Supplementary Materials

Evidence from Men for Ovary-Independent Effects of Genetic Risk Factors for Polycystic Ovary Syndrome

Jia Zhu, MD,¹,²,³ Natàlia Pujol-Gualdo, MSc,⁴,⁵ Laura B.L. Wittemans, PhD,⁶,⁷ Cecilia M. Lindgren, PhD,²,⁷,⁸ Triin Laisk, PhD,⁴ Joel N. Hirschhorn, MD, PhD,¹,²,³,⁹ Yee-Ming Chan, MD, PhD¹,²,³

¹Division of Endocrinology, Boston Children’s Hospital, Boston, MA, USA.
²Programs in Metabolism and Medical and Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA, USA.
³Department of Pediatrics, Harvard Medical School, Boston, MA, USA.
⁴Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia.
⁵Department of Obstetrics and Gynecology, PEDEGO Research Unit, Medical Research Centre, Oulu University Hospital, University of Oulu, Oulu, Finland.
⁶Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford OX3 ZFZ, UK.
⁷Nuffield Department of Women’s and Reproductive Health, University of Oxford, Oxford, UK.
⁸The Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7FZ, UK.
⁹Department of Genetics, Harvard Medical School, Boston, MA, USA.
Supplemental Methods

UK Biobank cohort

Participants underwent genotyping with one of two closely related arrays, and additional genotypes were imputed using the Haplotype Reference Consortium (HRC) data as the main reference panel (1). Analyses were conducted on genetic data release version 3 under UKB application 11898. All work complied with all relevant ethical regulations related to the UK Biobank, and all participants provided informed consent. Genetic principal components were derived on a subset of unrelated individuals of European ancestry (randomly selected without regard to PCOS case or control status) in the UK Biobank using FlashPCA2 (2). Because the largest PCOS GWAS meta-analysis was primarily conducted in individuals of European ancestry, and genetic risk scores are less valid with current methodologies when applied to populations with different ancestry due to differences in genetic architecture (3,4), the present analysis was restricted to individuals of European ancestry, which was determined by a combination of self-reported ancestry and principal components analysis (1,5). Related individuals were identified by genetic data, and kinship coefficients were reported for all pairs of relatives who were inferred to be third-degree relatives or closer (1). To minimize the possibility of confounding due to genetic relatedness, only one individual was included per group of relatives (third-degree or closer). If any individual had a diagnosis of PCOS, the individual with PCOS was preferentially included in the analysis to optimize the prevalence of PCOS.
Estonian Biobank replication cohort

Analyses in the EstBB were done under ethical approval 1.1-12/624 from the Estonian Committee on Bioethics and Human Research and data release N05 from the EstBB. The 200K data freeze was used for analyses. Genotyping for all participants was completed at the Core Genotyping Lab of the Institute of Genomics, University of Tartu, using Illumina GSAv1.0, GSAv2.0, and GSAv2.0_EST arrays. PLINK format files were created using Illumina GenomeStudio v2.0.4. Individuals were excluded from the analysis if their call-rate was <95% or if sex defined based on heterozygosity of the X chromosome did not match the sex in phenotype data. Before imputation, variants were filtered by call-rate <95%, HWE p-value <1x10^{-4} (autosomal variants only), and minor allele frequency <1%. Variant positions were updated to b37, and all variants were changed to the TOP strand orientation using GSAMD-24v1-0_20011747_A1-b37.strand.RefAlt.zip files from https://www.well.ox.ac.uk/~wrayner/strand/. Prephasing was completed using Eagle v2.3 software (number of conditioning haplotypes Eagle2 uses when phasing each sample was set to: --Kpbwt=20000) (6), and imputation was performed using Beagle v.28Sep18.793 with effective population size ne=20,000 (7). Population specific imputation reference of 2297 WGS samples was used (8). Genetic principal components were derived on the relatedness matrix of Estonian biobank participants. The cohort was restricted to men of European ancestry, which was determined by a combination of self-reported ancestry and principal components analysis (1,5). Related individuals were identified by genetic data, and kinship coefficients were reported for all pairs of relatives who were inferred to be third-degree relatives or closer.
For ascertainment of phenotypes, ICD codes were obtained via linking with the national Health Insurance Fund and other relevant databases (9).

**PCOS polygenic risk score calculation**

We compared three methods for calculating a polygenic risk score (PRS) for PCOS. Genetic variants with ambiguous strand (A/T or C/G) were removed from all polygenic risk score calculations. PRSice-2 uses a “clumping and thresholding” approach to clump genetic variants in close linkage disequilibrium, such that the remaining variants are independent of each other, and includes only those variants with a GWAS association $P$-value below a given threshold, with the threshold chosen to maximize the association of the risk score with PCOS (10). The PRSice-2 software also incorporates both directly genotyped SNP’s and dosage estimation from imputed data, which can improve predictive power of the resulting PRS (10). PRS-CS is a Bayesian approach that applies a continuous shrinkage model to modify effect sizes of SNPs to incorporate information on the strength of each variant’s association in the GWAS and the underlying linkage disequilibrium structure (11). PRS-CS utilizes a tuning (or global shrinkage) parameter, $\phi$, and we tested a range of $\phi$ values (1x10$^{-6}$ to 1) in women in the UK Biobank to identify the value that optimized the association of the PRS for PCOS. For imputed genotypes, PRS-CS does not incorporate dosage estimations for each variant. To account for the probabilities in the imputed genotypes with the PRS-CS software, a third PRS was generated using the PRS-CS software to calculate a weighted effect size for each SNP.
and the PRSice-2 software to incorporate the genotype dosage of each variant to produce a final risk score.

**Ascertainment of Phenotypes**

**PCOS IN WOMEN**

Primary care clinical events used to determine a diagnosis of PCOS included consultations, diagnoses, history, symptoms, procedures, laboratory tests, and/or administrative information.

PCOS cases in hospitalization records were determined by ICD-9/10 codes for either a diagnosis of PCOS (1) or symptoms of PCOS (2) as follows:

1. Diagnosis of PCOS
   - ICD-9 code: 256.4 (PCOS)
   - ICD-10 code of E28.2 (PCOS)

2. Symptoms of PCOS (both required):
   - Irregular menstruation:
     - ICD-9 code: 626 (disorders of menstruation)
     - ICD-10 code: N91.0-N91.5 (amenorrhea) or N92.5-N92.6 (irregular menstruation)
   - Hyperandrogenism:
     - ICD-9 code: 704.1 (hirsutism)
     - ICD-10 code: L68.0 (hirsutism)
OUTCOMES IN MEN

Cardiometabolic phenotypes: Obesity was defined as a BMI ≥ 30 kg/m². Coronary artery disease was based on a composite of myocardial infarction, ischemic heart disease, and/or coronary revascularization (12). Myocardial infarction was defined by self-report, a reported date of myocardial infarction, and/or ICD-9 codes of 410.9, 411.9, and 412.9 and ICD-10 codes of I21.0-21.4, I21.9, I22.0, I22.1, I22.8, I22.9, I23.0-23.3, I23.5-23.6, I23.8, I24.0, or I25.2 in hospitalization records. Coronary revascularization was based on OPCS-4 codes of K40.1-40.4, K41.1-41.4, K45.1-45.5, K49.1-49.2, K49.8-49.9, K50.2, K75.1-75.4, or K75.8-75.9. Type 2 diabetes mellitus was defined based on a previously reported composite algorithm of self-report, ICD-9/10 diagnosis codes from hospitalization records, medication use, and age at diagnosis (13,14). Those with age of diabetes onset less than 40 years were excluded. Controls were individuals aged 55 years or greater with no history of diabetes.

Hyperandrogenic phenotypes: Marked androgenic alopecia was based on self-report of hair/balding pattern on questionnaire based on the Norwood-Hamilton scale, a classification system of male pattern baldness ranging from stage 1 (no baldness) to stage 7 (complete baldness)(15,16). Because a self-report of stages 2 to 4 may represent variations in hairline and/or hair volume rather than true androgenic alopecia, cases were individuals who self-reported marked baldness (stage 5+), and controls were individuals who self-reported no sign of baldness (stage 1)(15). The free-androgen index (FAI), a measure of androgen production, was calculated by dividing the total testosterone level
(nmol/L) by the sex hormone-binding globulin (SHBG) level (nmol/L) and multiplying by the constant 100.

**Statistical analysis of BMI as a mediator**

Causal mediation analysis was used to assess the indirect effect of BMI on the association between the outcome of interest and the PRS. The significance of the indirect effect of BMI was tested using bootstrapping procedures. Indirect effects were computed for 1,000 bootstrapped samples, and the 95% CI was computed. A $p < 0.05$ was considered statistically significant.
Fig. S1. Validation of a PCOS genetic risk score in women in the UK Biobank

(A) Summary of ascertainment of PCOS cases in unrelated women of European ancestry in the UK Biobank by questionnaire (Self-report), report of PCOS diagnosis in primary care data or hospitalization records (Diagnosis), and/or report of irregular menses and hyperandrogenism in primary care data or hospitalization data (Symptoms). (B) Refinement of the prevalence of PCOS by age and menopause status. (C) Odds ratio of PCOS by quintile of the best-performing PCOS genetic risk score in women ≤50 years of age with no history of menopause. PCOS, polycystic ovary syndrome; CI, confidence interval.
Table S1. Association of candidate polygenic risk scores with PCOS in women in the UK Biobank

| Method                  | Cohort     | Tuning Parameter | $R^2$   | $p$     |
|-------------------------|------------|------------------|---------|---------|
| PRSice-2                | All        | $P = 1$          | 0.0020  | $5 \times 10^{-7}$ |
|                         | Subcohort  | $P = 0.036$      | 0.0014  | $2.6 \times 10^{-3}$ |
| PRS-CS                  | All        | $\phi = 1$       | 0.0020  | $5 \times 10^{-7}$ |
|                         | All        | $\phi = 1 \times 10^{-2}$ | 0.0033  | $1 \times 10^{-10}$ |
|                         | All        | $\phi = 1 \times 10^{-4}$ | 0.0044  | $1 \times 10^{-13}$ |
|                         | All        | $\phi = 1 \times 10^{-6}$ | 0.0032  | $3 \times 10^{-10}$ |
| PRS-CS + Dosage of Imputed SNPs | All      | $\phi = 1$       | 0.0022  | $2 \times 10^{-7}$ |
|                         | All        | $\phi = 1 \times 10^{-2}$ | 0.0034  | $9 \times 10^{-11}$ |
|                         | All        | $\phi = 1 \times 10^{-4}$ | 0.0045  | $1 \times 10^{-13}$ |
|                         | All        | $\phi = 1 \times 10^{-6}$ | 0.0031  | $7 \times 10^{-10}$ |
|                         | Subcohort  | $\phi = 1$       | 0.0018  | $6.4 \times 10^{-4}$ |
|                         | Subcohort  | $\phi = 1 \times 10^{-2}$ | 0.0033  | $3 \times 10^{-6}$ |
|                         | Subcohort  | $\phi = 1 \times 10^{-3}$ | 0.0046  | $3 \times 10^{-6}$ |
|                         | Subcohort  | $\phi = 1 \times 10^{-4}$ | **0.0052**  | **2 \times 10^{-9}** |
|                         | Subcohort  | $\phi = 1 \times 10^{-5}$ | 0.0050  | $8 \times 10^{-9}$ |
|                         | Subcohort  | $\phi = 1 \times 10^{-6}$ | 0.0049  | $1 \times 10^{-8}$ |

Polygenic risk scores were calculated in 206,852 unrelated women of European ancestry (“All”) and in a subcohort of 50,612 women who were ≤50 years of age with no history of menopause (“Subcohort”) in the UK Biobank. The first two scores were calculated using the PRSice-2 software using independent variants ($R^2 < 0.1$ with all other variants) with automatically optimized $p$-value thresholds for the PCOS phenotype. The remaining 15 scores were calculated using the PRS-CS algorithm, a Bayesian approach that weighs the effect size of each variant based on the level of statistical significance in the GWAS and a tuning parameter $\phi$ based on the underlying genetic architecture. Because the PRS-CS algorithm does not incorporate dosage information on imputed genotypes, additional scores were calculated using a modified algorithm using PRS-CS to generate modified
SNP effect sizes and PRSice-2 to calculated polygenic risk scores that incorporate dosage information (PRS-CS + Dosage of Imputed SNPs). The proportion of the variance explained ($R^2$) of PCOS was calculated for each polygenic risk score in the specified cohort, and the best-performing method (boldface) was used to calculate the polygenic risk score in men in the testing datasets.
References

1. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P, Marchini J. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203-209.

2. Abraham G, Qiu Y, Inouye M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. Bioinformatics. 2017;33(17):2776-2778.

3. Day F, Karaderi T, Jones MR, Meun C, He C, Drong A, Kraft P, Lin N, Huang H, Broer L, Magi R, Saxena R, Laisk T, Urbanek M, Hayes MG, Thorleifsson G, Fernandez-Tajes J, Mahajan A, Mullin BH, Stuckey BGA, Spector TD, Wilson SG, Goodarzi MO, Davis L, Obermayer-Pietsch B, Uitterlinden AG, Anttila V, Neale BM, Jarvelin MR, Fauser B, Kowalska I, Visser JA, Andersen M, Ong K, Stener-Victorin E, Ehrmann D, Legro RS, Salumets A, McCarthy MI, Morin-Papunen L, Thorsteinsdottir U, Stefansson K, andMe Research T, Styrkarsdottir U, Perry JRB, Dunaif A, Laven J, Franks S, Lindgren CM, Welt CK. Large-scale genome-wide meta-analysis of polycystic ovary syndrome suggests shared genetic architecture for different diagnosis criteria. PLoS Genet. 2018;14(12):e1007813.

4. Lam M, Chen CY, Li Z, Martin AR, Bryois J, Ma X, Gaspar H, Ikeda M, Benyamin B, Brown BC, Liu R, Zhou W, Guan L, Kamatani Y, Kim SW, Kubo M, Kusumawardhani A, Liu CM, Ma H, Periyasamy S, Takahashi A, Xu Z, Yu H, Zhu F, Schizophrenia Working Group of the Psychiatric Genomics C, Indonesia Schizophrenia C, Genetic Rosn-C, the N, Chen WJ, Faraone S, Glatt SJ, He L,
Hyman SE, Hwu HG, McCarroll SA, Neale BM, Sklar P, Wildenauer DB, Yu X, Zhang D, Mowry BJ, Lee J, Holmans P, Xu S, Sullivan PF, Ripke S, O'Donovan MC, Daly MJ, Qin S, Sham P, Iwata N, Hong KS, Schwab SG, Yue W, Tsuang M, Liu J, Ma X, Kahn RS, Shi Y, Huang H. Comparative genetic architectures of schizophrenia in East Asian and European populations. Nat Genet. 2019;51(12):1670-1678.

5. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904-909.

6. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Y AR, H KF, Schoenherr S, Forer L, McCarthy S, Abecasis GR, Durbin R, A LP. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016;48(11):1443-1448.

7. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81(5):1084-1097.

8. Mitt M, Kals M, Parn K, Gabriel SB, Lander ES, Palotie A, Ripatti S, Morris AP, Metspalu A, Esko T, Magi R, Palta P. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. Eur J Hum Genet. 2017;25(7):869-876.

9. Leitsalu L, Haller T, Esko T, Tammesoo ML, Alavere H, Snieder H, Perola M, Ng PC, Magi R, Milani L, Fischer K, Metspalu A. Cohort Profile: Estonian
Biobank of the Estonian Genome Center, University of Tartu. Int J Epidemiol. 2015;44(4):1137-1147.

10. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience. 2019;8(7).

11. Ge T, Chen CY, Ni Y, Feng YA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun. 2019;10(1):1776.

12. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, Kathiresan S. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50(9):1219-1224.

13. Eastwood SV, Mathur R, Atkinson M, Brophy S, Sudlow C, Flaig R, de Lusignan S, Allen N, Chaturvedi N. Algorithms for the Capture and Adjudication of Prevalent and Incident Diabetes in UK Biobank. PLoS One. 2016;11(9):e0162388.

14. Udler MS, Kim J, von Grotthuss M, Bonas-Guarch S, Cole JB, Chiou J, Christopher DAboboM, the I, Boehnke M, Laakso M, Atzmon G, Glaser B, Mercader JM, Gaulton K, Flannick J, Getz G, Florez JC. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. PLoS Med. 2018;15(9):e1002654.

15. Pirastu N, Joshi PK, de Vries PS, Cornelis MC, McKeigue PM, Keum N, Franceschini N, Colombo M, Giovannucci EL, Spiliopoulou A, Franke L, North KE, Kraft P, Morrison AC, Esko T, Wilson JF. GWAS for male-pattern baldness
identifies 71 susceptibility loci explaining 38% of the risk. Nat Commun. 2017;8(1):1584.

16. Norwood OT. Male pattern baldness: classification and incidence. South Med J. 1975;68(11):1359-1365.