Supplementary File III: Appraisal tool to assess risk of bias for included studies reporting PCOS prevalence

The 10 criteria used to assess the methodological quality of studies reporting PCOS prevalence are presented here. Explanations are provided for each question accordingly. These questions can be answered as 'yes', 'no', 'unclear' or 'not applicable'. Questions might be answered as 'yes' should be scored with 1 point, while all other answers should be assigned 0 points.

1. Was the sample representative of the target population?*
   Since testing every individual of the whole population is not practical or possible when conducting a study, a 'sampling' is required to be representative of the target population. In this context, the characteristics of target population should be defined with regard to demographic features, diagnosis, type of treatment and/or other confounding factors. It should also be noted that a sample might not be representative of the target population if participants are recruited from a certain group, such as those working for one organization/profession. Therefore, the method of random sampling, with or without stratification, should be employed and clearly recorded to ensure that the sample is representative of the target population.

2. Were study participants recruited in an appropriate way?*
   The process of 'recruitment' includes identifying eligible participants according to the study protocol, informing the potential participants about the study without provoking excessive or insufficient interest, maintaining ethical standards and following participants until study completion. Notably, the calling or advertising strategy should be performed with caution in order to avoid the disproportionate response of a particular subgroup of patients. The method and process of recruitment should be clearly reported, indicating whether all potential participants that should be included, and whether the whole population of interest had in fact been surveyed. If not, information might be provided as to whether a random sampling of a defined subset of the target population was used or whether a stratified random sampling using eligibility criteria was preferred to ensure that the sample was representative of the target population.

3. Was the sample size adequate?*
   Not only the sampling method but also the 'sample size' is critical for deciding whether the study group is representative of the target population. Although sample size calculation has been generally employed in interventional studies, it is also valid in prevalence studies. The mathematical method and attributed ratios that the authors use should be mentioned in their prevalence studies. However, large and national survey studies might be excepted from the requirement of sample size calculation.

\[ n = Z^2 \frac{P(1-P)}{d^2}, \]

where, \( n \) = sample size; \( Z = Z \) statistic for a level of confidence; \( P = \) expected prevalence or proportion (in proportion of one; if 20%, \( P = 0.2 \); \( d = \) precision (in proportion of one; if 5%, \( d = 0.05 \)).

According to the above-mentioned formula, we calculated the optimal sample size for PCOS prevalence studies. The calculated sample size of expected PCOS prevalence was accepted as 5.6% for the subset criteria of NIH and 10.0% for the subset criteria of Rotterdam. Detection precision was considered as 2.0% for NIH criteria and 3.0% for Rotterdam criteria. As a result, we found that the required minimal sample size for NIH criteria and Rotterdam criteria are 408 and 384 patients, respectively. We considered sample size to be adequate if the study population consists of more than 408 patients.

4. Were the study subjects and setting described in detail?*
   Regarding the proportion of hirsutism, various conditions might present diversity across different geographic regions and populations. Similarly to criterion 1, the study population should be described in sufficient detail. In this way, other researchers can determine if the study results are relevant to the population of interest to them.

5. Is the data analysis conducted with sufficient coverage of the identified sample?*
   Apart from recruitment, the authors should report the number of dropouts and their reasons. A high proportion of dropouts and refusals might call into question the validity of the findings of a given study. By documenting the reasons for dropouts and comparing their characteristics with other participants, readers can decide whether the reported prevalence rate is reliable or misestimated. Obviously, reasons unrelated to the study outcome and similarity between the characteristics of dropouts and respondents can justify those findings. By contrast, if the type of assessment/measurement is the leading aetiology for dropouts, then the results cannot be valid.

6. Was the same mode of data collection used for all subjects?*
   The approaches to collecting information from the subjects might be various. Face-to-face or telephone interviews and self-administered questionnaires are some options that might be encountered in prevalence studies. The linked knowledge of the type of data collection should be uniform and given in the methods section.

7. Was the hirsutism scoring and definition performed with standard and objective criteria based on population characteristics?*
   The amount of terminal hair growth should be assessed using the modified Ferriman–Gallway (mF–G) method in which the upper lip, chin, chest, upper and lower abdomen, thighs, upper and lower back and upper arms are scored. The cut-off level of hirsutism was defined as exceeding an mF–G score of 6 or 8.

8. Were reliable hyperandrogenaemia measurement methods used?*
   Biochemical hyperandrogenism (hyperandrogenaemia) involves any androgen—including total testosterone (TT), androstenedione (A), dehydroepiandrosterone sulphate (DHEAS) and/or the free androgen index (FAI) level—exceeding the respective 95th percentile of healthy, non-hirsute, eumenorrheic women without PCO.

9. Was oligo-anovulation defined according to correct terminology, not merely the patients' own reports?*
   Menstrual cycles \( \geq 35 \) or \( \leq 21 \) days were defined as oligo-anovulation. In patients with hirsutism or PCO appearance who had apparently regular menstrual bleeding, luteal phase (Days 21–24) progesterone levels should be determined. The threshold for the presence of ovulation was taken as \( >4 \) ng/ml.

10. Was the ultrasonography performed on the whole target population by measuring both antral follicle count (AFC) and ovarian volume to identify PCO?*
   An AFC of \( \geq 12 \) within a 2–9 mm diameter and/or ovarian volume of \( \geq 10 \) cm\(^3\) in at least a single ovary was defined as PCO.
This appraisal tool was inspired by recent protocols produced by Munn et al. (2014) and Hoy et al. (2012). The first five questions (Q1–Q5) were from Munn et al. while Q6 was from Hoy et al.

References
