to screening and treatment may be limited in some settings. However, concluded that simple clinical screening should be possible, but further assessment and treatment may be variable.

1.10.2 NICE economic evidence

Included studies

A single health economic search was performed by NICE to identify published economic evaluations of relevance to all review questions in this guideline. See the literature search strategy in Appendix A.

No economic studies were identified which were applicable to this review question (see economic study selection flow chart in Appendix B).

Excluded studies

No economic studies were reviewed at full text and excluded from this review.

Economic model

No original health economic modelling was conducted for review question 1.10.

1.10.3 NICE recommendations

The relevant recommendations for this section are Rec 1.18.1 and 1.18.2.

1.10.4 The committee's discussion and interpretation of the evidence

Clinical

The IG made two EBRs and two practice point recommendations on obstructive sleep apnoea. The committee decided to contextualise the two EBRs in this section of the IG as they felt the risk of obstructive sleep apnoea in woman with PCOS should be highlighted. A link to NICE guideline NG202 Obstructive sleep apnoea/hypopnoea syndrome was included in this section for further information.

Health economic

No health economic evidence was identified in the IG or in NICE's health economic search for obstructive sleep apnoea.

The contextualised recommendations made by the committee explained the higher prevalence of obstructive sleep apnoea for people with PCOS, noting that people with PCOS should be assessed for symptoms of sleep apnoea. If symptoms are present, referral for assessment should take place. The committee discussed that additional information on the signs of symptoms of sleep apnoea, in addition to the diagnostic tools and pathway, can be found in NICE's existing guidance on sleep apnoea (NG202). The committee therefore cross-referred to this guidance in their recommendation where they refer to assessing sleep apnoea.

The committee noted that the recommendations they contextualised are reflective of best clinical practice. However, the committee did acknowledge that these recommendations could increase awareness of sleep apnoea for those with PCOS

and therefore could result in a small increase in assessment and referrals. No substantial resource impact is anticipated as a result of these recommendations, as they are reflective of best practice, but there may be a small increase in costs to the NHS resulting from those where these recommendations do not reflect current practice due to the potential increased awareness of the risks for obstructive sleep apnoea for people with PCOS.

1.11 Endometrial hyperplasia and cancer

Review question 1.11: Are women with PCOS at increased risk of endometrial cancer?

1.11.1 Recommendations from the International evidence-based guideline for PCOS*

Evidence-based recommendation:

1.11.1 Healthcare professionals should be aware that premenopausal women with PCOS have markedly higher risk of developing endometrial hyperplasia and endometrial cancer.

Practice points:

1.11.2 Women with PCOS should be informed about the increased risk of endometrial hyperplasia and endometrial cancer, acknowledging that the overall chance of developing endometrial cancer is low, therefore routine screening is not recommended.

1.11.3 Long-standing untreated amenorrhea, higher weight, type 2 diabetes and persistent thickened endometrium are additional to PCOS as risk factors for endometrial hyperplasia and endometrial cancer.

1.11.4 Women with PCOS should be informed of preventative strategies including weight management, cycle regulation and regular progestogen therapy.

1.11.5 When excessive endometrial thickness is detected, consideration of a biopsy with histological analysis and withdrawal bleed is indicated.

1 **IG clinical evidence**

2 **Critical appraisal - ROBIS systematic review checklist**

Section	Question	Answer
Study eligibility criteria	Did the review adhere to pre-defined objectives and eligibility criteria?	Yes *(PICO was clear and detailed, search dates clearly listed)*
Study eligibility criteria	Were the eligibility criteria appropriate for the review question?	Yes *(detailed PICO available which is suitable for the review)*
Study eligibility criteria	Were eligibility criteria unambiguous?	Yes *(yes clearly defined)*
Study eligibility criteria	Were all restrictions in eligibility criteria based on study characteristics appropriate (e.g. date, sample size, study quality, outcomes measured)?	Yes *(appropriate and based on PICO, no date limit because new review)*
Study eligibility criteria	Were any restrictions in eligibility criteria based on sources of information appropriate (e.g. publication status or format, language, availability of data)?	Yes *(English language was only noted restriction which is appropriate to review)*
Study eligibility criteria	Concerns regarding specification of study eligibility criteria	Low *(all parameters appear appropriate to review)*
Identification and selection of studies	Did the search include an appropriate range of databases/electronic sources for published and unpublished reports?	Yes *(5 named databases, all appropriate for review question except PsycINFO is superfluous)*
Identification and selection of studies	Were methods additional to database searching used to identify relevant reports?	Probably yes *(not mentioned directly though states 2 original papers were identified from SR's in PRISMA and the general methods says that they sought relevant papers from their experts)*
Identification and selection of studies	Were the terms and structure of the search strategy likely to retrieve as many eligible studies as possible?	No *(subject headings not translated between databases. Pre-selected auto*

Section	Question	Answer
		indexed human limits used rather than tested methods. Structure is sound.)
Identification and selection of studies	Were restrictions based on date, publication format, or language appropriate?	Probably no *(date limits since 1990 applied. The protocol says it is a new review and has no date limits.)*
Identification and selection of studies	Were efforts made to minimise error in selection of studies?	Yes *(one reviewer selecting at title and abstract stage, 2 reviewers would have been better)*
Identification and selection of studies	Concerns regarding methods used to identify and/or select studies	High *(subject headings not translated. Use of auto indexed human limits.)*
Data collection and study appraisal	Were efforts made to minimise error in data collection?	Probably yes *(appraisal by one reviewer, ideally a second would have helped)*
Data collection and study appraisal	Were sufficient study characteristics available for both review authors and readers to be able to interpret the results?	Yes *(study table with study characteristics and structured extraction form used allowing interpretation of results)*
Data collection and study appraisal	Were all relevant study results collected for use in the synthesis?	Yes *(19 studies included in review, 16 included in GRADE tables)*
Data collection and study appraisal	Was risk of bias (or methodological quality) formally assessed using appropriate criteria?	Yes *(quality appraisal for each included study was completed which included a risk of bias check list in a structured form)*
Data collection and study appraisal	Were efforts made to minimise error in risk of bias assessment?	Probably yes *(this could have been minimised further by having an additional reviewer, but no other concerns for this area noted.)*
Data collection and study appraisal	Concerns regarding methods used to collect data and appraise studies	Low *(methods seem appropriate for review)*
Synthesis and findings	Did the synthesis include all studies that it should?	Yes *(17 original studies, 4 SRs. 2 additional papers identified from the SR's 16 studies were included in GRADE tables)*

Section	Question	Answer
Synthesis and findings	Were all pre-defined analyses reported or departures explained?	No information *(no information was given regarding the planned analysis of the results)*
Synthesis and findings	Was the synthesis appropriate given the nature and similarity in the research questions, study designs and outcomes across included studies?	Yes *(MA was used appropriately for this review question, study designs were appropriate for analysis)*
Synthesis and findings	Was between-study variation (heterogeneity) minimal or addressed in the synthesis?	Probably yes *(forest plots are present however the individual study details are blurry likely due to a copy and paste error, however heterogeneity appears reasonable)*
Synthesis and findings	Were the findings robust, e.g. as demonstrated through funnel plot or sensitivity analyses?	Yes *(funnel plots are present however the study details on the pasted images are blurry so it is difficult to interpret, likely a copy and paste error by the authors)*
Synthesis and findings	Were biases in primary studies minimal or addressed in the synthesis?	Probably yes *(no obvious discussion about this. The overall GRADE rating is very high given that a lot of the studies were rated moderate or very high risk of bias, as many were case control without knowing where they recruited healthy controls and self-reported)*
Synthesis and findings	Concerns regarding the synthesis and findings	Low *(methods for synthesis appear appropriate)*
Judging risk of bias	Concerns regarding specification of study eligibility	Low *(all parameters appear appropriate to review)*
Judging risk of bias	Concerns regarding methods used to identify and/or select studies	High *(subject headings not translated. Use of auto indexed human limits)*
Judging risk of bias	Concerns regarding methods used to collect data and appraise studies	Low *(methods seem appropriate for review)*
Judging risk of bias	Concerns regarding the synthesis and findings	Low *(methods for synthesis appear appropriate)*

Section	Question	Answer
Overall review ratings	Overall risk of bias	Low *(all areas low concern except for search terms)*
Overall review ratings	Applicability as a source of data	Fully applicable

Evidence to recommendation justification: The review found a higher prevalence of endometrial cancer in women with PCOS than in people without PCOS. This was inversely associated with age at menarche and positively associated with age at menopause. Endometrial cancer is very uncommon in premenopausal women. A reduced risk of 40% was seen with use of oral contraceptives regardless of the time since cessation. Subgroup analysis found women above the healthy weight range had a higher risk of developing endometrial cancer. Smoking also had a relationship of risk. Women with PCOS were found to have significantly higher odds of endometrial cancer and endometrial hyperplasia across all meta-analysis groups, compared to women without PCOS. This is supportive of the EBR. The evidence for endometrial hyperplasia alone was presented in 2 studies which had mixed outcomes. As such the only evidence for endometrial hyperplasia to support the recommendation came from studies that also included endometrial cancer. The IG deemed this sufficient to support the recommendation, however it would be more robust to have evidence from studies on endometrial hyperplasia alone. The strength of recommendation seems quite strong given the evidence available. No associations were found for age of first/last birth, breastfeeding, menopausal status, HRT use, or history of uterine fibromyomas. Research priorities include impact of androgen excess on endometrial cancer, incidence of endometrial cancer in different ethnic groups, and long-term history of endometrial cancer.

IG economic evidence

No health economic evidence was identified in the IG for review question 1.11 on endometrial hyperplasia and cancer.

The IG noted that the review panel who developed review 1.11 had mixed views on cost savings versus spending and was uncertain about the cost-effectiveness of screening – noting that endometrial cancer is common but typically uncommon in premenopausal women.

The IG also acknowledged that the knowledge base of non-gynaecologist health professionals regarding the assessment of the endometrium is variable and therefore the feasibility of implementing the IG recommendations will be dependent on the health system and resource variability.

1.11.2 NICE economic evidence

Included studies

A single health economic search was performed by NICE to identify published economic evaluations of relevance to all review questions in this guideline. See the literature search strategy in Appendix A.

No economic studies were identified which were applicable to this review question (see economic study selection flow chart in Appendix B).

Excluded studies

No economic studies were reviewed at full text and excluded from this review.

Economic model

No original health economic model was conducted for review question 1.11.

1.11.3 NICE recommendations

The relevant recommendations for this section are Rec 1.19.1 to 1.19.3.

1.11.4 The committee's discussion and interpretation of the evidence

Clinical

The IG made one EBR (1.11.1) and four practice point recommendations (1.11.2 – 1.11.5) on endometrial hyperplasia and cancer. The committee decided to contextualise all recommendations from the IG on this topic, except for practice point 1.1.3.

The committee contextualised one EBR on endometrial hyperplasia to raise awareness that women with premenopausal women with PCOS have a markedly higher risk of developing endometria hyperplasia and endometrial cancer compared to premenopausal women who do not have PCOS. The committee contextualised all but one of the practice points from this section of the IG to provide further information for women with PCOS who have not experienced menopause about the risks and discuss the strategies to prevent endometrial hyperplasia and endometrial cancer. The committee contextualised the practice point 1.11.5 on when to refer for biopsy with the addition of offering regular progestogen therapy or the combined contraceptive pill, for endometrial protection, from practice point 1.11.4.

Health economic

No health economic evidence was identified in the IG on endometrial hyperplasia and cancer. In addition, no health economic was identified in NICE's health economic literature search for this review question.

Ultimately the committee made three recommendations relating to endometrial hyperplasia and cancer, which were based on the four recommendations they decided to contextualise. The committee noted that the first two recommendations they made were concerned with explaining the potential risk of endometrial cancer, noting that the overall risk is low – and the steps people can take to reduce their overall risk. The committee acknowledged that these recommendations are reflective of current best practice and therefore the associated costs of implementing these

recommendations would be small. The committee also noted that providing information on how to reduce the overall risk of developing endometrial cancer has the potential to reduce the number of people with PCOS who go on to develop the disease. The committee acknowledged that the exact benefit of providing this information is difficult to quantify but given the small cost associated with providing this information, doing so is highly likely to be cost-effective. The committee wanted to reiterate that the overall risk of developing endometrial cancer is low but concluded that it was important all people with PCOS were made aware of the risks as to optimise health outcomes. The committee noted that risks can be lowered by optimising weight management, healthy eating, physical activity, cycle regulation and receiving regular progestogen therapy.

The final recommendation the committee contextualised on this topic was concerned with considering a biopsy for those when excessive endometrial thickness is detected either incidentally on an ultrasound scan or on an ultrasound scan for identifying PCOM. The committee acknowledged that this recommendation is also reflective current practice and therefore not anticipated to result in a significant resource impact.

1.12 Risks in relatives

Review question 1.12: What is the risk of PCOS and cardiometabolic outcomes (CVD, T2D) in relatives of women with PCOS?

1.12.1 Recommendations from the International evidence-based guideline for PCOS*

Evidence-based recommendation:

1.12.1 Healthcare professionals could consider that fathers and brothers of women with PCOS may have an increased prevalence of metabolic syndrome, type 2 diabetes, and hypertension.

Practice point:

1.12.2 The cardiometabolic risk in female first-degree relatives of women with PCOS remains inconclusive.

IG clinical evidence

Critical appraisal - ROBIS systematic review checklist

Section	Question	Answer
Study eligibility criteria	Did the review adhere to pre-defined objectives and eligibility criteria?	Yes *(PICO was clear and detailed, search dates clearly listed)*
Study eligibility criteria	Were the eligibility criteria appropriate for the review question?	Yes *(detailed PICO available which is suitable for the review. 4 questions were included in this review, of which each had their own inclusion criteria)*
Study eligibility criteria	Were eligibility criteria unambiguous?	Yes *(eligibility criteria clearly described in PICO)*

Section	Question	Answer
Study eligibility criteria	Were all restrictions in eligibility criteria based on study characteristics appropriate (e.g. date, sample size, study quality, outcomes measured)?	Probably yes *(restrictions appear appropriate and differ by different question's inclusion criterion such as different study type. This is a new review so no date limit, but this is not clearly reported)*
Study eligibility criteria	Were any restrictions in eligibility criteria based on sources of information appropriate (e.g. publication status or format, language, availability of data)?	Probably yes *(English language filter used (but not noted in PICO), no other limits noted)*
Study eligibility criteria	Concerns regarding specification of study eligibility criteria	Low *(all parameters are appropriate to review)*
Identification and selection of studies	Did the search include an appropriate range of databases/electronic sources for published and unpublished reports?	Yes *(7 named databases used, including 2 clinical trials databases for ongoing trials. Appears appropriate to review question, except PsycINFO superfluous)*
Identification and selection of studies	Were methods additional to database searching used to identify relevant reports?	Probably yes *(Other sources mentioned in PRISMA but does not detail any additional searches/manual search strategies however the general methods says that they sought relevant papers from their experts)*
Identification and selection of studies	Were the terms and structure of the search strategy likely to retrieve as many eligible studies as possible?	No *(structure sound but subject headings translated. A few free text terms missing in the relative's block. Trials registries search rather underdeveloped.)*
Identification and selection of studies	Were restrictions based on date, publication format, or language appropriate?	Yes *(only language applied)*
Identification and selection of studies	Were efforts made to minimise error in selection of studies?	Yes *(2 reviewers plus evidence team for consultation - appropriate for reducing error in study selection)*
Identification and selection of studies	Concerns regarding methods used to identify and/or select studies	Low *(minor comments on search quality).*

Section	Question	Answer
Data collection and study appraisal	Were efforts made to minimise error in data collection?	Yes *(2 reviewers plus evidence team)*
Data collection and study appraisal	Were sufficient study characteristics available for both review authors and readers to be able to interpret the results?	Yes *(detailed study characteristics table available, table is sufficiently populated for result interpretation and individual study extraction provided)*
Data collection and study appraisal	Were all relevant study results collected for use in the synthesis?	Probably yes *(26 studies included in MA and GRADE, 30 unique studies in SR, data from all studies appears to have been included but this is not detailed)*
Data collection and study appraisal	Was risk of bias (or methodological quality) formally assessed using appropriate criteria?	Yes *(quality appraisal templates used for all cross sectional and case control studies)*
Data collection and study appraisal	Were efforts made to minimise error in risk of bias assessment?	Probably yes *(yes- 2 reviewers plus evidence review team)*
Data collection and study appraisal	Concerns regarding methods used to collect data and appraise studies	Low *(methods seem appropriate for review)*
Synthesis and findings	Did the synthesis include all studies that it should?	Yes *(26 studies included in MA and GRADE, 30 unique studies in SR, data from all studies appears to have been included but this is not detailed)*
Synthesis and findings	Were all pre-defined analyses reported or departures explained?	No information *(no information was given regarding the planned analysis of the results)*
Synthesis and findings	Was the synthesis appropriate given the nature and similarity in the research questions, study designs and outcomes across included studies?	Yes *(Meta-analysis and descriptive analysis were used appropriately for this review question, study designs were appropriate for analysis. A lot of the studies could not be meta-analysed together due to differences in population)*
Synthesis and findings	Was between-study variation (heterogeneity) minimal or addressed in the synthesis?	Probably yes *(serious inconsistencies noted in results tables and downgraded in GRADE)*

Section	Question	Answer
Synthesis and findings	Were the findings robust, e.g. as demonstrated through funnel plot or sensitivity analyses?	Probably yes *(no funnel plot presumably due to not enough studies per synthesis to conduct one)*
Synthesis and findings	Were biases in primary studies minimal or addressed in the synthesis?	Probably yes *(no obvious discussion about this. GRADE ratings are very high for observational studies. No GRADE footnote 1-9)*
Synthesis and findings	Concerns regarding the synthesis and findings	Low *(no obvious issues raised)*
Judging risk of bias	Concerns regarding specification of study eligibility	Low *(all parameters are appropriate to review)*
Judging risk of bias	Concerns regarding methods used to identify and/or select studies	Low *(minor comments on search quality).*
Judging risk of bias	Concerns regarding methods used to collect data and appraise studies	Low *(methods seem appropriate for review)*
Judging risk of bias	Concerns regarding the synthesis and findings	Low *(no obvious issues raised except GRADE ratings are very high for observational and no GRADE footnotes for 1-9)*
Overall review ratings	Overall risk of bias	Low *(Study eligibility, search, methods of collecting data were appropriate for review. GRADE ratings high and missing footnotes)*
Overall review ratings	Applicability as a source of data	Fully applicable

Evidence to recommendation justification: The evidence in the review found that cardiometabolic risk in first degree relatives of women with PCOS was inconclusive. Small studies examined mothers and daughters with PCOS however this was limited due to low numbers of participants. An implementation consideration raised is how healthcare professionals should screen male first-degree relatives. As a result, greater research is needed in families of women with PCOS across all age groups,

particularly in mothers and daughters with PCOS. The review included 25 cross-sectional studies and 6 cohort studies, 1 study had a high risk of bias, the remaining were moderate risk of bias. For the outcome of metabolic syndrome, 6 studies were suitable for meta-analysis. This found first degree relatives of women with PCOS were more likely to suffer from a metabolic syndrome than first degree relatives of control women, moderate quality evidence due to larger number of studies. Further subgroup analyses were very low-quality evidence due to small sample sizes; however, they found that offspring of women with PCOS did not appear to have a higher incidence of diabetes, obesity or metabolic syndrome. Fathers and mothers of women with PCOS however were more likely to suffer from metabolic syndrome, with fathers also more likely to have hypertension and diabetes. Brothers of women with PCOS were more likely to have metabolic syndrome, but this was not seen in sisters. The authors note that not all studies included the outcomes of hypertension and diabetes. Daughters and sisters of women with PCOS did not appear to be at increased risk of PCOS themselves, however there were serious imprecision issues due to very small sample sizes, as such the practice point here is sufficient. The EBR although cautious in its wording may cause concern in fathers and brothers of women with PCOS when the evidence base for this does not appear particularly strong for diabetes or hypertension.

IG economic evidence

No health economic evidence was identified in the IG for review question 1.12 on risk in relatives.

The IG noted that there will be a cost for implementing screening of diabetes in relatives, but this may be offset by prevention of complications.

1.12.2 NICE Economic evidence

Included studies

A single health economic search was performed by NICE to identify published economic evaluations of relevance to all review questions in this guideline. See the literature search strategy in Appendix A.

No economic studies were identified which were applicable to this review question (see economic study selection flow chart in Appendix B)

Excluded studies

No economic studies were reviewed at full text and excluded from this review.

Economic model

No original health economic model was conducted for review question 1.12.

1.12.3 NICE recommendations

No recommendations were made for this section.

1.12.4 The committee's discussion and interpretation of the evidence

Clinical

The committee decided not to contextualise any recommendations from this section of the IG as the evidence in this area was inconclusive.

Health economic

No health economic was identified in the IG or NICE's health economic literature search for review question 1.12 on risks in relatives.

The committee decided not to contextualise the recommendations made in the IG on this topic as the evidence was inconclusive.

Appendix A Health economic literature review search strategy

The searches for the cost effectiveness evidence were run on 5 December 2024 and re-run on 25 March 2026. The following databases were searched: Medline (Ovid), Embase (Ovid; Econlit (Ovid) and the International HTA Database. Limits were applied to remove study types. The validated NICE cost utility filter was used on MEDLINE and Embase. English language limits were applied, and the search was run for evidence published since 2009.

A NICE Senior Information Specialist (SIS) conducted the searches. The MEDLINE strategy was quality assured by another NICE SIS. All translated search strategies were peer reviewed to ensure their accuracy. Both procedures were adapted from the 2015 PRESS Guideline Statement.

The Medline strategy is presented below

1 Polycystic Ovary Syndrome/

2 ((polycystic or poly cystic) adj4 ovar*).tw.

3 pco*.tw.

4 ((degenerat* or sclerocystic) adj4 ovar*).tw.

5 stein leventhal.tw.

6 Anovulation/

7 anovulat*.tw.

8 (oligo ovulat* or oligoovulat*).tw.

9 ((hyperandrogen* or hyper androgen*) adj4 ovar*).tw.

1 10 or/1-9 55812

2 11 Economics/

3 12 Value of life/

4 13 exp "Costs and Cost Analysis"/

5 14 exp Economics, Hospital/

6 15 exp Economics, Medical/

7 16 Economics, Nursing/

8 17 Economics, Pharmaceutical/

9 18 exp "Fees and Charges"/

10 19 exp Budgets/

11 20 budget*.ti,ab.

12 21 cost*.ti.

13 22 (economic* or pharmaco?economic*).ti.

14 23 (price* or pricing*).ti,ab.

15 24 (cost* adj2 (effective* or utilit* or benefit* or minimi* or unit* or estimat* or
16 variable*)).ab.

17 25 (financ* or fee or fees).ti,ab.

18 26 (value adj2 (money or monetary)).ti,ab.

19 27 or/11-26

1 28 10 and 27

2 29 letter.pt. or letter/

3 30 note.pt.

4 31 editorial.pt.

5 32 case report/ or case study/

6 33 (letter or comment*).ti.

7 34 or/29-33

8 35 randomized controlled trial/ or random*.ti,ab.

9 36 34 not 35

10 37 animals/

11 38 exp Animals, Laboratory/

12 39 exp Animal Experimentation/

13 40 exp Models, Animal/

14 41 exp Rodentia/

15 42 (rat or rats or mouse or mice or rodent*).ti.

16 43 or/37-42

17 44 43 not humans/

18 45 36 or 44

19 46 28 not 45

1 47 limit 46 to english language/

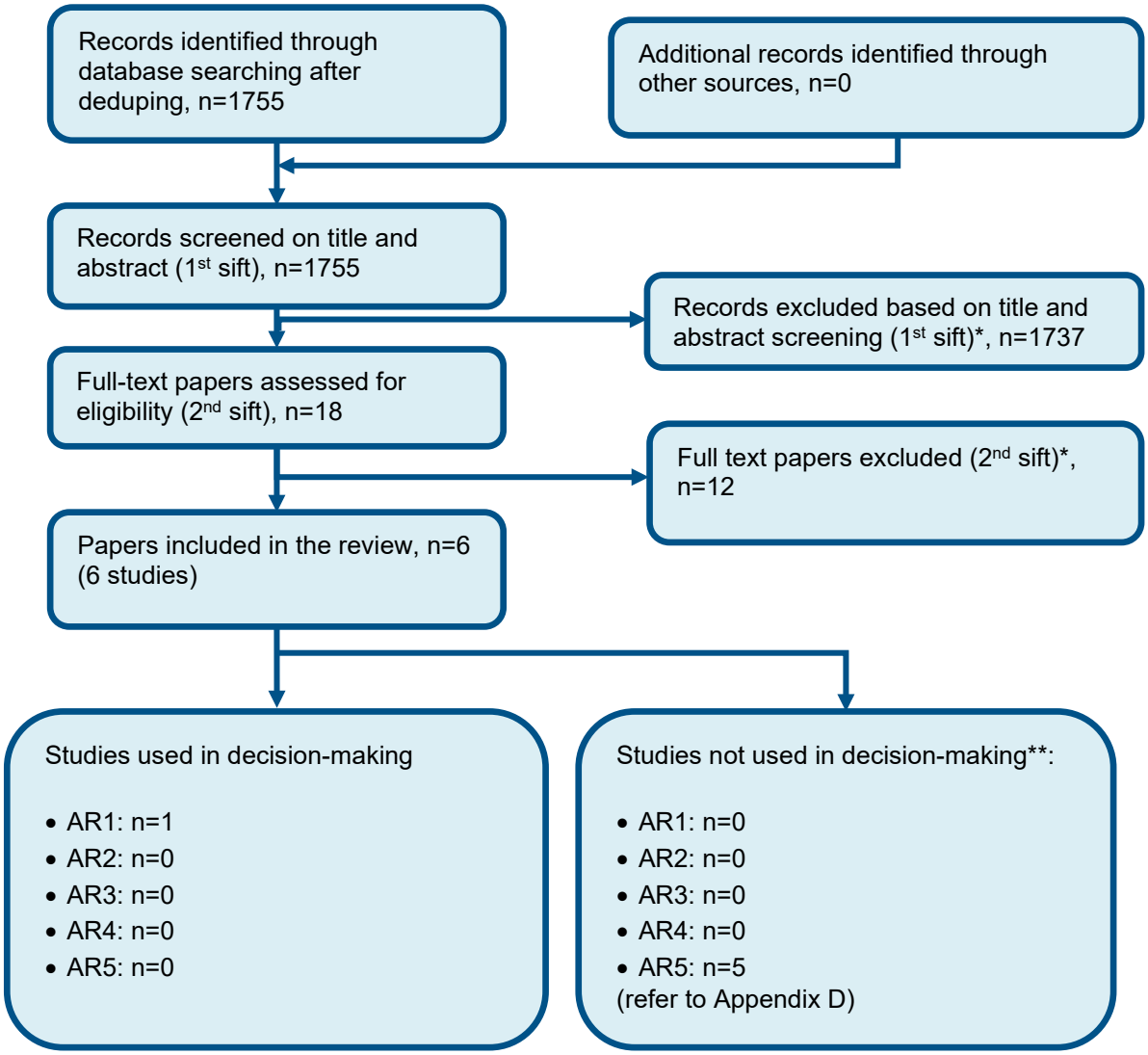
2 48 limit 47 to ed=20090101-20241

3 49 limit 47 to dt=20090101-20241205

4 50 48 or 49

5

Appendix B Health economic PRISMA diagram



* Not an economic evaluation, non-relevant population, intervention, comparison, design, setting or perspective; non-English language, not a full paper

**please refer to Review strategy described in the Economic review protocol in Methods document (Appendix B)

Appendix C Economic evidence tables

Garay, 2025

Table 4: Economic evidence study extraction table: Garay, 2025

Section	Details for Garay, 2025
Study details	**Economic analysis type:** Cost-consequences **Analysis design:** Decision analytic model (decision tree) **Country setting:** UK **Perspective:** NHS and PSS **Time horizon:** lifetime **Treatment duration:** NA **Discount rate per year:** Costs: NR; Outcomes: NR (discounting was applied for the lifetime cost of type 2 diabetes (Wang 2022) and the lifetime cost of stroke (Xiang-Ming (2018) but the rate at which these costs were discounted was not reported)
Interventions	**Intervention 1:** Transvaginal Ultrasound (TVUS) – which was requested if after an initial GP appointment only irregular menstrual cycles or hyperandrogenism was confirmed to assess polycystic ovary morphology (PCOM). **Intervention 2:** Anti-Müllerian hormone (AMH) – in this case the Elecsys AMH plus immunoassay using a cut-off of 3.2bg/mL for identifying PCOM. An AMH test was conducted when a person with suspected PCOS presented an initial GP appointment.
Population	**Population:** Women with suspected PCOS **Baseline characteristics** Age range = 25 – 40 years Male = 0%
Costs included	**Original currency & cost year:** 2022 UK pounds **Cost components incorporated:** Elecsys AMH Plus immunoassay, TVUS scan, GP appointment, obstetrics and gynaecology appointment, dermatology appointment, endocrinology appointment, lifestyle intervention, stroke cost and type 2 diabetes cost.
Outcomes included	**Primary health outcome(s) in economic analysis:** Cost of a PCOS diagnosis **Key events modelled:** cost of a PCOS diagnosis, cardiovascular risk, type 2 diabetes risk (and their associated costs).
Data Sources	**Effectiveness data:** Liu 2021, Gabrielli 2012, Lizneva 2016, HAMONIA (internal analysis by Roche to estimate the proportion of people who do not present with clinical hyperandrogenism but do have biochemical hyperandrogenism), Riestenberg 2022, Costa 2012, Lindström 2013, **Baseline / epidemiological data:** Sensitivity & specificity of AMH *APHRODITE study* (Dietz de Loos, 2021) Sensitivity and specificity of TVUS – meta-analysis from the IG **Quality-of-life weights:** NA **Costs and/or resource use:** AMH cost - Roche Diagnostics UK. Base-case is a list price. Ranges were calculated as +/- 15 GBP, TVUS scan by diagnostic imaging service, TVUS scan by gynaecology service, Outpatient appointment OBGYN, Outpatient appointment GP, Dermatology/endocrinology outpatient

Section	Details for Garay, 2025
	appointment (NHS reference costs 2021/2022 and unit costs for health and social care 2022). Cost of lifestyle interventions – estimated using unpublished data from the Sheffield City Council weight loss program Discounted lifetime cost of T2D – Wang 2022 Discounted lifetime costs of stroke – Xiang-Ming 2018
Results: costs	**Total costs of PCOS per year (per diagnosis):** TVUS: £1,536 AMH: £1,514 Incremental (2−1): Cost saving of £22 (95% CI: NR; p=NR) **Total costs of PCOS (per year) for those people presenting with signs and symptoms of PCOS at their GP practice** TVUS: £19,941,734 AMH: £19,657,705 Incremental (2−1): Cost saving of £284,029 **Diagnosis of PCOS costs components (per year)** <u>Consultations/consultation time</u> TVUS: £761,180 AMH: £574,751 Incremental (2−1): Cost savings of £186,429 <u>Laboratory costs</u> TVUS: £269,222 AMH: £269,222 Incremental (2−1): £0 <u>AFC with TVUS</u> TVUS: £512,938 AMH: £0 Incremental (2−1): £512,938 <u>Elecsys AMH Plus immunoassay</u> TVUS: £0 AMH: £403,934 Incremental (2−1): £403,934 **Other cost components (per year)** <u>Lifestyle interventions</u> TVUS: £2,512,189 AMH: £2,689,702

Section	Details for Garay, 2025
	Incremental (2−1): £177,513 <u>T2D</u> TVUS: £6,459,543 AMH: £6,392,000 Incremental (2−1): Cost saving of £67,542 <u>Stroke</u> TVUS: £9,426,662 AMH: £9,328,095 Incremental (2−1): Cost savings of £98,567
Results: health outcomes	**PCOS diagnoses per result type (per year):** **True positive** TVUS: 6,584 AMH: 6,721 Incremental (2−1): 136 (95% CI: NR; p=NR) **False positive** TVUS: 344 AMH: 697 Incremental (2−1): 353 (95% CI: NR; p=NR) **True negative** TVUS: 5,478 AMH: 5,125 Incremental (2−1): 353 fewer true negatives with AMH (95% CI: NR; p=NR) **False negative** TVUS: 581 AMH: 444 Incremental (2−1): 136 fewer false negatives with AMH (95% CI: NR; p=NR) **New cases of T2D (per year)** TVUS: 187 AMH: 185 Incremental (2−1): 2 fewer cases with AMH (95% CI: NR; p=NR) **New cases of stroke (per year)** TVUS: 205 AMH: 203 Incremental (2−1): 2.1 fewer cases with AMH

Section	Details for Garay, 2025
	(95% CI: NR; p=NR)
Results: cost effectiveness	NA
Results: Uncertainty	**Deterministic:** Deterministic sensitivity analysis conducted altering the base case values to an upper and lower limit to assess the total cost differences for TVUS and AMH. Impact on results are listed below in chronological order, with the changes where the input parameter has the greatest impact on results listed first. • Sensitivity of TVUS for PCOM • AMH unit cost • Relative risk reduction for T2D and stroke associated with lifestyle interventions • Unit cost of TVUS conducted by a radiologist • Sensitivity of AMH for PCOM • Unit cost for a GP appointment • Specificity for TVUS for PCOM • Specificity of AMH for PCOM • Unit cost for lifestyle interventions • Risk of stroke AMH was cost saving in all scenarios apart from when the sensitivity of TVUS was increased. The results are presented graphically so it is unclear at exactly what point TVUS becomes cost-effective, but TVUS highly likely to be cost-effective when the sensitivity is greater than or equal to 95%. The following scenario analyses were conducted: 1. TVUS requested for all PCOS suspicions versus with the AMH test (base-case): this scenario assumed that due to delays in TVUS access, GPs requested both laboratory tests and TVUS at the first consultation, and that there was a second consultation to discuss laboratory results, as well as a third to discuss the TVUS – a. Cost saving of £51 per PCOS diagnosis (per year) with AMH (HARMONIA data, cost saving of £43.80) 2. With TVUS for AFC (base-case) versus with the AMH test (if test is required only after HA): this scenario assumed that the Elecsys AMH Plus immunoassay was required only after the first set of laboratory tests, and only for women requiring a PCOM diagnosis to rule-out or rule-in PCOS a. Cost saving of £3.80 per PCOS diagnosis (per year) with AMH (HARMONIA data, cost saving of £7.10) 3. With TVUS for AFC (if GPs had a low referral rate to specialists) versus with the AMH test (base-case): this scenario assumed that 15% of GPs

Section	Details for Garay, 2025
	referred patients with suspicions of PCOS (10% to endocrinologists or dermatologists and 5% to gynaecologists) a. Cost saving of £83.90 per PCOS diagnosis (per year) with AMH (HARMONIA data, cost saving of £96.50) 4. With TVUS for AFC (if GPs had a high referral rate to specialists) versus with the AMH test (base-case): this scenario assumed that 30% of GPs referred patients with suspicions of PCOS (20% to endocrinologists or dermatologists and 10% to gynaecologists) a. Cost saving of £146 per PCOS diagnosis (per year) with AMH (HARMONIA data, cost saving of £162.70) 5. With TVUS for AFC (considering a 10% drop out rate before TVUS) versus with the AMH test (base-case, no drop out): this scenario assumed that 10% of patients drop out of the diagnosis pathway when a TVUS is requested to evaluate PCOM. In the Elecsys AMH Plus immunoassay scenario, the drop out rate was assumed to be 0% a. Cost saving of £37.40 per PCOS diagnosis (per year) with AMH (HARMONIA data, cost saving of £45.50) 6. With TVUS for AFC (considering a 25% drop out rate before TVUS) versus with the AMH test (base-case, no drop out): this scenario assumed that 25% of patients drop out of the diagnosis pathway when a TVUS is requested to evaluate PCOM. In the Elecsys AMH Plus immunoassay scenario, the drop out rate was assumed to be 0% a. Cost saving of £60.60 per PCOS diagnosis (per year) with AMH (HARMONIA data, cost saving of £68.30) 7. With TVUS for AFC (considering a 50% drop out rate before TVUS) versus with the AMH test (base-case, no drop out): this scenario assumed that 50% of patients drop out of the diagnosis pathway when a TVUS is requested to evaluate PCOM. In the Elecsys AMH Plus immunoassay scenario, the drop out rate was assumed to be 0% a. Cost saving of £99.30 per PCOS diagnosis (per year) with AMH (HARMONIA data, cost saving of £106.40) 8. Lower adherence rate for the lifestyle recommendations: this scenario assumed that only 50% of patients adhered to the lifestyle recommendations, reducing the efficacy of the lifestyle interventions by 50%. a. Cost saving of £15.50 per PCOS diagnosis (per year) with AMH (HARMONIA data, cost saving of £23.40) 9. An additional scenario analysis was performed to replace the calibration estimates with data from the HARMONIA study to assess the potential impact of the calibration methods on the estimation of PCOS characteristics in women with signs and symptoms of PCOS

Section	Details for Garay, 2025
	a. Cost saving of £30.30 with AMH per PCOS diagnosis (per year) in the base case analysis (results for HARMONIA data for other scenarios presented above) **Probabilistic:** Probabilistic sensitivity analysis presented graphically for all scenario analyses conducted, illustrating the additional costs per diagnosis and additional true positive results, per year.
Health inequalities assessment	NR
Comments	**Source of funding:** Roche Diagnostics International Ltd **Other:** AMH identified more false positives than TVUS, therefore a proportion of the cost savings might be offset by this.
Rating: Applicability	**Partially applicable** • UK study • No specific mention as to what discount rate was employed • Lifestyle interventions were provided to all women with a PCOS diagnosis, which is not in line with our contextualisation recommendations
Rating: Quality/ limitations	**Potentially serious limitations** • Although a lifetime horizon was employed, only T2D & stroke outcomes were modelled over the lifetime horizon • There was a discrepancy in the population for AMH and TVUS. The study did however note, that as PCOS phenotype B (i.e., PCOS without PCOM) is quite limited, this discrepancy is unlikely to have a notable impact on the results. • It was assumed that no TVUS was performed when AMH was conducted for PCOS assessment. The study did however note that some women will still require TVUS • The costs of lifestyle interventions employed in the model were high and derived from a small study based on a 12-week program conducted in Sheffield in the UK • The analysis assumed that lifestyle interventions reduced the risk of stroke and T2D. The risk reduction for the risk of stroke and T2D was based on a previous study on the prevention of T2D, rather than PCOS. These clinical estimates therefore may not be directly transferrable and accurately reflect outcomes in a PCOS population.

Abbreviations: CI= confidence interval; DA=deterministic analysis;
PSA=probabilistic sensitivity analysis; NA=not applicable; NR=not reported;
PSS= Personal Social Services

Appendix D Excluded health economic studies

None

Appendix E References

Berni TR, Morgan CL, Rees DA (2025) Rising Incidence, Health Resource Utilization, and Costs of Polycystic Ovary Syndrome in the United Kingdom. J Clin Endocrinol Metab 110(5): e1580–e1589.

Demis Bizuneh, A., Joham, A., Thien Tay, C., Kiconco, S., Earnest, A., Ram Dhungana, R., Suturina, L.V., Zhao, X., Gambineri, A., Ramezani Tehrani, F., Yildiz, B.O., Kim, JJ., Xu, L., Chigozie Makwe, C., Teede, H.J., Azziz, R. (2025) The PCOS Phenotype in Unselected Populations study: ethnic variation in population-based normative cut-offs for defining hirsutism, European Journal of Endocrinology, 192 (3), 228–239.

Garay, Osvaldo Ulises, Olziersky, Anna-Maria, Laven, Joop et al. (2025) Economic impact of the Elecsys anti-Mullerian hormone Plus immunoassay for anti-Mullerian hormone testing as part of polycystic ovary syndrome assessment in the United Kingdom. PLoS ONE [Electronic Resource] 20(6): e0326162

NHS Cost Collection 2024/2025 [online; accessed 29th April 2026]

Copyright for reproduced material

https://doi.org/10.26180/24003834.v1

https://doi.org/10.26180/23625288.v1

Suggested citation: Helena Teede et al.

International Evidence-based Guideline for the Assessment and Management of Polycystic Ovary Syndrome 2023. Monash University.

https://doi.org/10.26180/24003834.v1

Melbourne, Australia 2023

Monash University ISBN: 978-0-6458209-0-4