Clinical Research Article

Characterizing the Clinical and Genetic Spectrum of Polycystic Ovary Syndrome in Electronic Health Records

Ky’Era V. Actkins,1,2 Kritika Singh,2,3 Donald Hucks,2,3 Digna R. Velez Edwards,2,4,5 Melinda Aldrich,2,3,6,7 Jeeyeon Cha,8 Melissa Wellons,8 and Lea K. Davis2,3,7

1Department of Microbiology, Immunology, and Physiology, Meharry Medical College, Nashville, Tennessee 37208; 2Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee 37232; 3Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee 37232; 4Vanderbilt Epidemiology Center, Institute of Medicine and Public Health, Vanderbilt University Medical Center, Nashville, Tennessee 37232; 5Division of Quantitative Sciences, Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, Tennessee 37232; 6Department of Thoracic Surgery, Vanderbilt University Medical Center, Nashville, Tennessee 37232; 7Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee 37232; and 8Division of Diabetes, Endocrinology, and Metabolism, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee 37232

ORCID number: 0000-0002-1591-6914 (K. V. Actkins).

Abstract

Context: Polycystic ovary syndrome (PCOS) is one of the leading causes of infertility, yet current diagnostic criteria are ineffective at identifying patients whose symptoms reside outside strict diagnostic criteria. As a result, PCOS is underdiagnosed and its etiology is poorly understood.

Objective: We aim to characterize the phenotypic spectrum of PCOS clinical features within and across racial and ethnic groups.

Methods: We developed a strictly defined PCOS algorithm (PCOS\textsubscript{keyword-strict}) using the International Classification of Diseases, ninth and tenth revisions and keywords mined from clinical notes in electronic health records (EHRs) data. We then systematically relaxed the inclusion criteria to evaluate the change in epidemiological and genetic associations resulting in 3 subsequent algorithms (PCOS\textsubscript{coded-broad}, PCOS\textsubscript{coded-strict}, and PCOS\textsubscript{keyword-broad}). We evaluated the performance of each phenotyping approach and characterized prominent clinical features observed in racially and ethnically diverse PCOS patients.
Polycystic ovary syndrome (PCOS) is an endocrine disorder that is the leading cause of infertility in females. The genetic, environmental, and metabolic variables that contribute to its complex architecture also influence clinical heterogeneity among individuals with PCOS, resulting in a broad spectrum of symptoms. The heterogeneous clinical presentation of PCOS makes it difficult to diagnose, and an estimated 75% of females with PCOS remain undiagnosed (1, 2).

There are 3 commonly used diagnostic criteria for PCOS that each have specific symptom requirements for diagnosis. The first diagnostic definition was put forth by the National Institutes of Health (NIH), which required an ovulatory phenotype (eg, oligomenorrhea) and hyperandrogenism for diagnosis (3, 4). Subsequently, the second diagnostic definition for PCOS was created by the European Society for Human Reproduction and Embryology and the American Society for Reproductive Medicine, commonly referred to as the Rotterdam criteria. The Rotterdam diagnostic criteria were designed to expand the PCOS definition from the strict NIH definition by requiring any 2 of the following 3 symptoms: oligoanovulation, hyperandrogenism, or polycystic ovaries (3, 5). However, the Androgen Excess and PCOS Society argued that hyperandrogenism is the primary driver of PCOS and thus, created a third diagnostic criteria that requires both hyperandrogenism and ovulatory dysfunction (eg, oligoanovulation or polycystic ovaries) (4). PCOS diagnostic criteria were developed to improve and standardize diagnoses; however, patients identified by each approach may have differing symptomologies and only a small fraction of females will meet multiple diagnostic criteria (6, 7).

Implementing a broader PCOS definition could lead to false-positive diagnoses and prevalence overestimations (4, 8). However, restrictive definitions miss females with clinical manifestations outside the requirements (9), including metabolic dysfunction. Females with PCOS and metabolic dysfunction are at higher risk for a range of comorbidities, including insulin resistance, type 2 diabetes, cardiovascular disease, obesity, cancer, and psychiatric disorders (10, 11). Delayed screening and treatment for PCOS could exacerbate these metabolic conditions, particularly in underrepresented populations that have a greater baseline risk of metabolic disorders (12, 13). Studies have shown that African American and Hispanic females with PCOS have an increased risk for metabolic syndrome and other cardiovascular disorders, yet little is known about why comorbidities increase across diverse racial and ethnic groups and whether expanding diagnostic criteria for PCOS may improve health equity (14, 15).

Most prior studies included small numbers of minority females and many have used differing diagnostic criteria (1, 16-18). Without larger studies, it is difficult to accurately characterize the clinical presentation of PCOS and its consequences across populations. Adapting an inclusive phenotyping approach capable of identifying severe to mild PCOS presentations could improve study methods for marginalized communities.

The overarching hypothesis of this study is that clinically diagnosable PCOS represents the extreme tail (ie, clinical manifestation) of a spectrum of hormonal and metabolic dysregulation that has common manifestations in a large number of females. In our study, we developed a simple but stringent automated phenotyping algorithm that could be applied to electronic health records (EHRs) to identify females with confirmed PCOS. Next, we systematically relaxed the inclusion criteria to test the hypothesis that excess androgens, increased metabolic dysfunction, and PCOS genetic liability are present to a lesser degree among females with milder PCOS symptoms and symptoms outside the diagnostic criteria. Finally, we examined these features across diverse racial and ethnic groups within the EHR.

**Materials and Methods**

**Clinical data**

Vanderbilt University Medical Center (VUMC) is a tertiary care hospital in Nashville, Tennessee, with several clinics throughout Tennessee and the surrounding states. Medical records have been electronically documented at VUMC since the early 1990s (19). The VUMC EHR
houses more than 3 million EHRs in a deidentified clinical research database called the Synthetic Derivative (SD). The synthetic derivative includes demographics, International Classification of Diseases, ninth revision (ICD9) and tenth revision (ICD10) codes, procedural codes, clinical notes, medications, and laboratory values. The Vanderbilt University Institutional Review Board (IRB) approved this project (IRB 160 279).

Algorithm development

We identified females with PCOS within the synthetic derivative using a strict automated phenotyping algorithm that yielded a positive predictive value (PPV) of at least 90%. Our secondary analysis relaxed the requirements of our core algorithm to identify a larger sample of patients with a wider array of PCOS symptoms. In total, we developed 4 algorithms, which are illustrated in Fig. 1. The demographics of each resultant data set are reported in Table 1.

Before applying each algorithm, we filtered the entire synthetic derivative population to include only females between the ages of 11 to 44 years to enrich our population for females of reproductive age. We first deployed the PCOS algorithm, which required at least one ICD9 or ICD10 code for polycystic ovaries (ie, 256.4 and E28.2) OR hirsutism (ie, 704.1 and L68.0) AND irregular menses (ie, all 626 codes, all N91 codes, N92.5, and N92.6) (Supplementary Table 1) (20). Hirsutism was used as clinical evidence of hyperandrogenism. Individuals with any exclusion codes (Supplementary Table 1) present in their medical record were removed. Furthermore, the PCOS algorithm required the presence of at least one of the following keywords: “polycystic ovaries,” “PCOS,” or “PCO” in a clinical note. We included inpatient and outpatient note types designated as “outpatient note,” “gynecology clinic visit,” “nursing report,” “endocrine and diabetes clinic visit,” “pediatric endocrinology patient visit,” and “reproductive endocrinology clinic visit.” Next, to broaden the approach, we dropped the keyword requirements and retained only the ICD codes for the PCOS algorithm. We then loosened the requirement of the union of hirsutism and irregular menses codes and instead required the presence of ICD codes for polycystic ovaries OR hirsutism OR irregular menses. In addition to any code for these conditions, we required the presence of a PCOS-related keyword. The resultant algorithm was termed PCOS (Supplementary Table 1) (20). Last, we again dropped the keyword requirement, which resulted in the broadest approach (PCOS), including the presence of any PCOS symptoms (ie, ICD codes for polycystic ovaries OR hirsutism OR irregular menses).

To identify controls, the synthetic derivative was again filtered for females between the ages of 11 to 44 years. The control data set excluded individuals with one or more exclusion ICD codes or any inclusion codes used to define PCOS case criteria (Supplementary Table 1) (20).

Chart review methods for algorithm evaluation

We performed 2 comprehensive chart reviews to evaluate the performance of the phenotyping algorithms. The charts were reviewed by clinical domain experts (M.W. and J.C.) and trained reviewers (K.V.A., D.H., and K.S.). We first evaluated the PPV for each algorithm. We randomly selected 50 charts from each set of algorithm-defined cases (200 in total) for manual chart review by trained reviewers and clinical domain experts (21). We then calculated the proportion of true positive cases based on manual chart review out of the total algorithm-identified cases, to define the PPV for each algorithm (Table 2).

Next, we evaluated the sensitivity and specificity of the original PCOS (Supplementary Table 1) algorithm. For this review, we first identified a “medical home” population (defined by the presence of at least 5 ICD codes on separate days over at least a 3-year period [N = 752436]). Then, we filtered for females who met a data floor requirement (ie, at least one ICD10 code for PCOS, E28.2 [N = 9507]). From this sample, we then selected a new set of 200 charts at random and performed a manual chart review of all available information in the EHRs (ie, codes and clinical notes) to classify individuals into “PCOS cases” and “not PCOS cases” based on clinical gold standards for diagnosis. This manual review clinical determination was then compared to the PCOS (Supplementary Table 1) algorithm-defined determination to calculate sensitivity and specificity (Table 3).

Laboratory data

Weight and height measurements are routinely recorded during clinical visits and were extracted from the EHRs to calculate body mass index (BMI). Laboratory measurements were extracted and cleaned using QualityLab (22). In brief, 3075 quantitative laboratory tests were extracted from EHRs of 70,704 patients in the synthetic derivative. Laboratory values with nonnumeric values (n = 2865) and those evaluated on only one patient (n = 467) were removed. The remaining laboratory values were cleaned by filtering out duplicate entries and measurements with multiple units. Insulin, estradiol, and free testosterone hormone data were then extracted for further analysis. Median laboratory values were calculated for individuals with multiple clinical measurements. Median laboratory
values of BMI, insulin, estradiol, and free testosterone were used to characterize the clinical features of each algorithm dataset (Table 4 and Supplementary Table 2) (20). For each laboratory value, extreme outliers (ie, first and 99th percentile) were removed. Additionally, laboratory values were regressed on age at the laboratory and an inverse normal transformation was applied to the residual to result in an age-adjusted normalized laboratory value for subsequent genetic analyses (Supplementary Tables 4 and 5) (20). Sample sizes differed for subsequent analyses...
based on available clinical measurements (Supplementary Fig. 1) (20).

Genetic data
We included 40 802 females of European descent and 9418 females of African descent who were genotyped on the Illumina MEGAEX Array (23). There were 1 832 777 genotyped variants pre–quality control and preimputation. Full details of the quality control pipeline have been described in Dennis et al (22). In summary, we removed single-nucleotide variations (SNVs, formerly single-nucleotide polymorphisms [SNPs]) with a call rate of less than 0.98, individuals with a call rate of less than 0.98, and individuals with a discrepancy between chromosomal and reported sex. FlashPCA2 was used for principal component analysis (24, 25). Related samples with a pi-hat greater than 0.2 (approximate first-cousin relationships) were removed. The Michigan Imputation server was used to impute missing genotypes based on the Haplotype Reference Consortium panel and imputation accuracy was set at $R^2$ greater than 0.03 (26). For the European genotyped sample, 36 604 696 variants were included in the final data set and 36 513 773 variants were included in the African genotyped sample.
Table 3. Sensitivity and specificity for polycystic ovary syndrome algorithm data sets. A 2 × 2 contingency table showing the positive predictive value, negative predictive value, sensitivity, and specificity of the keyword strict (KS) algorithm. The KS algorithm was compared to the gold-standard definition. Chart review was conducted on 200 random records of women who met our data floor and medical home definition.

| Keyword strict algorithm | Case | Not case | Results |
|--------------------------|------|----------|---------|
| Gold Standard            |     |          |         |
| Case                     | 96   | 97       | 50% Sensitivity |
| Not case                 | 5    | 2        | 29% Specificity |
| Results                  | 95% PPV | 2% NPV  |         |

Abbreviations: NPV, negative predictive value; PPV, positive predictive value.

Construction of polycystic ovary syndrome polygenic risk scores

Polygenic risk scores (PRS) for PCOS were calculated for each individual using the PRS-CS software package, which applies a Bayesian continuous shrinkage model to SNV β estimates to adjust for linkage disequilibrium (27). With this framework, no P value thresholds are required to generate PRS. β estimates were drawn from the summary statistics of the largest meta-genome-wide association study (GWAS) of PCOS performed by Day and colleagues that included 5209 cases and 32 055 controls of European descent diagnosed using NIH Rotterdam criteria (28). To maximize our ability to assess the full polygenicity of PCOS, all SNVs (nEuropean = 783 989, nAfrican = 772 724) from the summary statistics data set were used to generate PCOSprs. PCOSprs were calculated for a total of 40 802 females of European descent and 9418 females of African descent in BioVU.

Statistical analysis

The prevalence of each algorithm phenotype was calculated using the proportion of cases identified by each algorithm out of the total number of females in the synthetic derivative between ages 11 to 44 years (see Table 2). The PPV was determined by calculating the proportion of “reviewer-defined” true positives out of the “algorithm-defined” positives (see Table 2). Sensitivity was defined as the number of “algorithm-defined” positives out of the total positives (ie, “algorithm-defined” positives plus “reviewer-defined” positives). Specificity was defined as the number of “algorithm-defined” negatives out of the total negatives (ie, “algorithm-defined” and “reviewer-defined” negatives) (Table 3).

We performed Wilcoxon rank sum tests to determine if the median BMI and laboratory measurements (eg, free testosterone, insulin, and estradiol) differed between the PCOS algorithm-identified cases and controls. We then tested the difference between cases defined by each algorithm and between cases and controls stratified by EHR-reported race. Finally, we evaluated differences between African American, White, and Hispanic females identified as cases by each algorithm (Table 4, Supplementary Table 2, Figs. 2-4, Supplementary Figs. 2 and 3) (20).

We performed logistic and linear regression analyses to test the association between PCOSprs and case/control designation and, separately, the association between PCOSprs and available hormone data (ie, estradiol, free testosterone, insulin). The analysis was performed in 2 populations based on genetic ancestry: European descent (n = 8925) and African descent (n = 3381). Analyses were filtered to individuals who met our medical home definition. Covariates included median age (calculated across the medical record for each individual) and the top 10 genetic principal components to adjust for genetic ancestry when fitting the case/control logistic regression model. A separate sensitivity analysis adjusting BMI was also performed (Supplementary Table 3) (20). For regression analysis of hormone data, covariates included the top 10 genetic principal components. BMI was included as a covariate in the sensitivity analysis. We then evaluated the model and variance explained by PCOSprs using the Nagelkerke pseudo-R² for the logistic regressions and R² for the linear regressions (Table 5, Supplementary Tables 3-5) (28). All analyses were performed in R 3.6.0.

We included 2 corrections for multiple testing. The first corrected threshold reflects adjustment for all independent tests (n = 30, P < .002, indicated by a in tables). The second corrected threshold is more stringent and reflects adjustment for all tests (n = 120, P < 4 × 10⁻³, indicated by b in tables) regardless of whether they were independent or performed on nested subsets of the data (see Fig. 1).

Results

Sample characteristics

PCOS data set demographics are reported in Table 1. Potential PCOS cases were selected from 704 970 females between ages 11 to 44 years in the synthetic derivative (see Fig. 1). Among them, 399 405 females had an EHR-reported race of White, 80 892 females reported as African American, and 30 445 females reported as Hispanic. The remaining females had a different EHR-reported race, with the largest majority being unknown. Although the sample sizes differed between the algorithms, each algorithm identified a similar proportion of positive cases across racial/ethnic groups. The most stringent
algorithm, PCOS\textsubscript{keyword-strict}, identified 4593 potential PCOS cases including 3143 White, 658 African American, and 204 Hispanic females (see Table \text{2}). As the algorithm requirements were relaxed, the number of individuals meeting symptom criteria increased. PCOS\textsubscript{keyword-broad} identified 6193 individuals including 4207 White, 899 African American, and 274 Hispanic females. The PCOS\textsubscript{coded-strict} approach identified 8340 individuals including 5526 White, 1217 African American, and 368 Hispanic females. At the broadest extreme, the PCOS\textsubscript{coded-broad} \((n = 41205)\) algorithm captured the largest and most diverse sample, with 26317 White, 7648 African American, and 2052 Hispanic females. The individuals identified by the keyword algorithms had the highest average medical record length at 4.69 years \((SD = 4.22)\) for PCOS\textsubscript{keyword-broad} and 4.66 years \((SD = 4.13)\) for PCOS\textsubscript{keyword-strict}. This was expected because the increased length of the record also increased the probability of PCOS documentation in the medical record. The PCOS\textsubscript{coded-strict} and PCOS\textsubscript{keyword-strict} algorithms captured the eldest set of patients who were in their mid-20s with a median age of 25.18 (SD = 7.63) and 25.09 (SD = 7.39) years, respectively. Comparatively, the control data set had the youngest sample, aged 23.85 \((SD = 9.47)\) years with the shortest EHR length at 3.92 \((SD = 3.9)\). As expected, the control sample had the

Table 4. Laboratory measurements for polycystic ovary syndrome algorithm cases and controls. Wilcoxon rank sum test was performed between cases and controls. The 25th and 75th percentile was reported for each algorithm

| Algorithm          | Type  | No.   | Median | 25th Percentile | 75th Percentile | \(P\)       |
|--------------------|-------|-------|--------|-----------------|-----------------|------------|
|                   |       |       |        |                 |                 | \(p < .002\) \(p < 4 \times 10^{-3}\) |
| Body mass index    |       |       |        |                 |                 |            |
| Coded broad        | Cases | 34603 | 26.54  | 22.21           | 33.50           | < 2.22E-16 |
|                    | Controls | 29140 | 24.03  | 20.68           | 29.30           |            |
| Coded strict       | Cases | 7555  | 33.605 | 27.28           | 40.36           | < 2.22E-16 |
|                    | Controls | 29140 | 24.03  | 20.68           | 29.30           |            |
| Keyword broad      | Cases | 5844  | 32.805 | 26.29           | 39.70           | < 2.22E-16 |
|                    | Controls | 29140 | 24.03  | 20.68           | 29.30           |            |
| Keyword strict     | Cases | 4358  | 34.08  | 27.63           | 40.76           | < 2.22E-16 |
|                    | Controls | 29140 | 24.03  | 20.68           | 29.30           |            |

In Table 4, we present the laboratory measurements for polycystic ovary syndrome algorithm cases and controls. Wilcoxon rank sum test was performed between cases and controls. The 25th and 75th percentile was reported for each algorithm.
lowest average clinical measurements for BMI, insulin, estradiol, and free testosterone (see Table 1).

Evaluation of algorithm performance
We found little difference between the PPVs calculated for the PCOS\textsubscript{keyword-strict} algorithm using 2 distinct chart review methods. PCOS\textsubscript{keyword-strict} yielded a PPV of 96% using our first method of chart review (see Table 2) and a PPV of 95% when it was tested against the clinician gold standard definition in our second manual review method (see Table 3). Despite this, the PCOS\textsubscript{keyword-strict} algorithm had a sensitivity of only 50% and specificity of 29% when evaluated using the gold-standard definition (see Table 3).

When compared to our broader algorithms, the performance of the PCOS\textsubscript{keyword-strict} algorithm was most similar to the PCOS\textsubscript{coded-strict} algorithm, which had a slightly higher PPV of 98% (see Table 2). The text mining criteria did not
significantly improve the accuracy of PCOS_{keyword-strict} over PCOS_{coded-strict}. The accuracy of the "broad" criteria decreased compared to the "strict" algorithms. Our broadest algorithm, PCOS_{coded-broad}, had the lowest performance, with a PPV of 30% (see Table 2). This is likely due to the large number of patients who had only an irregular menstruation ICD code with no other diagnosed PCOS symptoms. Interestingly, the addition of the text-mined PCOS keyword significantly improved the PPV of the broad criteria algorithm from 30% (PCOS_{coded-broad}) to 82% (PCOS_{keyword-broad}). Unlike the extreme broad criteria, the other algorithm phenotypes were similarly prevalent (18, 29, 30). This was also true for the race-stratified cohorts, for which there were little to no differences in prevalence across algorithm phenotypes.

Clinical feature characterization

We characterized the clinical profiles of cases and controls identified by each algorithm by calculating median BMI, insulin, estradiol, and free testosterone levels...
(see Table 4). Each algorithm identified set was significantly heavier compared to controls ($P < .05$). For all algorithms, African American and Hispanic females had higher baseline metabolic profiles for BMI, insulin, estradiol, and free testosterone measurements compared to White females. Moreover, the median BMI in the algorithm data sets increased with increasing algorithm stringency (Fig. 2A). African American females meeting the algorithm criteria consistently demonstrated the highest median BMI in the PCOS\textsubscript{coded-broad}, PCOS\textsubscript{coded-strict}, and PCOS\textsubscript{keyword-broad} algorithms (Supplementary Table 2) (20). The BMI of all 3 stratified race groups differed significantly from each other with the exception of White and Hispanic females in the strictest keyword algorithms (Fig. 3A). Within the Hispanic-defined group, the median values did not vary significantly between the PCOS\textsubscript{coded-strict}, PCOS\textsubscript{keyword-broad}, and PCOS\textsubscript{keyword-strict} groups (Supplementary Fig. 3A) (20).

Overall, females identified by each of the 4 PCOS algorithms demonstrated significantly higher insulin and free
testosterone levels compared to controls ($P < .05$) (Fig. 2B and 2D, Table 4). The median hormone values also increased with algorithm stringency, though the trend was not as pronounced as it was for the BMI measurements. BMI increased with algorithm stringency within each race stratified analysis. Insulin values increased with algorithm stringency for White and Hispanic females, and free testosterone values increased with algorithm stringency for White and African American females (Fig. 4, Supplementary Table 2) (20).

Measurements of free testosterone were significantly higher among females identified by the PCOS coded-strict, PCOS keyword-broad, and PCOS keyword-strict algorithms compared to the controls and the PCOS coded-broad algorithm (Supplementary Fig. 2D) (20). The insulin profile of African Americans was similar across females identified by each algorithm and also did not significantly differ from controls (Fig. 4B). However, African American females identified by the PCOS coded-broad and PCOS keyword-broad algorithms did exhibit higher estradiol levels than African American controls (Supplementary Table 2, Fig. 4C) (20). Additionally, African American females identified by the PCOS keyword-broad algorithm demonstrated significantly higher estradiol levels than White females identified by the same approach (Fig. 3C). Outside of these findings, there were no other significant differences between races for hormone values (see Fig. 3).

### Genetic validation of polycystic ovary syndrome case status

Increasing algorithm requirement stringency was also reflected in the performance of the PCOS$_{PRS}$, Odds ratios (ORs), which are reported per synthetic derivative of the PCOS$_{PRS}$, were higher among cases identified by algorithms using strict PCOS requirements and decreased as those requirements were relaxed. This trend was observed both for the European and African genotyped samples (Fig. 5, Table 5). BMI was tested as a covariate in the model and did not impact the results. (Supplementary Table 3) (20).

Among females identified as cases and controls in the European descent PCOS keyword-broad algorithm, we observed a significant association between PCOS$_{PRS}$ and case/control status (see Fig. 5 and Table 5). Cases identified by the PCOS keyword-broad algorithm yielded the highest OR of 1.28 (95% CI, 1.14-1.45, $P = 6.97$E-05) and a PRS pseudo-$R^2$ of 0.8%. Comparatively, the PCOS coded-strict algorithm yielded the second highest OR (OR = 1.40, 95% CI, 1.11-1.78, $P = .01$) with a PRS pseudo-$R^2$ of 0.5%.

Although none of the models for the African-descent genotyped samples passed our multiple correction thresholds, the PCOS coded-strict, PCOS keyword-broad, and the PCOS keyword-strict algorithms passed a false discovery rate of 0.05 and demonstrated higher point estimates compared to the European sample. The African descent PCOS keyword-strict algorithm yielded the highest OR (OR = 1.49, 95% CI, 1.11-2.01, $P = .01$) and had the greatest variance explained in case/control status by PCOS$_{PRS}$ model (PRS pseudo-$R^2$ = 1.1%). This was followed by the PCOS coded-strict algorithm with an OR of 1.40 (95% CI, 1.11-1.78, $P = .01$) and then, the PCOS keyword-broad algorithm with an OR of 1.34 (95% CI, 1.03-1.75, $P = .03$).

### Table 5. Genetic validation of polycystic ovary syndrome (PCOS) diagnosis. Logistic regressions were performed between PCOS case status and PCOS polygenic risk scores (PCOS$_{PRS}$). Regressions were adjusted for the first 10 principal components, and each algorithm data set was case matched for age and race.

| Algorithm          | Cases | Controls | Estimate | SE  | $P$   | OR   | LCI  | UCI  | PRS $R^2$ |
|--------------------|-------|----------|----------|-----|-------|------|------|------|------------|
| **European**       |       |          |          |     |       |      |      |      |            |
| Coded broad        | 1877  | 5052     | 0.06     | 0.03| .06   | 1.06 | 1.00 | 1.13 | 7.00E-04  |
| Coded strict       | 366   | 5052     | 0.17     | 0.06| 0.01  | 1.18 | 1.05 | 1.33 | 0.004      |
| Keyword broad      | 332   | 5052     | 0.25     | 0.06| 6.97E-05 | 1.28 | 1.14 | 1.45 | 0.008$^a$ |
| Keyword strict     | 245   | 5052     | 0.20     | 0.07| 0.01  | 1.22 | 1.06 | 1.41 | 0.005      |
| **African**        |       |          |          |     |       |      |      |      |            |
| Coded broad        | 908   | 1663     | 0.11     | 0.06| .07   | 1.12 | 0.99 | 1.26 | 0.002      |
| Coded strict       | 161   | 1663     | 0.34     | 0.12| 0.01  | 1.40 | 1.11 | 1.78 | 0.01       |
| Keyword broad      | 129   | 1663     | 0.30     | 0.13| 0.03  | 1.34 | 1.03 | 1.75 | 0.007      |
| Keyword strict     | 98    | 1663     | 0.40     | 0.15| 0.01  | 1.49 | 1.11 | 2.01 | 0.011      |

---

**Abbreviations:** LCI, lower confidence interval; OR, odds ratio; PRS, polygenic risk score; $R^2$, pseudo-$R^2$; UCI, upper CI.

$^aP < .002$. 

Association between polycystic ovary syndrome polygenic risk scores and hormone measurements

There were no significant associations between PCOS$_{PRS}$ and the hormone measurements (insulin, estradiol, and T
free testosterone) evaluated because of the large CIs. However, PCOS PRS explained 11% of the phenotypic variance for estradiol in the African-descent PCOS coded-strict cases and controls (Supplementary Table 4) (20). The highest $R^2$ observed among European-descent females was in estradiol levels measured in the PCOS coded-strict algorithm sample.

**Discussion**

Expanding study inclusion criteria has the potential to increase and diversify future study samples, which may significantly benefit underrepresented high-risk groups (29). This study describes an automated PCOS phenotyping method that captures varying degrees of PCOS symptoms, parallel dysregulated hormone levels, and increased genetic risk for the PCOS coded-strict, PCOS keyword-broad, and PCOS keyword-strict algorithms. Our results also further support the hypothesis that females with PCOS symptoms have increased metabolic dysregulation compared to controls when measured on BMI, insulin, and free testosterone values dependent on clinical definitions and their EHR-reported race. Finally, our study demonstrates the importance of including African-descent genotyped samples, in addition to European-descent populations, for genetic analysis.

Our 2 “strict definition” algorithms identified PCOS cases with high accuracy (PPV > 95%), and the PCOS keyword-strict algorithm performed just as well when compared to the gold-standard data set (PPV = 95%). Although the sensitivity and the specificity of the PCOS keyword-strict algorithm was low, the high PPV displays its high accuracy within an EHR setting, suggesting that these algorithms are not likely to be useful for clinical purposes but are useful for identifying PCOS study populations. Compared to the previously reported ICD-based PCOS algorithm, which had a PPV of 66% and 65% for the “broad” and “refined” algorithms, our 2 best performing algorithms improved case identification by more than 30% (9). The Castro et al (9) study and
our study both show the feasibility of using ICD-based algorithms to identify PCOS cases and the advantages of using EHR systems. Consistent with the previous study, we also found that the inclusion of text mining increased the proportion of true PCOS cases (PPV = 98%) but resulted in a smaller sample size. Removal of the text mining requirement only slightly affected the PPV (96%), indicating that studies performed in EHR systems with limited access to clinical notes and ultrasound images may still be able to identify females with PCOS with high accuracy. Conversely, we found that the addition of text mining to the broad algorithms significantly improved the accuracy from 30% to 82%, indicating that clinical suspicions of PCOS documented in medical records may also be used to identify females with PCOS even when few symptoms are coded.

The clinical and epidemiological features of our PCOS cohorts are consistent with previous studies that show females with PCOS have higher metabolic clinical characteristics than controls (31-33). We also observed no significant difference in the prevalence of PCOS definitions between minority and White individuals, nor were there any large differences between races for insulin, estradiol, and free testosterone clinical measurements outside of higher estradiol levels in African Americans in the PCOS\textsubscript{coded-strict} algorithm (34). This finding could illustrate an increase in aromatization caused by higher circulating androgen levels or may simply be a reflection of increased ovulatory activity observed among African American females (33, 35, 36). However, this proves difficult to definitively conclude because of the small sample size, the large variability of the data in the cases and controls, and the timing at which estradiol was measured, which could have been at any point during the individual’s menstrual cycle. It is also important to note that African American females had significantly higher BMI measurements than Hispanic and White females across both the control and PCOS definition data sets. African Americans are disproportionately affected by conditions such as obesity and type 2 diabetes, which could possibly be exacerbated by the onset of PCOS (37, 38). These conditions could also be influenced by the patient structure of the EHR database, which has a sicker population because of the tertiary care status of VUMC and its geographic location, where obesity rates are generally higher.

The spectrum of PCOS characteristics remain the root of its heavily debated diagnostic criteria, and PRS provided a unique way to validate the phenotypes identified by our algorithms outside the traditional clinical symptomology. PCOS has a complex genetic architecture evidenced by its high heritability (~70%) and its underlying polygenic background (39, 40). PCOS\textsubscript{PRS} exploits this by assessing the cumulative SNV-based risk across different clinical definitions and ancestries. Although it is ideal to calculate PRS using summary statistics obtained from population references that accurately portray our respective target group, to date there are no PCOS GWAS in African-descent populations. Thus, we were limited to using summary statistics from a European GWAS to build PRS both for our European- and African-descent samples, even though it is known that European-based PRS do not perform as robustly in non-European samples (41). Despite this, previous studies have shown that PCOS\textsubscript{PRS} models can detect cross-ancestry genetic risk in African and multiancestry cohorts using European-based PRS models (42). In our study, the PCOS\textsubscript{PRS} model explained little variance in case/control status (PRS pseudo-$R^2 \leq 1\%$), but the PCOS\textsubscript{PRS} was significantly higher among cases in the PCOS\textsubscript{coded-strict}, PCOS\textsubscript{keyword-broad}, and the PCOS\textsubscript{keyword-strict} algorithms compared to controls.

The limitations of our study include lower statistical power in the genetic analysis of African-descent cases and controls. This limitation is exacerbated by the underrepresentation of non-European populations in the PCOS GWAS, but also reflects the low representation of non-European samples in the VUMC EHRs. Eurocentric GWAS remain a significant limitation to interpretability of PRS in non-European samples because it is unrepresentative of the genetic diversity that make up those populations. Therefore, there is a high need for studies to increase the genetic data gathered on diverse ancestral populations to advance PCOS genomics in EHRs (41, 43).

In conclusion, we applied an automated phenotyping approach to examine variation in the symptomology of PCOS across racial and ethnic groups using both strict and broad classification criteria. Debate over current implemented PCOS diagnostic criteria, in part, reflects the complexity of PCOS symptoms, and EHRs provide significant advantages. First, research in the longitudinal EHR allows us to maximize the number of PCOS clinical manifestations observed over time (16, 30, 44). Second, EHRs contain a variety of clinical variables (laboratory measurements, imaging records, etc) that can be used to study the risk factors and health outcomes of PCOS patients. We created a simple EHR-based PCOS algorithm that identifies patients using ICD codes and text mining (PCOS\textsubscript{keyword-strict}) methods, and by relaxing the inclusion requirements we identified patients across the PCOS syndromic spectrum. This study approach provides a unique view of the shared clinical and genetic architecture that influences both PCOS as a discrete diagnosis and the symptoms of PCOS that affect a large number of females.
Acknowledgments

Financial Support: This work was supported by the National Institutes of Health (NIH Training Grants ST32GM007628-42 and ST32DK007061 to K.V.A. and J.C.), L.K.D. is supported by National Institutes of Health U54MD010722. This research was conducted in part using the resources from the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, Tennessee. Data sets for this project were obtained using the Synthetic Derivative at Vanderbilt University Medical Center, which is supported by multiple institutional, private, and federal grant sources. This includes the National Institutes of Health-funded Shared Instrumentation Grant S10RR025141 and Clinical and Translational Science Awards (CTSA) Grants UL1TR002243, UL1TR000445, and UL1RR024975. Genomic data are also supported by investigator-led projects that include U01HG004798, R01HG006378, U19HL065962, R01HD074711; and additional funding sources listed at https://victr.vumc.org/biovu-funding/.

Author Contributions: K.V.A. and L.K.D. developed the project concept, study design, and drafted the manuscript. K.V.A. performed the data analysis and interpretation and revised the manuscript. All authors contributed to the conception, study design, and drafting of the manuscript. K.V.A. and J.C. performed the chart review. All authors contributed to the interpretation of the data and the revision of the manuscript.

Additional Information

Correspondence and Reprint Requests: Lea K. Davis, PhD, Vanderbilt Genetics Institute, Division of Genetic Medicine, Department of Medicine, 511-A Light Hall, Vanderbilt University, 2215 Garland Ave, Nashville, TN 37232, USA. E-mail: lea.k.davis@gmail.com.

Disclosure Summary: The authors have nothing to disclose.

Data Availability: Restrictions apply to the availability of data generated or analyzed during this study to preserve patient confidentiality or because they were used under license. The corresponding author will on request detail the restrictions and any conditions under which access to some data may be provided.

References

1. Wolf WM, Wattick RA, Kinkade ON, Olfert MD. Geographical prevalence of polycystic ovary syndrome as determined by region and race/ethnicity. Int J Environ Res Public Health. 2018;15(11):2589.
2. Teede H, Gibson-Helm M, Norman RJ, Boyle J. Polycystic ovary syndrome: perceptions and attitudes of women and primary health care physicians on features of PCOS and renaming the syndrome. J Clin Endocrinol Metab. 2014;99(1):E107-E111.
3. Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. Fertil Steril. 2004;81(1):19-25.
4. Azizz R, Carmina E, Dewailly D, et al; Task Force on the Phenotype of the Polycystic Ovary Syndrome of The Androgen Excess and PCOS Society. The Androgen Excess and PCOS Society criteria for the polycystic ovary syndrome: the complete task force report. Fertil Steril. 2009;91(2):456-488.
5. Indran IR, Huang Z, Khin LW, Chan JKY, Viardot-Foucault V, Yong EL. Simplified 4-item criteria for polycystic ovary syndrome: a bridge too far? Clin Endocrinol (Oxf). 2018;89(2):202-211.
6. Stepno NK, Cassar S, Joham AE, et al. Women with polycystic ovary syndrome have intrinsic insulin resistance on euglycaemic-hyperinsulinaemic clamp. Hum Reprod. 2013;28(3):777-784.
7. Hickey M, Doherty DA, Atkinson H, et al. Clinical, ultrasound and biochemical features of polycystic ovary syndrome in adolescents: implications for diagnosis. Hum Reprod. 2011;26(6):1469-1477.
8. Copp T, Jansen J, Doust J, Mol BW, Dokras A, McCaffrey K. Are expanding disease definitions unnecessarily labelling women with polycystic ovary syndrome? BMJ. 2017;358:j3694.
9. Castro V, Shen Y, Yu S, et al. Identification of subjects with polycystic ovary syndrome using electronic health records. Reprod Biol Endocrinol. 2015;13(1):116.
10. Jacewicz-Święccka M, Kowalska I. Polycystic ovary syndrome and the risk of cardiometabolic complications in longitudinal studies. Diabetes Metab Res Rev. 2018;34(8):e3054.
11. Baldani DP, Skrgatic L, Ougoug R. Polycystic ovary syndrome: important underrecognised cardiometabolic risk factor in reproductive-age women. Int J Endocrinol. 2015;2015:786362.
12. Chan JL, Kar S, Vanky E, et al. Racial and ethnic differences in the prevalence of metabolic syndrome and its components of metabolic syndrome in women with polycystic ovary syndrome: a regional cross-sectional study. Am J Obstet Gynecol. 2017;217(2):189.e1-189.e8.
13. Engmann L, Jin S, Sun F, et al; Reproductive Medicine Network. Racial and ethnic differences in the polycystic ovary syndrome metabolic phenotype. Am J Obstet Gynecol. 2017;216(5):493.e1-493.e13.
14. Meyer ML, Sotres-Alvarez D, Steiner AZ, et al. Polycystic ovary syndrome signs and metabolic syndrome in premenopausal Hispanic/Latina women: the HCHS/SOL study. J Clin Endocrinol Metab. 2020;105(3):e447-e456.
15. Hillman JK, Johnson LN, Limaye M, Feldman RA, Sammel M, Dokras A. Black women with polycystic ovary syndrome (PCOS) have increased risk for metabolic syndrome and cardiovascular disease compared with white women with PCOS [corrected]. Fertil Steril. 2014;101(2):530-535.
16. March WA, Moore VM, Willson KJ, Phillips DI, Norman RJ, Davies MJ. The prevalence of polycystic ovary syndrome in a community sample assessed under contrasting diagnostic criteria. Hum Reprod. 2010;25(2):544-551.
17. Ding T, Hardiman PJ, Petersen I, Wang FF, Qu F, Baio G. The prevalence of polycystic ovary syndrome in reproductive-aged women of different ethnicity: a systematic review and meta-analysis. Oncotarget. 2017;8(56):96351-96358.
18. Bozdağ G, Mumusoglu S, Zengin D, Karabulut E, Yildiz BO. The prevalence and phenotypic features of polycystic ovary syndrome: a systematic review and meta-analysis. Hum Reprod. 2016;31(12):2841-2855.
19. Robinson JR, Wei WQ, Roden DM, Denny JC. Defining phenotypes from clinical data to drive genomic research. Annu Rev Biomed Data Sci. 2018;1:69-92.
20. Aktkis KV, Singh K, Hucks D, et al. Data from: Characterizing the clinical and genetic spectrum of polycystic ovary syndrome...
in electronic health records. figshare. Deposited August 18, 2020. doi:10.6084/m9.figshare.12821810.

21. Barnado A, Carroll RJ, Casey C, Wheless L, Denny JC, Crofford LJ. Phenome-wide association study identifies dsDNA as a driver of major organ involvement in systemic lupus erythematosus. *Lupus*. 2019;28(1):66-76.

22. Dennis JK, Sealock JM, Straub P, et al. Lab-wide association scan of polygenic scores identifies biomarkers of complex disease. [Published online ahead of print January 28, 2020.](https://doi.org/10.1101/2020.01.24.20018713)

23. Bien SA, Wojcik GL, Zubair N, et al; PAGE Study. Strategies for enriching variant coverage in candidate disease loci on a multiethnic genotyping array. *PloS One*. 2016;11(12):e0167758.

24. 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.

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

26. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48(10):1284-1287.

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

28. Day F, Karaderi T, Jones MR, et al; 23andMe Research Team. Large-scale genome-wide meta-analysis of polycystic ovary syndrome suggests shared genetic architecture for different diagnosis criteria. *PloS Genet*. 2018;14(12):e1007813.

29. Neven ACH, Laven J, Teede HJ, Boyle JA. A summary on polycystic ovary syndrome: diagnostic criteria, prevalence, clinical manifestations, and management according to the latest international guidelines. *Semin Reprod Med*. 2018;36(1):5-12.

30. Skiba MA, Islam RM, Bell R, Davis SR. Understanding variation in prevalence estimates of polycystic ovary syndrome: a systematic review and meta-analysis. *Hum Reprod Update*. 2018;24(6):694-709.

31. Lerchbaum E, Schwetz V, Rabe T, Giuliani A, Obermayer-Pietsch B. Hyperandrogenemia in polycystic ovary syndrome: exploration of the role of free testosterone and androstenedione in metabolic phenotype. *PloS One*. 2014;9(10):e108263.

32. Antonio L, Pauwels S, Laurent MR, et al. Free testosterone reflects metabolic as well as ovarian disturbances in subfertile oligomenorrheic women. *Int J Endocrinol*. 2018;2018:7956951.

33. Diamanti-Kandarakis E, Dunaf A. Insulin resistance and the polycystic ovary syndrome revisited: an update on mechanisms and implications. *Endocr Rev*. 2012;33(6):981-1030.

34. Ladson G, Dodson WC, Sweet SD, et al. Racial influence on the polycystic ovary syndrome phenotype: a black and white case-control study. *Fertil Steril*. 2011;96(1):224-229.e2.

35. Dumitrescu R, Mehedintu C, Briceag I, Purcarea VL, Hudita D. The polycystic ovary syndrome: an update on metabolic and hormonal mechanisms. *J Med Life*. 2015;8(2):142-145.

36. Marsh EE, Shaw ND, Klingman KM, et al. Estrogen levels are higher across the menstrual cycle in African-American women compared with Caucasian women. *J Clin Endocrinol Metab*. 2011;96(10):3199-3206.

37. Huddleston HG, Cedars MI, Sohn SH, Giudice LC, Fujimoto VY. Racial and ethnic disparities in reproductive endocrinology and infertility. *Am J Obstet Gynecol*. 2010;202(5):413-419.

38. Butts SF, Seifer DB. Racial and ethnic differences in reproductive potential across the life cycle. *Fertil Steril*. 2010;93(3):681-690.

39. Vink JM, Sadrzadeh S, Lambalk CB, Boomsma DI. Heritability of polycystic ovary syndrome in a Dutch twin-family study. *J Clin Endocrinol Metab*. 2006;91(6):2100-2104.

40. Gomes LG, Enéas De Carvalho A, Crespo RP, Bachega TASS, Mendonça BB, Gomes LG. An update of genetic basis of PCOS pathogenesis. *Arch Endocrinol Metab*. 2018;62(3):352-361.

41. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*. 2019;51(4):584-591.

42. Joo YY, Actkins K, Pacheco JA, et al. A polygenic and phenotypic risk prediction for polycystic ovary syndrome evaluated by phenotype-wide association studies. *J Clin Endocrinol Metab*. 2020;105(6):1918-1936.

43. Wojcik GL, Graff M, Nishimura KK, et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature*. 2019;570(7762):514-518.

44. Wang S, Alvero R. Racial and ethnic differences in physiology and clinical symptoms of polycystic ovary syndrome. *Semin Reprod Med*. 2013;31(5):365-369.