Characterizing the clinical and genetic spectrum of polycystic ovary syndrome in electronic health records

Short title: EHR-based analysis of PCOS

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

[1] Department of Microbiology, Immunology, and Physiology, Meharry Medical College, Nashville, TN

[2] Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN

[3] Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN

[4] Department of Thoracic Surgery, Vanderbilt University Medical Center, Nashville, TN

[5] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

[6] Vanderbilt Epidemiology Center, Institute of Medicine and Public Health, Vanderbilt University Medical Center, Nashville, TN

[7]Division of Quantitative Sciences, Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, TN

[8] Division of Diabetes, Endocrinology, and Metabolism, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN

Corresponding author

Lea Davis
Assistant Professor of Medicine
Vanderbilt Genetics Institute
Division of Genetic Medicine
Department of Medicine
511-A Light Hall
Vanderbilt University
2215 Garland Ave
Nashville, TN 37232
Office Tel: 615-875-9478

Keywords: polycystic ovary syndrome, phenotyping, hormones, electronic health record, polygenic risk scores

Acknowledgements

Funding/Support: KVA and JC are supported by NIH training grant 5T32GM007628-42 and 5T32DK007061. LKD is supported by U54MD010722.

This research was done in part using the resources from the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN. Datasets for this project were obtained using the Synthetic Derivative at Vanderbilt University Medical Center which is supported by multiple grant sources that are institutional, private, and federal. This includes the NIH funded Shared Instrumentation Grant S10RR025141 and CTSA grants UL1TR002243, UL1TR000445, and UL1RR024975.

Authors contributions: KVA and LKD drafted the manuscript. KVA, DH, and KS performed chart review. JC and MW supervised and reviewed chart review. All authors contributed to the interpretation of the data and the revision of the manuscript.

Disclosures

The authors have nothing to disclose.

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 under diagnosed 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 ($PCOS_{regex\text{-}strict}$) using International Classification of Diseases, 9th and 10th edition (ICD9/10) and regular expressions 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 three subsequent algorithms ($PCOS_{coded\text{-}broad}$, $PCOS_{coded\text{-}strict}$,$PCOS_{regex\text{-}broad}$). We evaluated the performance of each phenotyping approach and characterized prominent clinical features observed in racially and ethnically diverse PCOS patients.

Results: The best performing algorithm was our $PCOS_{coded\text{-}strict}$ algorithm with a positive predictive value (PPV) of 98%. Individuals classified as cases by this algorithm had significantly higher body mass index (BMI), insulin levels, free testosterone values, and genetic risk scores for PCOS, compared to controls. Median BMI was higher in African American women with PCOS compared to White and Hispanic women with PCOS.

Conclusions: PCOS symptoms are observed across a severity spectrum that parallels genetic burden. Racial and ethnic group differences exist in PCOS symptomology and metabolic health across different phenotyping strategies.

Introduction

Polycystic ovary syndrome (PCOS) is an endocrine disorder that is the leading cause of infertility in women. 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 heterogenous clinical presentation of PCOS makes it difficult to diagnose and an estimated 75% of women with PCOS remain undiagnosed (1,2).

There are three commonly used diagnostic criteria for PCOS which each have specific symptom requirements for diagnosis. The first diagnostic definition was put forth by the National Institute of Health (NIH) who required an ovulatory phenotype (i.e., oligomenorrhoea) 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, which is commonly referred to as Rotterdam criteria. The Rotterdam diagnostic criteria was designed to expand the PCOS definition from the strict NIH definition by requiring

116 any two out of the following three symptoms: oligoanovulation, hyperandrogenism, or
117 polycystic ovaries (3,5). However, the Androgen Excess and PCOS Society argued that
118 hyperandrogenism is the primary driver of PCOS and thus, created a third diagnostic
119 criteria that requires both hyperandrogenism and ovulatory dysfunction (i.e.,
120 oligoanovulation or polycystic ovaries) (4). PCOS diagnostic criteria were developed to
121 improve and standardize diagnosis, however, patients identified by each approach may
122 have differing symptomologies and only a small fraction of women will meet multiple
123 diagnostic criteria (6,7).

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

136 Most prior studies included small numbers of minority women and many have
137 used differing diagnostic criteria (1,16–18). Without larger studies, it is difficult to
138 accurately characterize the clinical presentation of PCOS and its consequences across

139 populations. Adapting an inclusive phenotyping approach capable of identifying severe
140 to mild PCOS presentations could improve study methods for marginalized
141 communities.

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

152

153 **Methods**

154

155 *Clinical data*

156 Vanderbilt University Medical Center (VUMC) is a tertiary care hospital in
157 Nashville, Tennessee, with several clinics throughout Tennessee and the surrounding
158 states. Medical records have been electronically documented at VUMC since the early
159 1990s (19). The VUMC EHR houses over 2.8 million EHRs in a de-identified clinical
160 research database referred to as the Synthetic Derivative (SD). The SD includes
161 demographics, International Classification of Diseases, Ninth Revision (ICD9) and

162 Tenth Revision (ICD10) codes, procedural codes (CPT), clinical notes, medications, and
163 laboratory values. The Vanderbilt University Institutional Review Board approved this
164 project (IRB 160279).
165
166 *Algorithm development*
167 We identified women with PCOS within the SD using a strict automated
168 phenotyping algorithm that yielded a positive predictive value (PPV) of at least 90%.
169 Our secondary analyses relaxed the requirements of our core algorithm to identify a
170 larger sample of patients with a wider array of PCOS symptoms. In total, we developed
171 four algorithms which are illustrated in **Figure 1**. The demographics of each resultant
172 dataset is reported in **Table 1**.
173 Before applying each algorithm, we filtered the entire SD population to include
174 only females between the ages of 11 to 44 to enrich our population for females of
175 reproductive age. We first deployed the $PCOS_{regex\text{-}strict}$ algorithm which required at least
176 one ICD9 or ICD10 code for: polycystic ovaries (i.e., 256.4 and E28.2) OR hirsutism
177 (i.e., 704.1 and L68.0) AND irregular menses (i.e., all 626 codes, all N91 codes, N92.5,
178 and N92.6) **(Supplementary Table 1)**. Hirsutism was used as clinical evidence of
179 hyperandrogenism. Individuals with any exclusion codes **(Supplementary Table 1)**
180 present in their medical record were removed. Furthermore, the $PCOS_{regex\text{-}strict}$ algorithm
181 required the presence of at least one of the following regular expressions, “polycystic
182 ovaries”, “PCOS”, or “PCO ” in a clinical note. We included inpatient and outpatient note
183 types designated as “Outpatient note”, “Gynecology clinic visit”, “Nursing report”,
184 “Endocrine and Diabetes Clinic Visit”, “Pediatric Endocrinology Patient Visit”, and

185 “Reproductive Endocrinology Clinic Visit”. Next, to broaden the approach, we dropped
186 the regular expression requirements and retained only the ICD codes for the $PCOS_{coded-}$
187 $_{strict}$ algorithm. We then loosened the requirement of the union of ICD codes and instead
188 required the presence of ICD codes for polycystic ovaries OR hirsutism OR irregular
189 menses. In addition to any code for these conditions, we required the presence of a
190 PCOS-related regular expression. The resultant algorithm was termed “$PCOS_{regex-broad}$”.
191 Lastly, we again dropped the regular expression requirement which resulted in the
192 broadest approach ($PCOS_{coded-broad}$) including only the presence of ICD codes for
193 polycystic ovaries OR hirsutism OR irregular menses.
194 To identify controls, the SD was again filtered for females between the ages of 11
195 to 44. The control dataset excluded individuals with one or more exclusion ICD codes or
196 any inclusion codes used to define PCOS case criteria **(Supplementary Table 1)**.
197
198 *Chart review methods for algorithm evaluation*
199 We performed two comprehensive chart reviews to evaluate the performance of
200 the phenotyping algorithms. The charts were reviewed by clinical domain experts (MW
201 and JC) and trained reviewers (KVA, DH, and KS). We first evaluated the positive
202 predictive value (PPV) for each algorithm. We randomly selected 50 charts from each
203 set of algorithm defined cases (200 in total) for manual chart review. We then calculated
204 the proportion of true positive cases based on manual chart review out of the total
205 algorithm identified cases, to define the PPV for each algorithm (**Table 2**).
206

207 Next, we evaluated the sensitivity and specificity of the original $PCOS_{regex\text{-}strict}$
208 algorithm. For this review, we first identified a “medical home” population (defined by the
209 presence of at least 5 ICD codes on separate days over at least a 3-year period (N =
210 752,436)). Then, we filtered for females who met a data floor requirement (i.e., at least
211 one ICD10 code for PCOS, E28.2 (N = 9,507)). We then selected 200 charts at random
212 and performed manual chart review of all available information in the EHR (i.e., codes
213 and clinical notes) to classify individuals into “PCOS cases” and “not PCOS cases”. This
214 manual review clinical determination was then compared to the $PCOS_{regex\text{-}strict}$ algorithm-
215 defined determination to calculate sensitivity and specificity (**Table 3**).
216
217 *Laboratory data*
218 Weight and height measurements are routinely recorded during clinical visits and
219 were extracted from the EHR to calculate body mass index (BMI). Laboratory
220 measurements were extracted and cleaned using QualityLab (20). In brief, 3,075
221 quantitative laboratory tests were extracted from EHRs of 70,704 patients in the SD.
222 Laboratory values with non-numeric values (n=2,865) and labs evaluated on only one
223 patient (n=467) were removed. The remaining labs were cleaned by filtering out
224 duplicate entries and measurements with multiple units. Insulin, estradiol, and free
225 testosterone hormone data was then extracted for further analysis. Median laboratory
226 values were calculated for individuals with multiple clinical measurements. Median
227 laboratory values of BMI, insulin, estradiol, and free testosterone were used to
228 characterize the clinical features of each algorithm dataset (**Table 4 and**
229 **Supplementary Table 2**). For each lab, extreme outliers (i.e., 1st and 99th percentile)

were removed. Additionally, lab values were regressed on age at lab and an inverse normal transformation was applied to the residual to result in an age-adjusted normalized lab value for subsequent analysis including genetic analysis (**Supplementary Table 5**). Sample sizes differed for subsequent analyses based on available clinical measurements (**Supplementary Figure 1**).

Genetic data

We included 40,006 females of European descent and 7,736 females of non-European descent who were genotyped on the Illumina MEGA^{EX} Array (Illumina Inc, San Diego, CA) (21). We removed SNPs with a call rate < 0.98, individuals with a call rate of < 0.98, and individuals with a discrepancy between chromosomal and reported sex. EIGENSTRAT was used for principal component analysis (PCA) (22,23). Related samples with a pi-hat > 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>0.03 (24).

Construction of PCOS polygenic risk scores ($PCOS_{PRS}$)

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 SNP beta estimates (25). Beta estimates were drawn from summary statistics of the largest meta-GWAS of PCOS performed by Day et al. which included 5,209 cases and 32,055 controls of European descent diagnosed using NIH or Rotterdam criteria (26).

253 *$PCOS_{PRS}$* were calculated for a total of 36,487 females of European descent, 5,582 non-
254 European descent females, and 29,511 females from multiple ancestries in BioVU.
255
256 *Statistical Analysis*
257 The prevalence of each algorithm was calculated using the proportion of cases
258 identified by each algorithm among total number of EHRs (**Table 2**). The PPV was
259 determined by calculating the proportion of "reviewer-defined" true positives among
260 "algorithm-defined" positives (**Table 2**). Sensitivity was defined as the number of
261 "algorithm-defined" positives out of the total positives (i.e., "algorithm-defined" positives
262 plus "reviewer-defined" positives). Specificity was defined as the number of "algorithm-
263 defined" negatives out of the total negatives (i.e., "algorithm defined and "reviewer-
264 defined" negatives) (**Table 3**).
265 We performed Wilcoxon rank-sum tests to determine if the median BMI and
266 laboratory measurements (e.g., free testosterone, insulin, and estradiol) differed
267 between the PCOS algorithm-identified cases and controls. We then tested the
268 difference between cases defined by each algorithm and between cases and controls
269 stratified by EHR-reported race. Finally, we evaluated differences between African
270 American, White, and Hispanic women identified as cases by each algorithm (**Table 4,**
271 **Supplementary Table 2, Figures 2-4, Supplementary Figures 2-3**).
272 We performed logistic and linear regression analyses to test the association
273 between $PCOS_{PRS}$ and case/control designation and, separately, the association
274 between $PCOS_{PRS}$ and available hormone data (i.e., estradiol, free testosterone,
275 insulin). The analysis was performed in three populations based on genetic ancestry:

European descent (n=6,930), non-European (n = 2,050), and multi-ancestry (n=7,135). Our multi-ancestry cohort includes all genotyped individuals (European and non-European descent). Covariates included median age (calculated across the medical record for each individual) and top ten principal components of ancestry for the case/control logistic regression model. A separate sensitivity analysis adjusting BMI was also performed (**Supplementary Table 3**). For regression analysis of hormone data, covariates included the top ten principal components of ancestry. We then evaluated the model and variance explained by $PCOS_{PRS}$ using Nagelkerke's pseudo R^2 for the logistic regressions and R^2 for the linear regressions (**Table 5**, **Supplementary Table 3-4**). All of the analyses were done in R 3.3.3.

We included two corrections for multiple testing. The first corrected threshold reflects adjustment for each of the independent tests (n = 30, $p < 0.002$, indicated by * in tables). The second corrected threshold is more stringent and reflects adjustment for each of the tests (n = 120, $p < 4x10^{-3}$, indicated by ** in tables) regardless of whether they were independent or performed on nested subsets of the data (**Figure 1**).

Results

Sample Characteristics

PCOS dataset demographics are reported in **Table 1**. Although the sample sizes differed between the algorithms, each algorithm identified a similar proportion of individuals across racial/ethnic groups. The most stringent algorithm, $PCOS_{regex\text{-}strict}$,

identified 4,593 potential PCOS cases including 3,143 White, 658 African American and 204 Hispanic women (**Table 2**). As the algorithm requirements were relaxed, the number of identified PCOS cases increased. $PCOS_{regex\text{-}broad}$ identified 6,193 individuals including 4,207 White, 899 African American, and 274 Hispanic women. The $PCOS_{coded\text{-}strict}$ approach identified 8,340 individuals including 5,526 White, 1,217 African American, and 368 Hispanic women. At the broadest extreme, the $PCOS_{coded\text{-}broad}$ (n = 41,205) algorithm captured the largest and most diverse sample with 26,317 White, 7,648 African American, and 2,052 Hispanic women.

The individuals identified by the regex algorithms had the highest average medical record length at 4.69 years (SD = 4.22) for $PCOS_{regex\text{-}broad}$ and 4.66 years (SD = 4.13) for $PCOS_{regex\text{-}strict}$. This was expected as the increased length of record also increased the probability of PCOS documentation in the medical record. The $PCOS_{coded\text{-}strict}$ and $PCOS_{regex\text{-}strict}$ 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), respectively. Comparatively, the control dataset had the youngest sample aged 23.85 (SD = 9.47) with the shortest EHR record length at 3.92 (SD = 3.9). As expected, the control sample had the lowest average clinical measurements for BMI, insulin, estradiol, and free testosterone.

Evaluation of algorithm performance

We found little difference between the PPVs calculated for the $PCOS_{regex\text{-}strict}$ algorithm using two distinct chart review methods. $PCOS_{regex\text{-}strict}$ yielded a PPV of 96%

322 using our first method of chart review (**Table 2**) and a PPV of 95% when it was tested
323 against the gold standard definition in our second manual review method (**Table 3**).
324 Despite this, $PCOS_{regex\text{-}strict}$ algorithm had a sensitivity of only 50% and specificity of
325 29% when evaluated using the gold standard definition (**Table 3**).
326 When compared to our broader algorithms, the performance of the $PCOS_{regex\text{-}strict}$
327 algorithm was most similar to the $PCOS_{coded\text{-}strict}$ algorithm which had a slightly higher
328 PPV of 98% (**Table 2**). The text mining criteria did not significantly improve the accuracy
329 of $PCOS_{regex\text{-}strict}$ over $PCOS_{coded\text{-}strict}$. The accuracy of the "broad" criteria decreased
330 compared to the "strict" algorithms. Our broadest algorithm, $PCOS_{coded\text{-}broad}$, had the
331 lowest performance with a PPV of 30% (**Table 2**). This is likely due to the large number
332 of patients who only had an irregular menstruation ICD code with no other diagnosed
333 PCOS symptoms. Interestingly, the addition of the text-mined PCOS keyword
334 significantly improved the PPV of the broad criteria algorithm from 30% ($PCOS_{coded\text{-}broad}$)
335 to 82% ($PCOS_{regex\text{-}broad}$). Unlike the extreme broad criteria, the other algorithms had a
336 PCOS prevalence that reflected the prevalence observed in the population (18,27,28).
337 This was also true for the race stratified cohorts where there were little to no differences
338 in prevalence.
339
340 *Clinical feature characterization*
341 We characterized the clinical profiles of cases and controls identified by each
342 algorithm by calculating median BMI, insulin, estradiol, and free testosterone levels
343 (**Table 4**). Each algorithm identified set was significantly heavier compared to controls
344 ($p < 0.05$). For all algorithms, African American and Hispanic women had higher

345 baseline metabolic profiles for BMI, insulin, estradiol, and free testosterone
346 measurements compared to White women (1,2). Moreover, the median BMI in the
347 algorithm datasets increased with increasing algorithm stringency (**Figure 2a**). African
348 American women meeting algorithm criteria consistently demonstrated the highest
349 median BMI in the $PCOS_{coded\text{-}broad}$, $PCOS_{coded\text{-}strict}$, and $PCOS_{regex\text{-}broad}$ algorithms
350 (**Supplementary Table 2**). The BMI of all three stratified race groups differed
351 significantly from each other with the exception of White and Hispanic women in the
352 strictest regex algorithms (**Figure 3a)**. Within the Hispanic defined group, the median
353 values did not vary significantly between the $PCOS_{coded\text{-}strict}$, $PCOS_{regex\text{-}broad}$, and
354 $PCOS_{regex\text{-}strict}$ groups (**Supplementary Figure 3a**).
355 Overall, women identified by each of the four PCOS algorithms demonstrated
356 significantly higher insulin and free testosterone levels compared to controls ($p < 0.05$)
357 (**Figures 2b, 2d, Table 4**). The median hormone values also increased with algorithm
358 stringency, though the trend was not as pronounced as it was for the BMI
359 measurements. BMI increased with algorithm stringency within each race stratified
360 analysis. Insulin values increased with algorithm stringency for White and Hispanic
361 women, and free testosterone values increased with algorithm stringency for White and
362 African American women (**Figure 4**, **Supplementary Table 2**).
363 Measurements of free testosterone were significantly higher among women
364 identified by the $PCOS_{coded\text{-}strict}$, $PCOS_{regex\text{-}broad}$ and $PCOS_{regex\text{-}strict}$ algorithms compared
365 to the controls and the $PCOS_{coded\text{-}broad}$ algorithm (**Supplementary Figure 2d**). The
366 insulin profile of African Americans was similar across women identified by each
367 algorithm and also did not significantly differ from controls (**Figure 4b**). However,

368 African American women identified by the $PCOS_{coded\text{-}broad}$ and $PCOS_{regex\text{-}broad}$ algorithms
369 did exhibit higher estradiol levels than African American controls (**Figure 4c**).
370 Additionally, African American women identified by the $PCOS_{regex\text{-}broad}$ algorithm
371 demonstrated significantly higher estradiol levels than White women identified by the
372 same approach (**Figure 3c**). Outside of these findings, there were no other significant
373 differences between races for hormone values (**Figure 3**).
374
375 *Genetic validation of PCOS case status*
376
377 Increasing algorithm requirement stringency was also reflected in the
378 performance of the $PCOS_{PRS}$. Odds ratios (ORs), which are reported per standard
379 deviation of the $PCOS_{PRS}$ were higher among cases identified by algorithms using strict
380 PCOS requirements and decreased as those requirements were relaxed. This trend
381 was observed for both the European and multi-ancestry genotyped samples (**Figure 5,**
382 **Table 5**). BMI was tested as a covariate in the model and did not impact the results.
383 (**Supplementary Table 3**).
384 Among European genotyped cases identified by the $PCOS_{regex\text{-}broad}$ algorithm we
385 observed an OR of 1.29 (p = 4.6E-5) and the greatest variance explained in
386 case/control status by the $PCOS_{PRS}$ model (pseudo-r^2 = 0.83%, p = 4.6E-5) (**Table 5**).
387 When the $PCOS_{PRS}$ model was applied to the multi-ancestry sample, the $PCOS_{coded\text{-}broad}$
388 yielded the highest OR (OR = 1.49, p = 8.7E-5), followed by $PCOS_{regex\text{-}strict}$ (OR = 1.44,
389 p = 0.003) and $PCOS_{regex\text{-}broad}$ (OR = 1.41, p = 0.001). The $PCOS_{coded\text{-}strict}$ algorithm
390 (pseudo-r^2 = 0.7%, p = 8.7E-5) also had the most variance explained by the model.

391 Similar to the European genotyped sample, only the $PCOS_{regex\text{-}broad}$ and the $PCOS_{coded\text{-}strict}$ algorithm passed at least one of our two multiple correction tests.

393 There were no significant associations between $PCOS_{PRS}$ and case/control
394 status within the non-European genotyped sample. We are underpowered in this
395 analysis, which is evidenced by large confidence intervals in comparison to the
396 European and multi-ancestry genotyped sample (**Figure 5, Table 5**).

397

398 *Association between* $PCOS_{PRS}$ *and hormone measurements*

399 There was no association between $PCOS_{PRS}$ and any of the hormone measures
400 (insulin, estradiol, and free testosterone) evaluated. (**Supplementary Tables 4-6**).

401

402 **Discussion**

403

404 Expanding study inclusion criteria has the potential to increase and diversify
405 future study samples which may significantly benefit underrepresented high risk groups
406 (27). This study describes an automated PCOS phenotyping method that captures
407 varying degrees of PCOS symptoms, parallel dysregulated hormone levels, and
408 increased genetic risk for the $PCOS_{coded\text{-}strict}$, $PCOS_{regex\text{-}broad}$, and $PCOS_{regex\text{-}strict}$
409 algorithms. Our results also further support the hypothesis that women with PCOS
410 symptoms have increased metabolic dysregulation compared to controls when
411 measured on BMI, insulin, and free testosterone values dependent on clinical definitions
412 and their EHR-reported race. Finally, our study demonstrates the importance of

413 including multiethnic genotyped samples in addition to European populations, for
414 genetic analysis.

415 Our two “strict definition” algorithms identified PCOS cases with high accuracy
416 (PPV>95%) and the $PCOS_{regex\text{-}strict}$ algorithm performed just as well when compared to
417 the gold standard dataset (PPV = 95%). Although the sensitivity and the specificity of
418 the $PCOS_{regex\text{-}strict}$ algorithm was low, the high PPV displays its high accuracy within an
419 EHR setting. Compared to the previously reported ICD based PCOS algorithm, which
420 had a PPV of 66% and 65% for their “broad” and “refined” algorithm, our two best
421 performing algorithms improved case identification by over 30% (9). Both the Castro et
422 al. study and our study show the feasibility of using ICD based algorithms to identify
423 PCOS cases and the advantages of using EHR systems. Consistent with the previous
424 study, we also found that inclusion of text mining increased the proportion of true PCOS
425 cases (PPV = 98%) but resulted in a smaller sample size. Removal of the text mining
426 requirement only slightly impacted the PPV (96%), indicating that studies performed in
427 EHR systems with limited access to clinical notes and ultrasound images may still be
428 able to identify women with PCOS with high accuracy. Conversely, we found that the
429 addition of text mining to the broad algorithms significantly improved the accuracy from
430 30% to 82%, indicating that clinical suspicions of PCOS documented in medical records
431 may also be used to identify women with PCOS even when few symptoms are coded.

432 The clinical and epidemiological features of our PCOS cohorts are consistent
433 with previous studies that show women with PCOS have higher metabolic clinical
434 characteristics than controls (29–31). We also observed no significant difference in the
435 prevalence of PCOS definitions between minority and White individuals, nor were there

436 any large differences between races for insulin, estradiol, and free testosterone clinical
437 measurements (32). However, African American had significantly higher BMI
438 measurements than Hispanic and White women across both the control and PCOS
439 definition datasets. African Americans are disproportionately affected by conditions such
440 as obesity and type 2 diabetes, which could possibly be exacerbated by the onset of
441 PCOS (33,34). These conditions could also be influenced by the patient structure of the
442 EHR database which will have a sicker population due to the tertiary care status of
443 VUMC and due to its geographic location where obesity rates are generally higher.
444 Polygenic risk scores provide the unique ability to examine the genetic
445 susceptibility of PCOS across different clinical definitions and ancestries. However, it is
446 important to note that, as of to date, there are no PCOS GWAS in African descent
447 populations, and that PRS built from Eurocentric populations will not perform as robustly
448 in non-European samples (35). Despite this, previous studies have shown that
449 $PCOS_{PRS}$ models can detect cross-ancestry genetic risk in African and multi-ancestry
450 cohorts using European based PRS models (36). In our study, the $PCOS_{PRS}$ model
451 explained little variance in case/control status (pseudo-$r^2 < 1\%$), but the $PCOS_{PRS}$ was
452 significantly higher among cases in the $PCOS_{coded\text{-}strict}$, $PCOS_{regex\text{-}broad}$, and the
453 $PCOS_{regex\text{-}strict}$ algorithms compared to controls.
454 Eurocentric GWAS remain a significant limitation to interpretability of PRS in non-
455 European samples and studies including genetic data gathered on diverse ancestral
456 populations are needed to advance PCOS genomics in the EHR (35,37). The limitations
457 of our study include low statistical power in the genetic analysis of non-European cases
458 and controls. This limitation is primarily a function of the under-representation of non-

459 European populations in the PCOS GWAS, but also reflects the low representation of
460 non-European samples in the VUMC EHR.

461 In conclusion, we applied an automated phenotyping approach to examine
462 variation in the symptomology of PCOS across racial and ethnic groups using both strict
463 and broad classification criteria. Debate over current implemented PCOS diagnostic
464 criteria, in part, reflects the complexity of PCOS symptoms and EHRs provide significant
465 advantages. First, research in the longitudinal EHR allows us to maximize the number
466 of PCOS clinical manifestations observed over time (16,28,38). Secondly, EHRs contain
467 a variety of clinical variables (e.g., laboratory measurements, imaging records, etc.) that
468 can be used to study the risk factors and health outcomes of PCOS patients. We
469 created a simple EHR-based PCOS algorithm that identifies patients using ICD codes
470 and text mining ($PCOS_{regex\text{-}strict}$) methods and by relaxing the inclusion requirements we
471 identified patients across the PCOS syndromic spectrum. This study approach provides
472 a unique view of the shared clinical and genetic architecture that influences both PCOS
473 as a discrete diagnosis, and the symptoms of PCOS that affect a large number of
474 women.

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;(1):99.
3. **ESHRE TR, Group A-SPCW.** Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. *Fertil. Steril.* 2004;81(1):19–25.
4. **Azziz R, Carmina E, Dewailly D, Diamanti-Kandarakis E, Escobar-Morreale F, Futterweit W, Janssen OE, Legro RS, Norman RJ, Taylor AE, Witchel SF, (Task Force on the Phenotype of the Polycystic Ovary Syndrome of The Androgen Excess and PCOS Society)∗, Ector H, Escobar-Morreale F, Futterweit W, Janssen OE, Legro RS, Norman RJ, Taylor AE, Witchel SF.** 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 JK, 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. **Stepto NK, Cassar S, Joham AE, Hutchison SK, Harrison CL, Goldstein RF, Teede HJ.** Women with polycystic ovary syndrome have intrinsic insulin resistance on euglycaemic-hyperinsulaemic clamp. *Hum. Reprod.* 2013;28(3):777–784.
7. **Hickey M, Doherty DA, Atkinson H, Sloboda DM, Franks S, Norman RJ, Hart R.** 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, McCaffery K.** Are expanding disease definitions unnecessarily labelling women with polycystic ovary syndrome? *BMJ* 2017;358. doi:10.1136/bmj.j3694.
9. **Castro V, Shen Y, Yu S, Finan S, Pau CT, Gainer V, Keefe CC, Savova G, Murphy SN, Cai T, Welt CK.** Identification of subjects with polycystic ovary syndrome using electronic health records. *Reprod. Biol. Endocrinol.* 2015;13(1):1–8.
10. **Jacewicz-Święcka M, Kowalska I.** Polycystic ovary syndrome and the risk of cardiometabolic complications in longitudinal studies. *Diabetes. Metab. Res. Rev.* 2018;34(8). doi:10.1002/dmrr.3054.
11. **Pavicic Baldani D, Skrgatic L, Ougouag R.** Polycystic Ovary Syndrome: Important Underrecognised Cardiometabolic Risk Factor in Reproductive-Age Women. 2015. doi:10.1155/2015/786362.
12. **Chan JL, Kar S, Vanky E, Morin-Papunen L, Piltonen T, Puurunen J, Tapanainen JS, Maciel GAR, Hayashida SAY, Soares Jr JM.** 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-

521 sectional study. *Am. J. Obstet. Gynecol.* 2017;217(2):189. e1-189. e8.
522 13. **Engmann L, Jin S, Sun F, Legro RS, Polotsky AJ, Hansen KR, Coutifaris C,**
523 **Diamond MP, Eisenberg E, Zhang H, Santoro N.** Racial and ethnic differences
524 in the polycystic ovary syndrome metabolic phenotype. 2017.
525 doi:10.1016/j.ajog.2017.01.003.
526 14. **Meyer ML, Sotres-Alvarez D, Steiner AZ, Cousins L, Talavera GA, Cai J,**
527 **Daviglus ML, Loehr LR.** Polycystic ovary syndrome signs and metabolic
528 syndrome in premenopausal Hispanic/Latina women: The HCHS/SOL study. *J.*
529 *Clin. Endocrinol. Metab.* 2020. doi:10.1210/clinem/dgaa012.
530 15. **Hillman JK, Johnson LNC, Limaye M, Feldman RA, Sammel M, Dokras A.**
531 Black women with polycystic ovary syndrome (PCOS) have increased risk for
532 metabolic syndrome and cardiovascular disease compared with white women with
533 PCOS. *Fertil. Steril.* 2014;101(2):530–535.
534 16. **March WA, Moore VM, Willson KJ, Phillips DIW, Norman RJ, Davies MJ.** The
535 prevalence of polycystic ovary syndrome in a community sample assessed under
536 contrasting diagnostic criteria. *Hum. Reprod.* 2010;25(2):544–551.
537 17. **Ding T, Hardiman PJ, Petersen I, Wang F-F, Qu F, Baio G.** The prevalence of
538 polycystic ovary syndrome in reproductive- aged women of different ethnicity: a
539 systematic review and meta-analysis. *Oncotarget* 2017;8(56):96351–96358.
540 18. **Bozdag G, Mumusoglu S, Zengin D, Karabulut E, Yildiz BO.** The prevalence
541 and phenotypic features of polycystic ovary syndrome: A systematic review and
542 meta-Analysis. *Hum. Reprod.* 2016;31(12):2841–2855.
543 19. **Robinson JR, Wei W-Q, Roden DM, Denny JC.** Defining phenotypes from
544 clinical data to drive genomic research. *Annu. Rev. Biomed. Data Sci.* 2018;1:69–
545 92.
546 20. **Dennis JK, Sealock JM, Straub P, Hucks D, Actkins K, Faucon A, Goleva SB,**
547 **Nirachou M, Singh K, Morley T, Ruderfer DM, Mosley JD, Chen G, Davis LK.**
548 Lab-wide association scan of polygenic scores identifies biomarkers of complex
549 disease. *medRxiv* 2020:2020.01.24.20018713.
550 21. **Bien SA, Wojcik GL, Zubair N, Gignoux CR, Martin AR, Kocarnik JM, Martin**
551 **LW, Buyske S, Haessler J, Walker RW, Cheng I, Graff M, Xia L, Franceschini**
552 **N, Matise T, James R, Hindorff L, Le Marchand L, North KE, Haiman CA,**
553 **Peters U, Loos RJF, Kooperberg CL, Bustamante CD, Kenny EE, Carlson**
554 **CS.** Strategies for Enriching Variant Coverage in Candidate Disease Loci on a
555 Multiethnic Genotyping Array. *PLoS One* 2016;11(12):167758.
556 22. **Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D.**
557 Principal components analysis corrects for stratification in genome-wide
558 association studies. *Nat. Genet.* 2006;38(8):904–909.
559 23. **Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ.** Second-
560 generation PLINK: rising to the challenge of larger and richer datasets.
561 *Gigascience* 2015;4(7):7.
562 24. **Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew**
563 **EY, Levy S, McGue M, Schlessinger D, Stambolian D, Loh PR, Iacono WG,**
564 **Swaroop A, Scott LJ, Cucca F, Kronenberg F, Boehnke M, Abecasis GR,**
565 **Fuchsberger C.** Next-generation genotype imputation service and methods. *Nat.*
566 *Genet.* 2016;48(10):1284–1287.

25. **Ge T, Chen C-Y, Ni Y, Feng Y-CA, Smoller JW.** Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* 2019;10(1776):1–10.
26. **Day F, Karaderi T, Jones MR, Meun C, He C, Drong A, Kraft P, Lin N, Huang H, Broer L.** Large-scale genome-wide meta-analysis of polycystic ovary syndrome suggests shared genetic architecture for different diagnosis criteria. *PLoS Genet.* 2018;14(12):e1007813.
27. **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.
28. **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;7(1). doi:10.1093/humupd/7.1.1.
29. **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. 2014. doi:10.1371/journal.pone.0108263.
30. **Antonio L, Pauwels S, Laurent MR, Vanschoubroeck D, Jans I, Billen J, Claessens F, Decallonne B, De Neubourg D, Vermeersch P, Vanderschueren D.** Free Testosterone Reflects Metabolic as well as Ovarian Disturbances in Subfertile Oligomenorrheic Women. *Int. J. Endocrinol.* 2018. doi:10.1155/2018/7956951.
31. **Diamanti-Kandarakis E, Dunaif A.** Insulin Resistance and the Polycystic Ovary Syndrome Revisited: An Update on Mechanisms and Implications. 2012. doi:10.1210/er.2011-1034.
32. **Ladson G, Dodson WC, Sweet SD, Archibong AE, Kunselman AR, Demers LM, Williams NI, Coney P, Legro RS.** Racial Influence on the Polycystic Ovary Syndrome Phenotype: A Black White Case-Control Study. *Fertil Steril* 2011;96(1):224–229.
33. **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.
34. **Butts SF, Seifer DB.** Racial and ethnic differences in reproductive potential across the life cycle. *Fertil. Steril.* 2010;93(3):681–690.
35. **Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ.** Clinical use of current polygenic risk scores may exacerbate health disparities. 2019;51(4):584–591.
36. **Joo YY, Actkins K, Pacheco JA, Basile AO, Carroll R, Crosslin DR, Day F, Denny JC, Velez Edwards DR, Hakonarson H, Harley JB, Hebbring SJ, Ho K, Jarvik GP, Jones M, Karderi T, Mentch FD, Meun C, Namjou B, Pendergrass S, Ritchie MD, Stanaway IB, Urbanek M, Walunas TL, Smith M, Chisholm RL, Kho AN, Davis L, Hayes MG.** A polygenic and phenotypic risk prediction for Polycystic Ovary Syndrome evaluated by Phenome-wide association studies. *J. Clin. Endocrinol. Metab.* 2020. doi:10.1210/clinem/dgz326.
37. **Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, Highland**

HM, Patel YM, Sorokin EP, Avery CL, Belbin GM, Bien SA, Cheng I, Cullina S, Hodonsky CJ, Hu Y, Huckins LM, Jeff J, Justice AE, Kocarnik JM, Lim U, Lin BM, Lu Y, Nelson SC, Park S-SLSL, Poisner H, Preuss MH, Richard MA, Schurmann C, Setiawan VW, Sockell A, Vahi K, Verbanck M, Vishnu A, Walker RW, Young KL, Zubair N, Acuña-Alonso V, Ambite JL, Barnes KC, Boerwinkle E, Bottinger EP, Bustamante CD, Caberto C, Canizales-Quinteros S, Conomos MP, Deelman E, Do R, Doheny K, Fernández-Rhodes L, Fornage M, Hailu B, Heiss G, Henn BM, Hindorff LA, Jackson RD, Laurie CACC, Laurie CACC, Li Y, Lin D-YY, Moreno-Estrada A, Nadkarni G, Norman PJ, Pooler LC, Reiner AP, Romm J, Sabatti C, Sandoval K, Sheng X, Stahl EA, Stram DO, Thornton TA, Wassel CL, Wilkens LR, Winkler CA, Yoneyama S, Buyske S, Haiman CA, Kooperberg C, Le Marchand L, Loos RJFF, Matise TC, North KE, Peters U, Kenny EE, Carlson CS. *Genetic analyses of diverse populations improves discovery for complex traits*. Nature Publishing Group; 2019:514–518.

38. **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.

633 **Figure Legends**
634
635
636 **Figure 1. PCOS algorithm pipeline.** This schematic illustrates the algorithm
637 development steps for the controls and the four PCOS algorithms. The Venn diagram
638 illustrates how the algorithm datasets are nested within one another. Each color
639 corresponds with one of the algorithm datasets. ICD = International classification of
640 disease, 9th and 10th revision.
641
642 **Figure 2. Laboratory measurements for PCOS algorithm cases and controls.**
643 Boxplots of (a) BMI measurements, (b) insulin levels, (c) estradiol levels, and (d) free
644 testosterone levels for PCOS algorithm cases and controls. Colored boxes represent
645 races. Boxes represent the individuals with lab measurements in the 25th and 75th
646 percentile. Lines above and below the boxes represent the 95th and 5th quartiles. Lines
647 within each box mark the median. Wilcoxon rank-sum tests were performed between
648 algorithm cases and controls and Brackets display the p-values of the statistical test.
649 BMI= body mass index.
650
651 **Figure 3. Comparison of laboratory measurements between White, African**
652 **American, and Hispanic PCOS algorithm cases and controls.** Boxplots of (a) BMI
653 measurements, (b) insulin levels, (c) estradiol levels, and (d) free testosterone levels for
654 White, African American, and Hispanic PCOS cases and controls. Colors correspond
655 with each race. Boxes represent the 25th and 75th quartiles. Lines above and below the
656 boxes represent the 95th and 5th quartiles. Lines within each box mark the median.
657 Wilcoxon rank-sum tests were performed between races within the PCOS cases and

658 control datasets. Brackets display the p-values of the statistical test. BMI= body mass
659 index; W = White; AA = African American; HIS = Hispanic.
660
661 **Figure 4. Race stratified laboratory measurements for PCOS algorithm cases and**
662 **controls.** Boxplots of (a) BMI measurements, (b) insulin levels, (c) estradiol levels, and
663 (d) free testosterone levels of race stratified PCOS algorithm datasets. Each color
664 corresponds with a race. Boxes represent the 25th and 75th quartiles. Lines above and
665 below the boxes represent the 95th and 5th quartiles. Lines within each box mark the
666 median. Wilcoxon rank-sum tests were performed between algorithm cases and
667 controls within each race. Brackets display the p-values of the statistical test. BMI= body
668 mass index; CN = controls; CB = coded broad; CS = coded strict; RB = regex broad; RS
669 = regex strict.
670
671 **Figure 5. Genetic validation of PCOS algorithms.** Logistic regressions were
672 performed between PCOS case status and PCOS polygenic risk scores ($PCOS_{PRS}$).
673 The regression models were adjusted for median age and the first 10 principal
674 components. P-values are displayed for each regression model. Colors correspond to
675 the genotyped sample used in the analysis. OR = odds ratio.

Table 1. PCOS algorithm descriptive statistics. Descriptive statistics of women identified as PCOS cases and controls are provided. SD = standard deviation; EHR = electronic health record. NH = non-Hispanic, HIS = Hispanic.

	Controls	Coded Broad	Coded Strict	Regex Broad	Regex Strict
N	29140	41205	8340	6193	4593
NH White	22569	25977	5418	4154	3101
NH African American	3278	7555	1195	887	648
Hispanic	1421	2053	368	274	204
HIS White	1014	1462	264	203	154
HIS African American	33	69	13	11	8
Average Age (SD)	23.85 (9.47)	24.96 (7.97)	25.18 (7.63)	24.71 (7.52)	25.09 (7.39)
Average EHR Record Length (SD)	3.92 (3.9)	4.52 (4.54)	4.23 (4.15)	4.69 (4.22)	4.66 (4.13)
Medical Home (%)	65%	64%	60%	66%	66%
Average Clinical Values (SD)					
Body Mass Index	25.71 (7.43)	28.61 (8.44)	34.12 (9.10)	33.50 (9.14)	34.54 (9.07)
Insulin (mcU/mL)	20.08 (18.99)	24.86 (22.83)	27.17 (23.05)	25.34 (21.94)	27.27 (21.58)
Estradiol (pg/mL)	82.22 (69.03)	89.92 (69.74)	75.92 (61.13)	71.64 (56.56)	75.57 (60.49)
Free Testosterone (pg/mL)	3.58 (2.45)	4.70 (2.94)	5.70 (3.23)	5.41 (3.01)	5.73 (3.09)

Table 2. Prevalence for PCOS algorithms and chart review results. Prevalence of PCOS was calculated for everyone identified in each algorithm (overall) and for White, African American, and Hispanic women identified in each algorithm. Chart review was conducted on 50 random records identified by each PCOS algorithm (n = 200) and algorithm performance was measured through PPV. PPV = positive predictive value.

	Coded Broad	Coded Strict	Regex Broad	Regex Strict
Overall				
N	41205	8340	6193	4593
Prevalence	46%	9.3%	6.9%	5.1%
PPV	30%	98%	82%	96%
White				
N	26317	5526	4207	3143
Prevalence	50%	10.5%	8%	4.8%
African American				
N	7648	1217	899	658
Prevalence	65.9%	10.5%	7.8%	5.7%
Hispanic				
N	2052	368	274	204
Prevalence	48%	8.6%	6.4%	4.8%

Table 3. Sensitivity and specificity for PCOS algorithm datasets. A 2x2 contingency table showing the PPV, NPV, sensitivity, and specificity of the regex strict (RS) algorithm. The RS algorithm was compared to the gold standard definition. Chart review was conducted on 200 random records from females who met our data floor and medical home definition. PPV = positive predictive value; NPV = negative predictive value.

		Regex Strict Algorithm		
		Case	Not Case	**Results**
Gold Standard	Case	96	97	50% Sensitivity
	Not case	5	2	29% Specificity
	Results	95% PPV	2% NPV	

Table 4. Laboratory measurements for PCOS algorithm cases and controls. Wilcoxon rank sum test was performed between cases and controls. The 25th and 75th percentile was reported for each algorithm. BMI = body mass index.

Algorithm	Type	N	Median	25th Percentile	75th Percentile	P	
BMI							
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		
Regex Broad	Cases	5844	32.805	26.29	39.70	< 2.22E-16	**
	Controls	29140	24.03	20.68	29.30		
Regex Strict	Cases	4358	34.08	27.63	40.76	< 2.22E-16	**
	Controls	29140	24.03	20.68	29.30		
Insulin							
Coded Broad	Cases	257	17.4	10.00	24.10	0.0001	**
	Controls	14475	13.65	5.50	20.90		
Coded Strict	Cases	150	19.05	11.20	28.70	1.50E-06	**
	Controls	14475	13.65	5.50	20.90		
Regex Broad	Cases	134	18.1	10.52	26.06	0.00088	*
	Controls	14475	13.65	5.50	20.90		
Regex Strict	Cases	109	19.4	11.53	32.53	2.40E-05	**
	Controls	14475	13.65	5.50	20.90		
Estradiol							
Coded Broad	Cases	196	63.5	38.00	116.00	0.0081	
	Controls	18799	55.5	32.00	108.00		
Coded Strict	Cases	67	54	38.00	96.65	0.79	
	Controls	18799	55.5	32.00	108.00		
Regex Broad	Cases	58	54.5	38.50	90.88	0.86	
	Controls	18799	55.5	32.00	108.00		
Regex Strict	Cases	48	55.5	39.50	93.20	0.64	
	Controls	18799	55.5	32.00	108.00		
Free Testosterone							
Coded Broad	Cases	159	4.4	2.25	6.58	0.0016	*
	Controls	125	3	1.70	4.95		
Coded Strict	Cases	75	5.9	3.00	7.70	3.80E-06	**
	Controls	125	3	1.70	4.95		
Regex Broad	Cases	73	5.2	2.90	7.40	3.50E-05	**
	Controls	125	3	1.70	4.95		
Regex Strict	Cases	53	5.6	3.60	7.50	1.40E-05	**
	Controls	125	3	1.70	4.95		

Table 5. Genetic validation of PCOS diagnosis. Logistic regressions were performed between PCOS case status and PCOS polygenic risk scores ($PCOS_{PRS}$). The regressions were adjusted for median age and the first 10 principal components. SE = standard error; OR = odds ratio; LCI = lower confidence interval; UCI = upper confidence interval.

Algorithm	Cases	Controls	Estimate	SE	P	OR	LCI	UCI	r^2	
European										
Coded Broad	1878	5052	0.04	0.03	0.235	1.04	0.98	1.10	0.0003	
Coded Strict	366	5052	0.19	0.06	0.002	1.21	1.07	1.36	0.0046	*
Regex Broad	332	5052	0.26	0.06	4.6E-5	1.29	1.14	1.46	0.0083	**
Regex Strict	245	5052	0.21	0.07	0.004	1.23	1.07	1.42	0.0049	
Non-European										
Coded Broad	741	1309	0.09	0.10	0.354	1.10	0.90	1.33	0.0006	
Coded Strict	126	1309	0.32	0.20	0.111	1.38	0.93	2.04	0.0039	
Regex Broad	103	1309	0.28	0.22	0.208	1.32	0.86	2.04	0.0027	
Regex Strict	77	1309	0.21	0.25	0.404	1.24	0.75	2.03	0.0014	
Multi-Ancestry										
Coded Broad	2228	4907	0.08	0.05	0.114	1.08	0.98	1.20	0.0005	
Coded Strict	419	4907	0.40	0.10	8.7E-5	1.49	1.22	1.81	0.0068	**
Regex Broad	382	4907	0.34	0.10	0.001	1.41	1.15	1.73	0.0050	*
Regex Strict	281	4907	0.36	0.12	0.003	1.44	1.13	1.82	0.0050	

Figure 1. PCOS algorithm pipeline. This schematic illustrates the algorithm development steps for the controls and the four PCOS algorithms. The Venn diagram illustrates how the algorithm datasets are nested within one another. Each color corresponds with one of the algorithm datasets. ICD = International classification of disease, 9th and 10th revision.

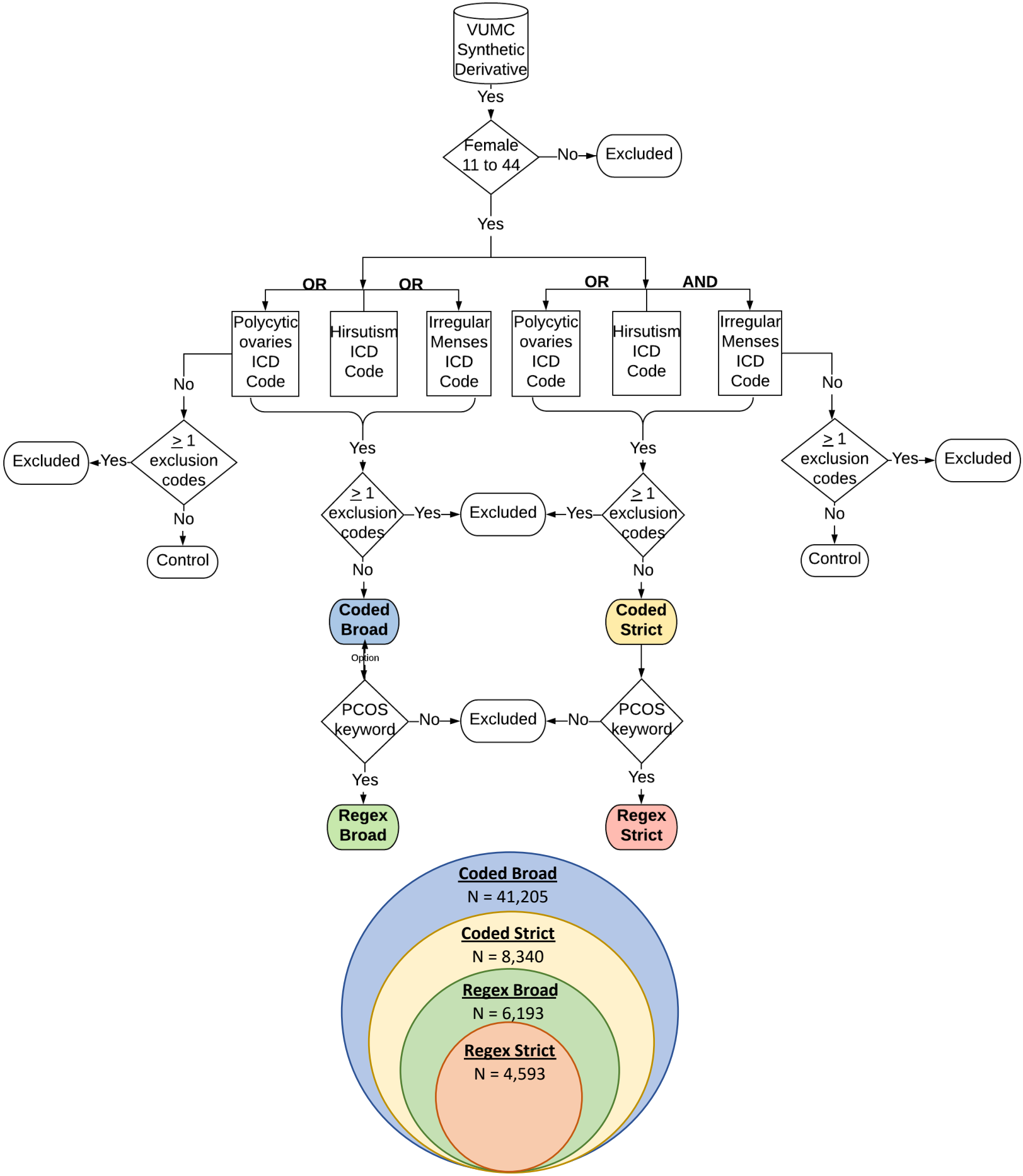

Figure 2. Laboratory measurements for PCOS algorithm cases and controls. Boxplots of (a) BMI measurements, (b) insulin levels, (c) estradiol levels, and (d) free testosterone levels for PCOS algorithm cases and controls. Colored boxes represent races. Boxes represent the individuals with lab measurements in the 25th and 75th percentile. Lines above and below the boxes represent the 95th and 5th quartiles. Lines within each box mark the median. Wilcoxon rank-sum tests were performed between algorithm cases and controls and Brackets display the p-values of the statistical test. BMI= body mass index.

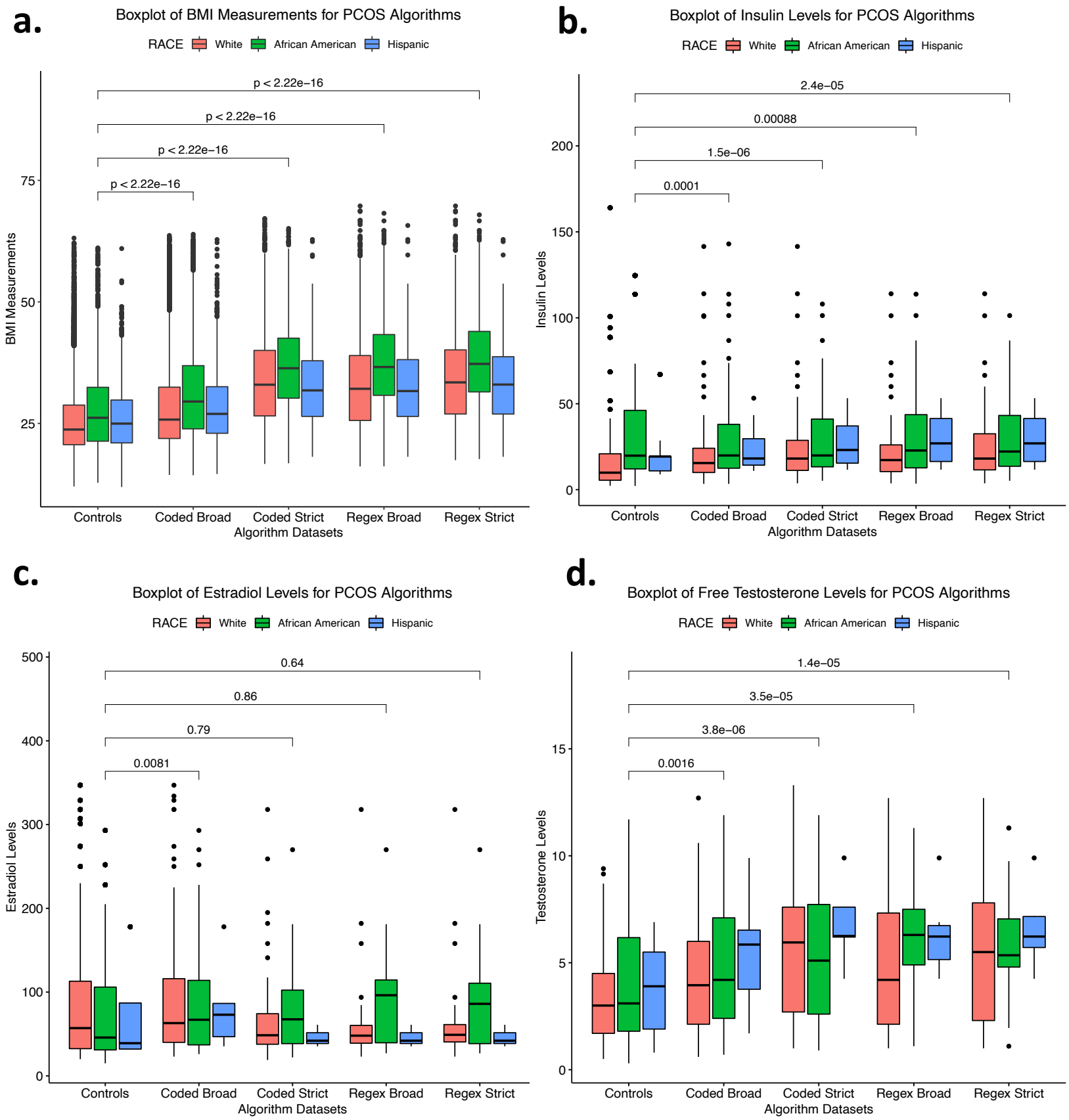

Figure 3. Comparison of laboratory measurements between White, African American, and Hispanic PCOS algorithm cases and controls. Boxplots of (a) BMI measurements, (b) insulin levels, (c) estradiol levels, and (d) free testosterone levels for White, African American, and Hispanic PCOS cases and controls. Colors correspond with each race. Boxes represent the 25th and 75th quartiles. Lines above and below the boxes represent the 95th and 5th quartiles. Lines within each box mark the median. Wilcoxon rank-sum tests were performed between races within the PCOS cases and control datasets. Brackets display the p-values of the statistical test. BMI= body mass index; W = White; AA = African American; HIS = Hispanic.

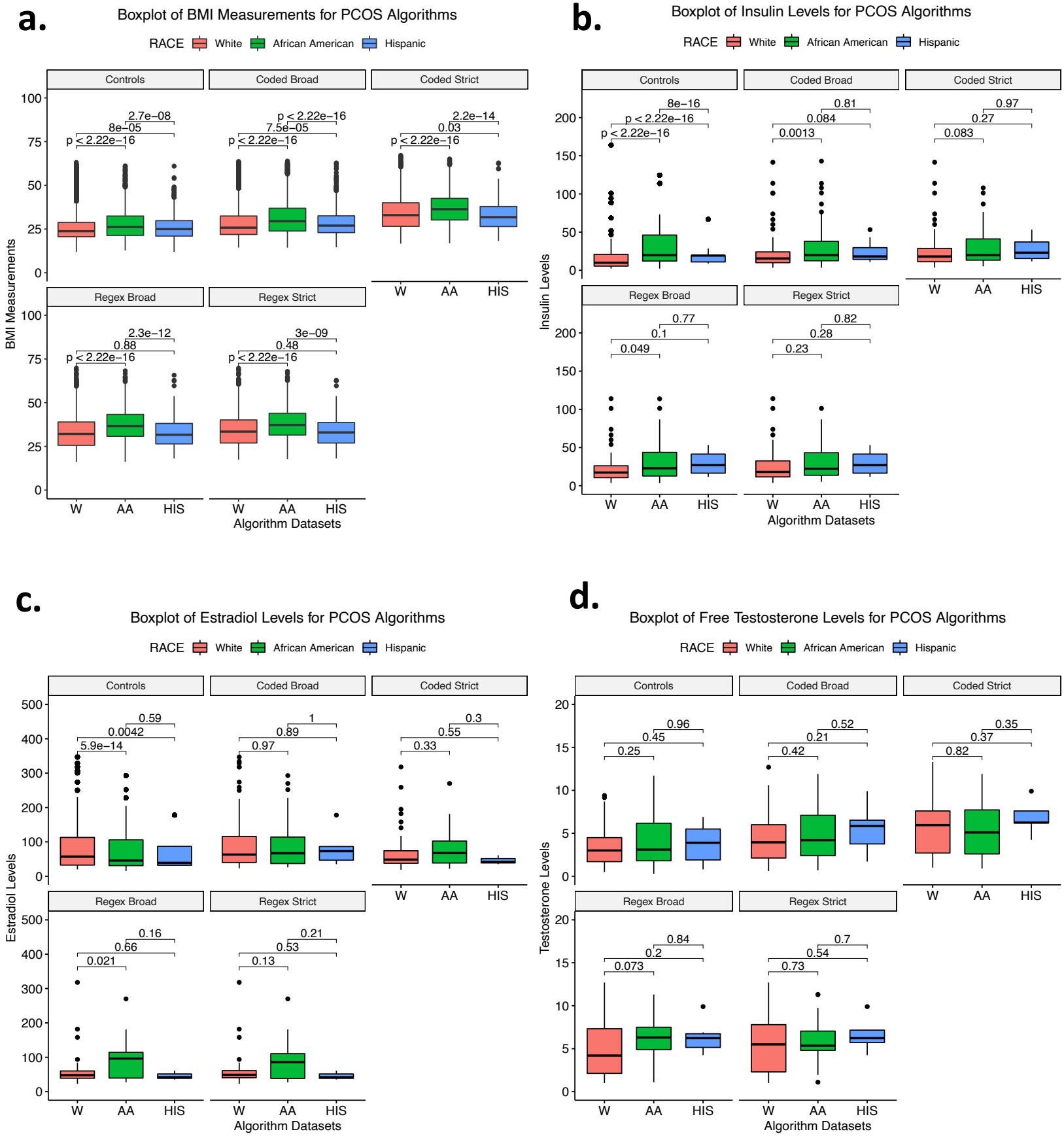

Figure 4. Race stratified laboratory measurements for PCOS algorithm cases and controls. Boxplots of (a) BMI measurements, (b) insulin levels, (c) estradiol levels, and (d) free testosterone levels of race stratified PCOS algorithm datasets. Each color corresponds with a race. Boxes represent the 25th and 75th quartiles. Lines above and below the boxes represent the 95th and 5th quartiles. Lines within each box mark the median. Wilcoxon rank-sum tests were performed between algorithm cases and controls within each race. Brackets display the p-values of the statistical test. BMI= body mass index; CN = controls; CB = coded broad; CS = coded strict; RB = regex broad; RS = regex strict.

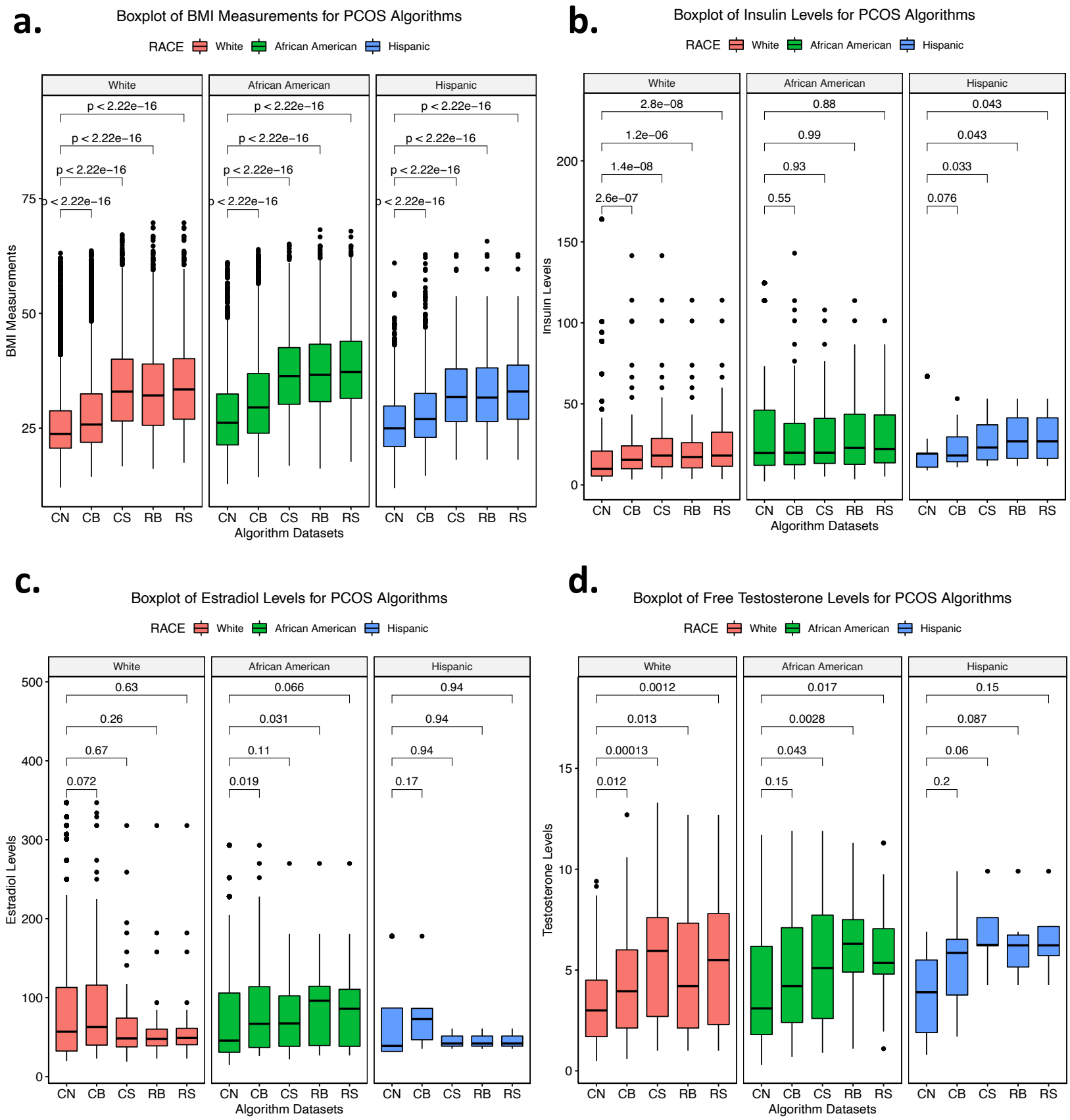
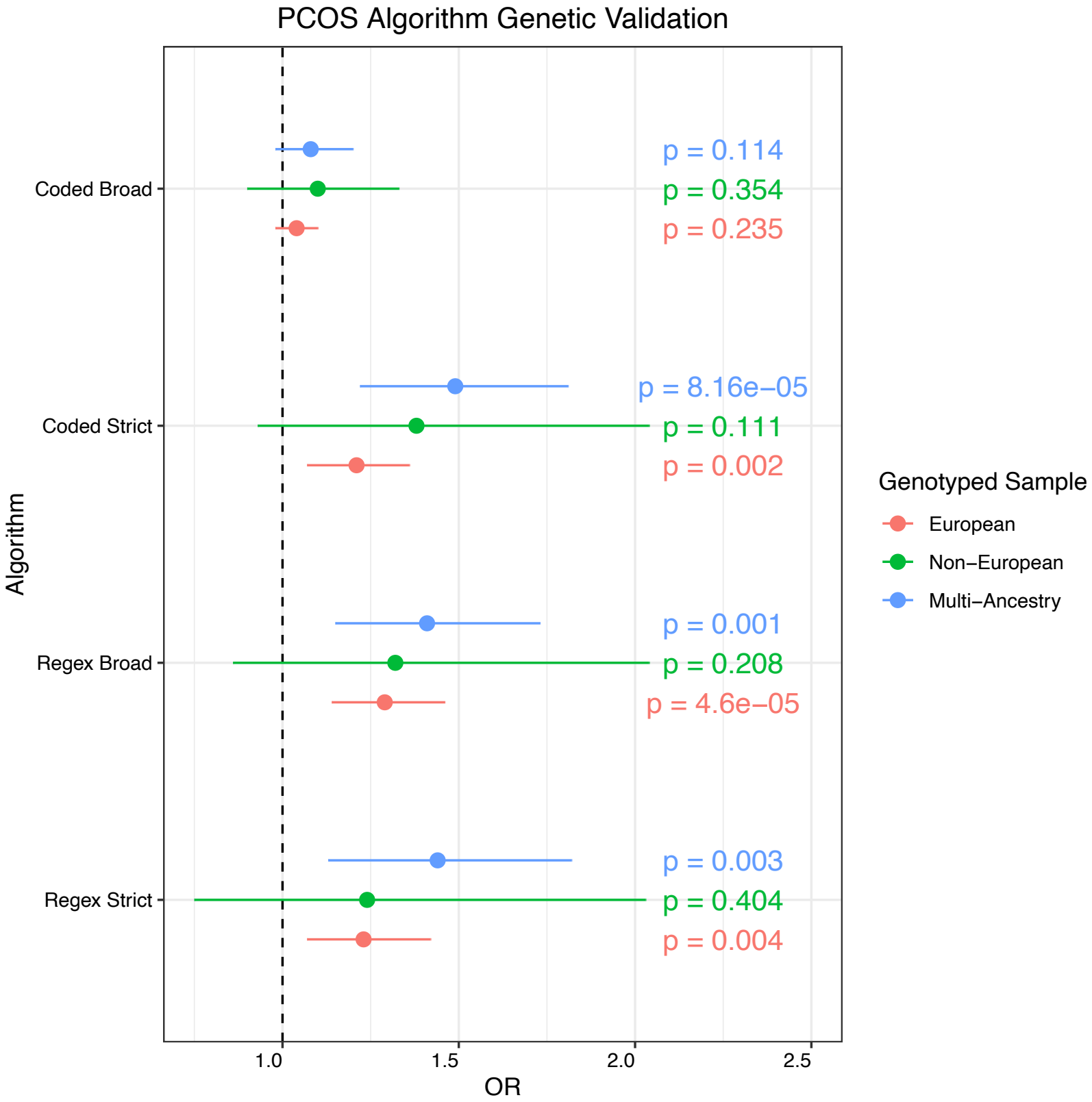
Figure 5. Genetic validation of PCOS algorithms. Logistic regressions were performed between PCOS case status and PCOS polygenic risk scores ($PCOS_{PRS}$). The regression models were adjusted for median age and the first 10 principal components. P-values are displayed for each regression model. Colors correspond to the genotyped sample used in the analysis. OR = odds ratio.