

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|--|
| n/a | Confirmed |
| ☐ | ☒ The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☒ | ☐ A description of all covariates tested |
| ☐ | ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☐ | ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

- | | |
|-----------------|---|
| Data collection | No special software or code was used for data collection. Data preparation was done by SPSS version 26. |
| Data analysis | The statistical analysis of clinical data was performed using SPSS version 26. The unsupervised clustering models were implemented by R (4.0.3). The codes can be requested by writing to corresponding authors for scientific research and non-commercial use. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

- All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
 - A description of any restrictions on data availability
 - For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Individual-level patient records of the clinical trial (clinicaltrials.gov: NCT06124391) are not publicly available due to national legal restrictions. The de-identified data can be requested by contacting the corresponding authors (ZC at chen zijiang@hotmail.com or HZ at hanzh80@sdu.edu.cn). All data access requests will be

reviewed. Data can be shared only for non-commercial academic purpose and will require a formal material transfer agreement. The formally signed data transfer agreement is required under the supervision of the corresponding laws and regulations.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	As this study focused on PCOS, all participants were female.
Reporting on race, ethnicity, or other socially relevant groupings	Race/ethnicities were self-reported or based on electronic health records. When validated the proposed four PCOS subtypes, we conducted subgroup analysis based on different races/ethnicities in the America cohort. This cohort included Caucasian, African American, and other races and ethnicities, which were based on participants' country or area of birth.
Population characteristics	All participants in the cohorts were between 18 and 45 years old. PCOS was diagnosed using the Rotterdam criteria.
Recruitment	Participants were recruited from nine clinical centers in China, North America, South America, Singapore, and Europe. Inclusion criteria was age 18-45 with PCOS diagnosed by the Rotterdam criteria. Exclusion criteria were congenital adrenal hyperplasia, androgen-secreting tumors, Cushing's syndrome. The control women were only included for the IVF outcome analysis, recruited from the patients who underwent IVF due to the male factors.
Ethics oversight	Study protocols were approved by the relevant Ethics Committee of each cohorts.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	A total of 52,002 PCOS cases and 4,165 control women were included in the study, where the sample size was determined by the data availability.
Data exclusions	PCOS women who have received such therapies at the time of enrollment, and therefore and did not meet the Rotterdam diagnostic criteria in terms of clinical manifestations or test parameters were excluded.
Replication	Replications was not relevant in this study. We used independent cohorts.
Randomization	Randomization was not relevant in this study. Participants were allocated into different subtypes by unsupervised clustering method.
Blinding	During the data processing, all data was de-identified to remove patient related information.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	For information purpose, all cohorts were registered as an international multicenter study on clinicaltrials.gov (NCT06124391). The American validation cohort data is from a clinical trial study which is registered with ClinicalTrial.gov NCT00719186.
Study protocol	We have provided the full protocol in the manuscript and the flow chart in Figure 1.
Data collection	Participants in the discovery cohort were collected during December 2013 and June 2020. Demographic and medical history, body weight and height, endocrine and metabolic laboratory indicators were recorded. The IVF protocols and outcomes were collected from the medical records. The follow-up data were collected by telephone interview and physical examination.
Outcomes	For the prospective follow-up study, the outcomes were difference in the remission of PCOS and progression of other chronic complications, such as type 2 diabetes, hypertension, and dyslipidemia. For the in vitro fertilization results, the primary outcomes were difference of live birth rate, pregnancy rate, and pregnancy loss among four subtypes. The secondary outcomes were the maternal and neonatal complications.

Plants

Seed stocks	<i>Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.</i>
Novel plant genotypes	<i>Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.</i>
Authentication	<i>Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.</i>