Predicting polycystic ovary syndrome (PCOS) with machine learning algorithms from electronic health records

Zahra Zad, BS[1*]; Victoria S. Jiang, MD[2*]; Amber T. Wolf, BA[3]; Taiyao Wang, PhD[1]; Jay Jojo Cheng, BA[4]; Ioannis Ch. Paschalidis, PhD[1,5]; Shruthi Mahalingaiah, MD, MS[2,6~]

[1]Division of Systems Engineering, Center for Information and Systems Engineering (CISE), Boston University, 15 St. Mary's Street, Brookline, MA 02446, USA

[2]Division of Reproductive Endocrinology and Infertility, Department of Obstetrics and Gynecology, Massachusetts General Hospital, 55 Fruit Street, Yawkey 10, Boston, MA 02114, USA

[3]Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Place, New York, NY 10029, USA

[4] Department of Biostatistics and Medical Informatics, University of Wisconsin, West Johnson Street, Madison, WI 53792, USA

[5]Department of Electrical & Computer Engineering, Department of Biomedical Engineering, and Faculty for Computing & Data Sciences, Boston University, 8 St. Mary's Street, Boston, MA 02215, USA

[6]Department of Environmental Health, Harvard T.H. Chan School of Public Health, 665 Huntington Avenue, Boston, MA 02115, USA

* Zahra Zad and Victoria S. Jiang contributed equally to this manuscript and are co-first authors of this manuscript.

Keywords: Polycystic ovary syndrome (PCOS), disease prediction, predictive analytics, machine learning.

~Corresponding Author & reprint requests:

Shruthi Mahalingaiah, MD, MS. **ORCID: 0000-0002-5527-5787**

Email: shruthi@hsph.harvard.edu

Study funding/competing interest(s): This study was partially supported by National Science Foundation grants CCF-2200052, IIS-1914792, and DMS-1664644, by the NIH under grants R01 GM135930 and UL54 TR004130, and by the Boston University Kilachand Fund for Integrated Life Science and Engineering

Disclosure Summary: The authors declare no conflict of interest and nothing to disclose.

Structured Abstract

Context. Predictive models have been used to aid early diagnosis of PCOS, though existing models are limited to fertility clinic populations.

Objective. Build a predictive model based on an outpatient population at risk for PCOS to facilitate earlier diagnosis and risk prediction.

Design. Retrospective cohort study from a SafetyNet hospital's electronic medical records (EMR) from 2003-2016.

Setting. Hospital outpatient clinics

Patients or Other Participants. 30,601 women aged 18-45 years without concurrent endocrinopathy who had any visit to Boston Medical Center for primary care, obstetrics/gynecology, endocrinology, family medicine, or general internal medicine.

Intervention(s). None

Main Outcome Measure(s). Four prediction outcomes based on Rotterdam criteria using ICD-9 codes for PCOS, irregular menstruation, hyperandrogenism, and PCOM on ultrasound.

Results. We developed predictive models using four machine learning methods: logistic regression, supported vector machine, gradient boosted trees, and random forests. Hormone values (FSH, LH, estradiol, and SHBG) were combined to create a multilayer perceptron score using a neural network classifier. Prediction of PCOS prior to clinical diagnosis in an out-of-sample test set of patients achieved AUC of 85%, 81%, 80%, 82%, respectively in Models I, II, III and IV. Significant positive predictors of PCOS diagnosis across models included hormone levels and obesity; negative predictors included gravidity and positive bHCG.

Conclusions. Among an at-risk population, machine learning algorithms were used to predict PCOS. This approach may guide early detection of PCOS within EMR-interfaced populations to facilitate counseling and interventions that may reduce long-term health consequences, however, additional studies including an entire health system patient population are necessary for model validation.

Introduction

Polycystic ovary syndrome (PCOS) is the most common type of ovulation disorder and endocrinopathy among reproductive age women. PCOS is a diagnosis of exclusion after other endocrinopathies known to affect ovulation have been evaluated including thyroid, adrenal, and pituitary related disease. Based on the Rotterdam criteria, PCOS is diagnosed when two of the three following criteria are exhibited: clinical or biochemical hyperandrogenism, oligo-anovulation, and polycystic ovary morphology (PCOM) on transvaginal or transabdominal ultrasound. PCOS has a population prevalence of 5-15%, depending on the diagnostic criteria used (1).

PCOS is associated with multiple health issues and increased morbidity and mortality, including a high chronic disease burden that is also very costly for individuals with PCOS and insurers (2). PCOS is the leading cause of anovulatory infertility in reproductive-aged women. In fact, over 90% of anovulatory women who present to infertility clinics have PCOS (3). PCOS patients have an increased risk of endometrial hyperplasia and endometrial cancer (4) due to anovulatory cycles leading to long periods of exposure to the effects of unopposed estrogen. PCOS has been associated with the development of metabolic syndrome (5), diabetes (6), cerebrovascular disease and hypertension (7), compared to women without PCOS. Despite these serious health consequences, PCOS frequently goes undiagnosed due to the wide range of symptom severity on presentation, leading to delayed treatment and potentially more severe clinical sequelae due to lack of preventive care, health management, and counseling (4).

Predictive models can play a significant role in aiding earlier diagnosis of PCOS, though several include only those women presenting for fertility care. One model used serum anti-Müllerian hormone (AMH) and androstenedione levels, menstrual cycle length, and BMI to predict the development of PCOS in Chinese women (8). Another model used only AMH and BMI to predict a diagnosis of PCOS or other ovulatory dysfunction disorders (9). Other studies have created predictive models for certain outcomes among women with PCOS such as pregnancy outcomes (10,11) and insulin resistance (12). In this study, we use clinical and socioeconomic predictive variables among 30,601 women aged 18 to 45 years who

had testing/clinical findings that may facilitate a diagnosis of PCOS to develop and evaluate methods for systematic identification of PCOS using existing electronic health records (EHR) data with the ultimate goal of earlier detection and treatment of PCOS for improved long-term health outcomes.

Materials/Methods

Data acquisition

The dataset was created by querying de-identified patient data from female patients aged 18 to 45 years who had or were considered at risk for PCOS diagnosis by having had any one of the three testing procedures for PCOS in their EHR. Included within the initial sample were those patients who had any visit to Boston Medical Center (BMC) for primary care, obstetrics and gynecology, endocrinology, family medicine, or general internal medicine and received: 1) a pelvic/transvaginal ultrasound for any reason, 2) androgen lab assessment, or had clinical symptoms of androgen excess, 3) an ICD-9 label for irregular periods, or 4) a PCOS diagnosis, between October 2003 to December 2016 within the BMC Clinical Data Warehouse (CDW). The start-date was selected to reflect the first day that ICD-9 codes were used and recorded at BMC. The end date reflected cessation of use of the ICD-9 codes and transition to ICD-10 codes within BMC. Exclusion criteria included diagnosis of concurrent endocrinopathy, such as thyroid disorders, hyperaldosteronism, Cushing's syndrome, other adrenal gland disorders, or malignancy based on ICD-9 codes as listed in Supplementary Table 1.

Ethical approval

The study was approved by the Institutional Review Board of Boston University School of Medicine and the Harvard T.H. Chan School of Public Health (Protocol # H35708) and is considered non-human subjects research.

Reference label definitions

Individual predictors

110 Time-varying predictor variables with a date stamp before that of the outcome of interest were
111 included in our models. We considered the following predictor variables:
112 *Socioeconomic and lifestyle demographic variables*: age, race (White/Caucasian, Black/African
113 American, Hispanic/Latina, Asian, Native Hawaiian/Pacific Islander, Middle Eastern, Other/Unknown),
114 smoking status (yes/no), marital status (single, married, separated, divorced, widowed, other),
115 homelessness (yes/no), and highest level of education (8th grade or less, some high school, high school
116 graduate, some college/technical/vocational training, graduated college/technical school/vocational
117 training, declined to answer, other).
118 *Anthropometrics:* Body mass index (BMI, kg/m^2) was either calculated from height and weight or
119 abstracted as the listed BMI variable associated with each visit. BMI was then categorized into three
120 categories: normal (BMI < 25 kg/m^2); overweight (BMI between 25-30 kg/m^2); and obese (BMI > 30
121 kg/m^2). To further capture the obesity population in the absence of height/weight/BMI data, the obese
122 category also included any patient with an ICD-9 code for unspecified obesity (278.00), morbid obesity
123 (278.01), localized adiposity (278.1), and/or a history of gastric bypass.
124 *Cardiovascular health*: To include blood pressure as a predictor variable, we defined a
125 categorical hypertension variable by using systolic (SBP) and diastolic (DBP) blood pressure readings
126 and ICD-9 diagnostic codes for unspecified essential hypertension (401.9), benign essential hypertension
127 (401.1), and essential primary hypertension (401.0). Blood pressure was categorized into three groups:
128 normal, defined by no ICD-9 codes for hypertension recorded and SBP < 120 mmHg, and DBP < 80
129 mmHg; elevated, defined by no ICD-9 codes for hypertension recorded and SBP was 120-129 mmHg or
130 DBP < 80 mmHg; hypertension, defined by any ICD-9 code for hypertension recorded or SBP $\geq$ 140
131 mmHg or DBP $\geq$ 90 mmHg.
132 *Reproductive endocrine predictive variables*: beta human chorionic gonadotropin (bHCG) level
133 (negative bHCG < 5 mIU/mL, positive bHCG $\geq$ 5 mIU/mL), HIV status (negative/positive), age at
134 menarche, pelvic inflammatory disease diagnosis (614.9), history of hysterosalpingogram, and gravidity
135 (history of present or prior pregnancy within obstetric history). Endocrine and metabolic lab values

136 included: TSH, glycosylated hemoglobin (A1c) as a marker for diabetes, low-density lipoprotein (LDL),
137 high density lipoprotein (HDL), and diagnosis of hypercholesterolemia (272.0).
138
139 **Combined predictors**
140 Expecting a nonlinear relationship between many reproductive hormones and a PCOS diagnosis,
141 we used a multilayer perceptron (MLP) neural network to map follicle-stimulating hormone (FSH),
142 luteinizing hormone (LH), sex hormone binding globulin (SHBG), and estradiol (E2) values to a
143 composite metric we call MLP score. The MLP score was repetitively trained and the hyperparameters
144 were tuned to generate a predictive probability associated with PCOS diagnosis for each predictive
145 model, as described with further detail below.
146
147 **Outcomes**
148 *Defining PCOS:* PCOS diagnosis was assigned for any patient who had an ICD-9 code for PCOS
149 (256.4) or met the Rotterdam criteria (13), according to which a positive diagnosis is made in the
150 presence of two out of the following three features: (i) irregular menses (IM) as defined by rare menses,
151 oligo-ovulation, or anovulation; (ii) hyperandrogenism (HA) as defined by clinical or biochemical
152 androgen excess; and (iii) polycystic ovarian morphology (PCOM) noted on transabdominal or
153 transvaginal ultrasound. Based on these three criteria, we defined three auxiliary variables IM, HA, and
154 PCOM to use in the definition of our labels. PCOM was captured through diagnostic radiology text
155 reports from ovarian ultrasound imaging for the subset that had ultrasound imaging (14).
156 *Defining Irregular Menstruation (IM):* IM was defined with the following ICD-9 codes: absence
157 of menstruation (626.0), scanty or infrequent menstruation (626.1), irregular menstrual cycle (626.4),
158 unspecified disorders of menstruation and abnormal bleeding from female genital tract (626.9), and
159 infertility, female associated with anovulation (628.0) (3).
160 *Defining Hyperandrogenism (HA):* HA was assigned to a patient if any of the androgen lab
161 testing for bioavailable testosterone, free testosterone, or total testosterone was greater than clinical

162 thresholds of 11 ng/dL, 5 pg/mL, 45 ng/dL, respectively. In addition, HA was assigned if ICD-9 codes
163 for hirsutism (704.1) or acne (706.1 or 706.0) were recorded for a patient.
164 *Defining Ultrasound characteristics for polycystic ovarian morphology (PCOM):* Among those
165 with an ultrasound in this dataset, PCOM was identified on ultrasound reports using natural language
166 processing (NLP) with complete methods detailed by Cheng and Mahalingaiah (14), to report PCOM as
167 identified (PCOM present), unidentified (PCOM absent), or indeterminate (PCOM unidentifiable based
168 on source report data).
169
170 We considered four models to predict the following: **Model I**: patients with ICD-9 diagnosis of
171 PCOS (256.4) within the EHR; **Model II**: patients diagnosed with PCOS by Rotterdam criteria having IM
172 and HA without a specific ICD-9 PCOS code; **Model III**: patients diagnosed with PCOS by Rotterdam
173 criteria having two out of the three conditions IM/HA/PCOM and without a specific ICD-9 PCOS code;
174 **Model IV**: all patients with PCOS using either Model I or Model III criteria. ICD-9 codes were abstracted
175 from the billing code and diagnosis code associated with each encounter within the EHR. Model I
176 included all patients who were diagnosed with PCOS. Model II and its superset Model III was composed
177 of patients who did not have a PCOS diagnosis code but met diagnostic criteria of PCOS based on
178 Rotterdam criteria, representing the patient population with undiagnosed PCOS. Model IV essentially
179 captures all women who were diagnosed or met criteria for PCOS within our population. Supplementary
180 Table 2 details model definitions and includes the count and percent of patients in each category. The date
181 of diagnosis was assigned by the date of PCOS ICD-9 code (256.4) for Model I, the date of the latest
182 diagnostic criteria met for Model II and III, and the earlier date associated with Model I and Model III, for
183 Model IV.
184
185 **Predictive models**
186 **Classification methods**

We explored a variety of supervised classification methods, both linear and nonlinear. Linear methods included logistic regression (LR) and support vector machines (SVM) (15) and were fitted with an additional regularization term: an L1-norm of the coefficient vector to inject robustness (16) and induce sparsity. Regularization added a penalty to the objective function, thereby minimizing the sum of a metric capturing fitness to the data and a penalty term that is equal to some multiple of a norm of the model parameters. Sparsity was motivated by the earlier works (17–20), where it was shown that sparse classifiers can perform almost as well as very sophisticated classification methods. Nonlinear methods, including gradient boosted trees (GBT/XGBoost) (21) and random forests (RF) (22) which produce large ensembles of decision trees, may yield better classification performance, but are not interpretable or explainable to enable a safety check by a clinician. Specifically, the RF is a large collection of decision trees and it classifies by averaging the decisions of these trees. The GBT/XGBoost, also called gradient boosting machine (GBM), similarly combines decisions by many decision trees. We used LightGBM which is a fast, high-performance GBM framework (23). We tuned GBM's hyperparameters through cross-validation.

Performance metrics

To assess model performance, we obtained the Receiver Operating Characteristic (ROC) curve. The ROC is created by plotting the true positive rate, which is indicative of sensitivity or recall, against the false positive rate (equal to one minus specificity) at various thresholds. The c-statistic or the area under the ROC curve (AUC), is used to evaluate the prediction performance. A perfect predictor is defined by generating an AUC score of 1, and a predictor which makes random guesses has an AUC score of 0.5. We also used the weighted-F1 score to evaluate the models. The weighted-F1 score is the average of the F1 scores of each class weighted by the number of participants in each class. The class-specific F1 scores are computed as the harmonic mean of precision and recall of a classifier which predicts the label of the given class. The weighted-F1 score is between 0 to 1, and a higher value

212 represents a better model. The AUC is more easily interpretable, and the weighted F1-score is more
213 robust to class imbalance (24).
214
215 **Statistical feature selection (SFS)**
216 Categorical variables were converted into dummy/indicator variables. To avoid collinearity, we
217 dropped the missing or unclassified data (NaN) category. For continuous variables, missing values were
218 imputed by the median value for that variable. A summary of the missing variables for each model is
219 provided in Supplementary Table 3. Variables with very low variability (SD<0.0001) were assessed for
220 removal from the models, however none were noted in any model. We applied statistical feature selection
221 (SFS) to reduce the less informative features and simplify the models. For each of the four models'
222 outcomes, the chi-squared test was applied for binary variables and the Kolmogorov-Smirnov statistic for
223 continuous variables; the variables for which we could not reject the null hypothesis of the same
224 distribution for each class (p-value >0.01) were removed. Representative aggregated patient-level
225 statistics for each model are shown in Supplementary Table 4. We also removed one from each pair of
226 highly correlated variables (with absolute value of the correlation coefficient > 0.8) to avoid redundant
227 variables. Highly correlated variables and the retained variable are provided in Supplementary Table 5.
228 For all models we standardized the corresponding features by subtracting the mean and scaling to unit
229 variance.
230
231 **Training-test splitting**
232 We split the dataset into five random parts, where four parts were used as the training set, and the
233 remaining part was used for testing. We used the training set to tune the model hyperparameters via 5-fold
234 cross-validation, and we evaluated the performance metrics on the testing set. We repeated training and
235 testing five times, each time with a different random split into training/test sets. The mean and standard
236 deviation of the metrics on the test sets over the five repetitions are reported.
237

238 **Development of the MLP score**

239 For every model, there was a considerable difference between the AUC of linear models and non-
240 linear models. To improve the performance of our linear models, we utilized nonlinear models to capture
241 intricate relationships between features. We utilized Gradient Boosted Trees (GBT) to find which features
242 most commonly appeared together among decision trees. We found FSH, LH, SHBG, and estradiol levels
243 to be a meaningful group of features which are all reproductive hormones and continuous variables that
244 appeared together among trees for all our models. We subsequently used these four features as input
245 features into a multilayer perceptron (MLP) neural network model with three hidden layers, each
246 employing the rectified linear unit (ReLU) activation function. The neural network was trained using the
247 training set to classify PCOS. We used the output probability of the MLP model, which we called "MLP
248 score," as a new feature into our original predictive models.

249

250 **Recursive feature elimination (RFE)**

251 We also used a recursive feature elimination approach with L1-penalized logistic regression (L1-
252 regularized RFE) to extract the most informative features and develop parsimonious models. Specifically,
253 after running the L1-penalized logistic regression (L1-LR), we obtained weights associated with the
254 variables (i.e., the coefficients of the model, denoted by β), and we eliminated the variable with the
255 smallest absolute weight in each turn. We iterated in this fashion, eliminating one variable at a time, to
256 select a model that maximizes a metric equal to the mean AUC minus the standard deviation (SD) of the
257 AUC in a validation dataset (using 5-fold cross-validation on the training set to obtain an average of this
258 metric over five repetitions).

259

260 **Final predictive models**

261 We computed the performance of the following models: L1-penalized logistic regression (LR-
262 L1), support vector machine (SVM-L1), random forests (RF), and gradient boosted trees
263 (GBT/XGBoost). We calculated each variable's LR coefficient with a 95% confidence interval (β

264 [95%CI]), the correlation of the variable with the outcome (Y-correlation), the p-value of each variable
265 (p-value), the mean of the variable (Y1-mean) in the PCOS labeled patients, the mean of the variable (Y0-
266 mean) in the patients without the PCOS label, and the mean and standard deviation of the variable over all
267 patients (All-mean and All-SD). Ranking predictor variables by the absolute value of their coefficients in
268 the logistic regression model amounts to ranking these variables by how much they affect the predicted
269 probability of the outcome. A positive coefficient implies that the larger the value of the variable within
270 the range specified by the data, the higher the chance of having a PCOS diagnosis as defined by the model
271 outcome.
272
273 **Results**
274 **Results of data acquisition and data pre-processing**
275 After inclusion and exclusion criteria were applied to all 65,431 women within the initial data
276 pool, 30,601 patient records were available for this analysis and defined populations are included in
277 Figure 1. There were 1,329 patients (4.5%) with a PCOS ICD-9 diagnosis code (Model I). 1,465 patients
278 had records with PCOM results as present, absent, or unidentifiable. There were 1,056 patients (3.6%)
279 with undiagnosed PCOS (Model II), and a total of 1,116 (3.8%) of patients with no ICD 256.4 indication
280 and two out of IM/HA/PCOM positive criteria (Model III). Finally, there were 2,445 PCOS patients
281 (8.0%) in the combined analysis (Model IV). The total number of records in each model are included in
282 Supplementary Table 2. In the total cohort, the patients were predominantly Black/African American
283 (40.3%) and White (26.5%), with an average age of 33.6 years (SD = 6.6). Complete demographic
284 characteristics are described in Table 1.
285 There were 43 categorical variables and 12 continuous variables retained as predictors after the
286 data pre-processing procedures. There were four pairs of highly correlated variables and one variable
287 from each correlated pair included in the final model as noted in Supplemental Table 5. Supplementary
288 Table 4 describes all 51 variables used by the predictive models.
289

290 **Model Performance**

291 Tables 2, 3, 4 and 5 display the parsimonious models that use the MLP score (LR-L2-MLP score)
292 and show the most significant variables in the prediction of the outcome for Models I, II, III, and IV,
293 respectively. All p-values were less than 0.05, which was set as the significance level.

294 For Model I, the parsimonious predictive model achieved an AUC (SD) of 82.3% (1.7). The MLP
295 score (β = 0.71) and obesity (β = 0.45) were positively correlated with PCOS diagnosis. Pregnancy
296 (gravidity β = -0.53; positive pregnancy test β = -0.50), normal BMI (β = -0.24), smoking (β = -0.18), age
297 (β = -0.16), and Hispanic race (β = -0.10) were inversely correlated with PCOS diagnosis as shown in
298 Table 2.

299 For Model II, the parsimonious predictive model achieved an AUC (SD) of 77.6% (1.3). The
300 MLP score (β = 0.61), obesity (β = 0.21), normal BMI (β = 0.15), normal blood pressure (β = 0.16),
301 negative pregnancy test (β = 0.12), and normal HDL (β = 0.08) were positively correlated with
302 undiagnosed PCOS. Age (β = -0.27), pregnancy (gravidity β = -0.26; positive pregnancy test β = -0.19),
303 and Hispanic race (β = -0.18) were inversely correlated with undiagnosed PCOS as show in Table 3.

304 For Model III, the parsimonious predictive model achieved an AUC (SD) of 77.4% (1.6). The
305 MLP score (β = 0.60), obesity (β = 0.19), normal blood pressure (β = 0.17), normal BMI (β = 0.14), Black
306 race (0.13), negative pregnancy test (β = 0.12), and normal HDL (β = 0.09) were positively correlated
307 with undiagnosed PCOS. Age (β = -0.25), pregnancy (gravidity β = -0.24; positive pregnancy test β = -
308 0.20), and Hispanic race (β = -0.15) were inversely correlated with undiagnosed PCOS as show in Table
309 4.

310 For Model IV, the parsimonious predictive model achieved an AUC (SD) of 79.1% (1.1). The
311 MLP score (β = 0.7), obesity (β = 0.31), normal BMI (β = 0.15), hypertension (β = 0.07) and some higher
312 degree of education, such as college or vocational/technical school (β = 0.06) were positively correlated
313 with PCOS diagnosis. Age (β = -0.21), pregnancy (gravidity β = -0.37; positive pregnancy test β = -0.34;
314 negative pregnancy test β = -0.05), Hispanic race (β = -0.12), and smoking (β = -0.08) were inversely
315 correlated with PCOS diagnosis as shown in Table 5.

316 GBT models had the highest performance. Predictions of PCOS in a test set of patients not used
317 during algorithm training achieved 85%, 81%, 80%, and 82% AUC for Models I, II, III, and IV,
318 respectively. We also report the performance with the logistic regression model (LR-L1) after SFS and
319 the performance when using our developed MLP score alongside variables selected via recursive feature
320 elimination (LR-L2-MLP score). Supplementary Table 6 displays features for each model, associated
321 with LR-L1 algorithm after SFS. As we hypothesized, developing models using the MLP score (LR-L2-
322 MLP score) leads to improvement of the performance of linear models (LR-L1) for Models I, II, III, and
323 IV, respectively from 79%, 72%, 73%, and 75% AUC to 82%, 78%, 77%, and 79% AUC. Table 6 details
324 the models with the best performance (highest AUC) using all 51 features before and after statistical
325 feature selection (SFS). In Table 6, the means and standard deviations of AUC and weighted-F1 scores on
326 the test set over the five repetitions are listed. Supplementary Table 7 displays the performance of all
327 models and all algorithms, before and after statistical feature selection (SFS).
328
329 **Discussion**
330 Evaluating an at-risk population for PCOS is essential for early diagnosis and initiating multi-
331 disciplinary care with the goal of reducing health risks (endometrial hyperplasia/cancer), infertility and
332 pregnancy complications, and chronic disease burden including cardiometabolic disorders associated with
333 PCOS. Retrospective analysis of the at-risk population within an urban health center allows for
334 assessment of factors predictive of diagnosis. Of note, the study sample represents a population of
335 patients who had any visit to BMC for primary care, obstetrics and gynecology, endocrinology, family
336 medicine, or general internal medicine and does not represent a random sample. While this is not a
337 population level assessment, our model is applicable to patients with high suspicion for PCOS who
338 interact with the healthcare system.
339 The ranked list of variables, from the most predictive to the least predictive of the PCOS
340 outcome, informed the main drivers of the predictive models. For example, non-gravidity, high levels of
341 LH, low levels of FSH, obesity, and higher BMI increase the likelihood of PCOS. These variables are

342 consistent with key variables from other models and in the pathophysiology of PCOS. The overall
343 predictive accuracy was high for all models, suggesting that a predictive model may assist in early
344 detection of PCOS within those at risk in an electronically interfaced medical record. Furthermore, we
345 found that non-linear models had superior predictive capacity compared to linear models for all four
346 model outcomes, potentially allowing for inclusion of non-linear reproductive hormone relationships.
347 When assessing patients who received a diagnosis of PCOS (Model I), the most predictive factors
348 related to diagnosis were hormone levels (as captured by the MLP score) and obesity, a clinical factor in
349 supporting a PCOS diagnosis. Specifically, there is a non-linear relationship between reproductive
350 hormones such as FSH, LH, and estradiol. Often these hormonal lab tests are obtained randomly in those
351 with oligomenorrhea, and it is also common to find an elevated FSH to LH ratio. A concern may also be
352 the misclassification of hypothalamic amenorrhea into the group classified as PCOS where the FSH and
353 LH levels would be low or suppressed, or in the setting of premature ovarian insufficiency, notable by an
354 elevated FSH and low estradiol. The MLP score allows for the diversity of relationships of these hormone
355 levels and was trained using a neural network to appropriately classify PCOS. Additionally, prior
356 pregnancy (gravidity) and a positive pregnancy test were negatively associated with a diagnosis of PCOS,
357 consistent with the underlying increased risk of infertility due to oligo-ovulation which many patients
358 with PCOS experience. Additionally, normal BMI and smoking, a known ovarian toxicant, were
359 negatively associated with the presence of a PCOS diagnosis, which may indicate patient characteristics
360 that increase risk of a delayed PCOS diagnosis. These identified predictor variables demonstrate the
361 robustness of the model towards predicting phenotypic traits of patients with PCOS, which is aligned with
362 the performance accuracy. While the significant factors such as hormone levels, gravidity, bHCG, and
363 obesity identified in the model are already known to be associated with PCOS, the true impact of our
364 model lies within the implementation of such a tool within the EMR to alert clinicians to the diagnosis
365 and subsequently allow patients to access necessary healthcare to either prevent or manage co-morbidities
366 related to PCOS.

367 When assessing patients who met diagnostic criteria without the ICD-9 label of PCOS (Models II
368 and III), predictive factors both supported the underlying PCOS diagnosis and alluded towards factors
369 that may contribute to missing the diagnosis despite meeting Rotterdam criteria. Similar to Model I,
370 gravidity and a positive pregnancy test were negatively associated with Models II and III diagnosis, while
371 obesity was positively associated with Models II and III diagnosis, consistent with the PCOS diagnosis
372 within Model I. Interestingly, distinct positive predictors among Models II and III were normal BMI,
373 normal blood pressure, and normal HDL. These patients may present as the "lean" phenotype of PCOS or
374 those with mild features, leading to underdiagnosis of PCOS. Diagnosing "lean" PCOS can be more
375 nuanced, potentially delaying diagnosis or requiring more specialized consultation (25). By evaluating
376 key factors among the underdiagnosed population, this predictive model can serve as a safety net among
377 hospital systems to aid in timely diagnosis and identification of diagnostic barriers through EHR data
378 analytics. Addressing diagnostic gaps such as PCOS underdiagnosis among the "lean" population can
379 lead to more timely diagnosis and downstream follow-up for known clinical sequelae associated with
380 PCOS.

381 Consistent with Models I, II, and III, positive pregnancy test and gravidity were both negatively
382 associated with PCOS diagnosis in Model IV while obesity and presence of hypertension were both
383 positively associated with the Model IV combined PCOS outcome. Some higher degree of education,
384 such as college or vocational/technical school, was also positively associated with the outcomes of
385 undiagnosed PCOS and combined PCOS (Models II, III, and IV), which may suggest that education
386 status and patient's self-advocacy for seeking care within a medical system may be implicated specifically
387 in under-diagnosed individuals. Of note, we dropped insurance status after finding that the null was a
388 strong predictor of PCOS, though it is interesting to note that 83% of 331 patients in this dataset with
389 missing insurance have PCOS. Insurance status alludes to socioeconomic barriers such as financial access
390 to care, which can result in a delay in timely diagnosis through either inability to seek evaluation or
391 follow through with testing. Persistence in seeking treatment within a fractionated health care system can

392 be challenging financially and psychologically, as patients may need multiple evaluation or specialist's
393 consultation to reach the right diagnosis.

394 Other predictive models have been created for earlier detection of PCOS as well as for predicting
395 long-term health outcomes among women with a diagnosis of PCOS. One such model was created from
396 11,720 ovarian stimulation cycles at Peking University Third Hospital. The model used serum
397 antimullerian hormone (AMH) and androstenedione levels, BMI, and menstrual cycle length to predict a
398 diagnosis of PCOS. The algorithm was then developed into an online platform that is able to calculate
399 one's risk of PCOS given certain indicators that are inputted into the model, allowing for better screening
400 abilities in the clinic (26). Another study created a similar model, taking into account AMH and BMI to
401 predict a diagnosis of PCOS or other ovulatory dysfunction disorders among 2,322 women (9). They
402 found that in women with higher BMIs and lower AMH levels could be used to predict PCOS compared
403 to normal-weight or underweight women. Deshmukh et al. created a simple four-variable model which
404 included free androgen index (FAI), 17-hydroxyprogesterone, AMH, and waist circumference for
405 predicting risk of PCOS in a cross-sectional study involving 111 women with PCOS and 67 women
406 without PCOS (27). Lastly, Joo et al. used polygenic and phenotypic risk scores to develop a PCOS risk
407 prediction algorithm (28). They found high degrees of association between PCOS and various metabolic
408 and endocrine disorders including obesity, type 2 diabetes, hypercholesterolemia, disorders of lipid
409 metabolism, hypertension, and sleep apnea (28). Our model is unique because it includes a general
410 hospital-sourced population with a heterogenous population as opposed to solely in the context of fertility
411 care. Furthermore, the parsimonious and interpretable models were very close in achieving full model
412 predictive accuracy, performing also relatively closely to the best-performing non-linear models.
413 Essentially, our parsimonious models "isolate" nonlinearities in hormone levels (captured by the MLP
414 score) and linearly combine that score with other variables. Our variables are easily accessible in an
415 electronic health dataset, rendering the models helpful for clinical prediction. However, our study did not
416 evaluate AMH as a predictive variable because it was not widely utilized during the time window of this
417 data extraction corresponding with ICD-9 codes.

418 In addition to the goal of improved screening for PCOS, other models have been created to
419 predict long-term clinical outcomes in women with PCOS, such as ovulation, conception, and live birth
420 (10,11). Given the increased risk of insulin resistance in women with PCOS, Gennarelli et al. created a
421 mathematical model to predict insulin sensitivity based on variables such as BMI, waist and hip
422 circumferences, truncal-abdominal skin folds, and serum concentrations of androgens, SHBG,
423 triglycerides, and cholesterol (12). Models to predict non-alcoholic fatty liver disease (NAFLD) risk
424 among young adults with PCOS have also been generated (29). Combining earlier detection with more
425 accurate risk stratification of clinical sequalae through predictive modeling can significantly improve the
426 long-term health outcomes of women with PCOS. Application of our models to predict other down-
427 stream health risks after the diagnosis of PCOS is a future area of research.
428 Beyond the long-term health impacts of PCOS, the condition also carries a significant economic
429 cost for our healthcare system. A study by Riestenberg et al (2022) recently estimated the total economic
430 burden of PCOS, as well as the cost specifically for pregnancy-related complications and long-term health
431 morbidities (2). The authors estimated the annual economic burden of PCOS to be $8 billion as of 2020 in
432 the United States. Furthermore, the excess cost of pregnancy-related comorbidities such as gestational
433 hypertension, gestational diabetes, and preeclampsia attributable to PCOS totals $375 million USD
434 annually. Outside of pregnancy, the cost of long-term comorbidities associated with PCOS including
435 stroke and type 2 diabetes mellitus was estimated at $3.9 billion USD. Meanwhile, the cost for diagnostic
436 evaluation of PCOS was less than 2% of the total economic burden. This estimated financial burden
437 suggests that predictive models aiding earlier diagnosis could not only reduce long-term health
438 consequences of PCOS but also alleviate significant healthcare costs associated with the condition.
439 Strengths of our model include the large sample size of patients used to generate these predictive
440 models which are highly accurate and robust in PCOS diagnosis prediction and serve as one of the largest
441 to date (30). Although not included, we also assessed a multi-imputation method to see if multi-imputing
442 missing values led to improved accuracy. This method produces similar results, with linear models
443 becoming slightly worse and non-linear models slightly improving. Therefore, since there was no

444 considerable improvement in models' performance and, more importantly, multi-imputation ultimately
445 led to loss of interpretability of the data, it was not justifiable to use multi-imputation in our models.
446 Additionally, utilizing the MLP score in this novel application allowed for non-linear
447 interpretation of the variables. Most models evaluate reproductive hormones as individual variables
448 within linear models, which does not account for the high inter- and intra-patient variability. By using
449 non-linear mapping of the hormone values, we were able to generate a composite variable allowing for a
450 linear function that correlates with the likelihood of an accurate prediction. While using the MLP score
451 and incorporating the non-linear relationship between reproductive hormones (FSH, estradiol, LH, and
452 SHBG) allowed for improved accuracy of each of the models, it can be somewhat cumbersome to
453 implement in other populations and does not quantify the true correlative relationship between each
454 individual variable comprising the MLP score.
455 Of note, our models were generated using a population who was at risk for PCOS diagnosis based
456 on clinical parameters, rather than the entire patient population at BMC. The purpose of this study was to
457 evaluate the at-risk population and not delineate the presence of PCOS in the population at large, or the
458 entire population presenting for care—as we would need to perform an evaluation on an unselected
459 population to determine the prevalence of PCOS (31). Utilizing this at risk PCOS population within a
460 large urban health center allowed us to focus on identifying gaps in diagnosis. Within our at-risk cohort,
461 1,116 individuals were identified by the model without the ICD-9 code that met Rotterdam PCOS
462 diagnostic criteria (Model III), suggesting the predictive value of our models to identify at risk groups
463 within a large health system and reduce delays in diagnosis. A random population of patients would have
464 lowered or potentially eliminated any modeling signal and potentially lead to minimal association
465 indicators for PCOS, significantly limiting predictive power. Clinically, to meet Rotterdam criteria for
466 PCOS diagnosis, patients would need to undergo clinical work-up including clinical evaluation for
467 concurrent endocrinopathy, amenorrhea evaluation, androgen testing, and ultrasound, without which
468 diagnosis would not be assigned. If used clinically, this model would need these key pieces of clinical and

469 laboratory information to accurately assess patient's PCOS risk, thus justifying our utilization of an at-
470 risk population for model training purposes.

471 Informative presence in EHR data is defined as data that is present and informed with respect to
472 the health outcome, in this case PCOS, as well as behavioral patterns of interaction with healthcare
473 institutions which may be additionally impacted by marginalization (32). This is an important
474 consideration for interpreting predictive models using EHR data (32,33). Even among those evaluated, we
475 were able to extract over 1000 patients who were undiagnosed with PCOS among the population,
476 suggesting the predictive value of the modelling in identifying diagnosis gaps even among small
477 populations within a large health system. Additional studies need to be conducted in other patient
478 populations or unselected community-based populations to validate the use of these models, especially
479 expanding to the entire population within a health system to evaluate the accuracy of our models. This
480 novel model is the first use of a machine learning algorithm on a large EHR dataset to assess PCOS risk
481 and to guide early detection and interventions of PCOS, pregnancy planning, and pre-pregnancy
482 interventions. If the predicted probability of PCOS is high, facilitating an earlier diagnosis, this may
483 support earlier access to specialized physicians for both fertility care/family planning along with
484 preventative care for the cardiovascular and gynecological sequelae of longstanding PCOS.

Acknowledgements

We would like to acknowledge Linda Rosen, the research manager of the Clinical Data Warehouse at the Boston Medical Center for procuring the dataset, and Alexis Veiga, the research assistant who provided administrative support for this project.

Author contributions

ZZ performed the analysis and co-wrote the manuscript with a focus on the methods. VJ interpreted the findings and drafted the initial manuscript. AW conducted a literature review, contributed to interpretation of data, writing, and editing the manuscript. TW initiated the analytical approach to the research question. JC curated the initial dataset and reviewed the analysis and manuscript drafts. SM and ICP designed the study, oversaw analysis, interpretation of the findings, and manuscript drafting and revision process. All authors met ICJME criteria for authorship.

Data availability

All datasets generated during and/or analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request.

Competing interests

The authors declare no competing interests.

504 **References**

505 1. Azziz R, Carmina E, Dewailly D, Diamanti-Kandarakis E, Escobar-Morreale HF, Futterweit W, et al.
506 The Androgen Excess and PCOS Society criteria for the polycystic ovary syndrome: the complete task
507 force report. Fertil Steril. 2009 Feb;91(2):456–88.

508 2. Riestenberg C, Jagasia A, Markovic D, Buyalos RP, Azziz R. Health Care-Related Economic Burden of
509 Polycystic Ovary Syndrome in the United States: Pregnancy-Related and Long-Term Health
510 Consequences. J Clin Endocrinol Metab. 2022 Jan 18;107(2):575–85.

511 3. Sirmans SM, Pate KA. Epidemiology, diagnosis, and management of polycystic ovary syndrome. Clin
512 Epidemiol. 2013 Dec 18;6:1–13.

513 4. Barry JA, Azizia MM, Hardiman PJ. Risk of endometrial, ovarian and breast cancer in women with
514 polycystic ovary syndrome: a systematic review and meta-analysis. Hum Reprod Update.
515 2014;20(5):748–58.

516 5. Lim SS, Kakoly NS, Tan JWJ, Fitzgerald G, Bahri Khomami M, Joham AE, et al. Metabolic syndrome
517 in polycystic ovary syndrome: a systematic review, meta-analysis and meta-regression. Obes Rev.
518 2019;20(2):339–52.

519 6. Anagnostis P, Tarlatzis BC, Kauffman RP. Polycystic ovarian syndrome (PCOS): Long-term metabolic
520 consequences. Metabolism. 2018;86:33–43.

521 7. Wekker V, Van Dammen L, Koning A, Heida KY, Painter RC, Limpens J, et al. Long-term
522 cardiometabolic disease risk in women with PCOS: a systematic review and meta-analysis. Hum
523 Reprod Update. 2020;26(6):942–60.

524 8. Xu H, Feng G, Alpadi K, Han Y, Yang R, Chen L, et al. A Model for Predicting Polycystic Ovary
525 Syndrome Using Serum AMH, Menstrual Cycle Length, Body Mass Index and Serum
526 Androstenedione in Chinese Reproductive Aged Population: A Retrospective Cohort Study. Front
527 Endocrinol. 2022;13:821368.

528 9. Vagios S, James KE, Sacha CR, Hsu JY, Dimitriadis I, Bormann CL, et al. A patient-specific model
529 combining antimüllerian hormone and body mass index as a predictor of polycystic ovary syndrome
530 and other oligo-anovulation disorders. Fertil Steril. 2021;115(1):229–37.

531 10. Kuang H, Jin S, Hansen KR, Diamond MP, Coutifaris C, Casson P, et al. Identification and replication
532 of prediction models for ovulation, pregnancy and live birth in infertile women with polycystic ovary
533 syndrome. Hum Reprod. 2015;30(9):2222–33.

534 11. Jiang X, Liu R, Liao T, He Y, Li C, Guo P, et al. A Predictive Model of Live Birth Based on Obesity
535 and Metabolic Parameters in Patients With PCOS Undergoing Frozen-Thawed Embryo Transfer. Front
536 Endocrinol. 2021;12.

537 12. Gennarelli G, Holte J, Berglund L, Berne C, Massobrio M, Lithell H. Prediction models for insulin
538 resistance in the polycystic ovary syndrome. Hum Reprod. 2000;15(10):2098–102.

539 13. ESHRE TR, Group ASPCW. Revised 2003 consensus on diagnostic criteria and long-term health risks
540 related to polycystic ovary syndrome. Fertil Steril. 2004;81(1):19–25.

14. Cheng JJ, Mahalingaiah S. Data mining polycystic ovary morphology in electronic medical record ultrasound reports. Fertil Res Pract. 2019;5(1):1–7.

15. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction. Vol. 1. Springer series in statistics Springer, Berlin; 2001.

16. Chen R, Paschalidis IC. Distributionally Robust Learning. Found Trends® Optim. 2020 Dec 22;4(1–2):1–243.

17. Brisimi TS, Xu T, Wang T, Dai W, Adams WG, Paschalidis IC. Predicting Chronic Disease Hospitalizations from Electronic Health Records: An Interpretable Classification Approach. Proc IEEE. 2018 Apr;106(4):690–707.

18. Brisimi TS, Xu T, Wang T, Dai W, Paschalidis IC. Predicting diabetes-related hospitalizations based on electronic health records. Stat Methods Med Res. 2019 Dec 1;28(12):3667–82.

19. Chen R, Paschalidis ICh. Robust Grouped Variable Selection Using Distributionally Robust Optimization. J Optim Theory Appl. 2022 Sep 1;194(3):1042–71.

20. Chen R, Paschalidis IC, Hatabu H, Valtchinov VI, Siegelman J. Detection of unwarranted CT radiation exposure from patient and imaging protocol meta-data using regularized regression. Eur J Radiol Open. 2019;6:206–11.

21. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. San Francisco, California, USA: Association for Computing Machinery; 2016 [cited 2020 Jun 19]. p. 785–94. (KDD '16). Available from: https://doi.org/10.1145/2939672.2939785

22. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.

23. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: A highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017;30.

24. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. Brock G, editor. PLOS ONE. 2015 Mar 4;10(3):e0118432.

25. Toosy S, Sodi R, Pappachan JM. Lean polycystic ovary syndrome (PCOS): an evidence-based practical approach. J Diabetes Metab Disord. 2018 Nov 13;17(2):277–85.

26. Xu H, Feng G, Alpadi K, Han Y, Yang R, Chen L, et al. A Model for Predicting Polycystic Ovary Syndrome Using Serum AMH, Menstrual Cycle Length, Body Mass Index and Serum Androstenedione in Chinese Reproductive Aged Population: A Retrospective Cohort Study. Front Endocrinol. 2022;13.

27. Deshmukh H, Papageorgiou M, Kilpatrick ES, Atkin SL, Sathyapalan T. Development of a novel risk prediction and risk stratification score for polycystic ovary syndrome. Clin Endocrinol (Oxf). 2019;90(1):162–9.

576 28. Joo YY, Actkins K, Pacheco JA, Basile AO, Carroll R, Crosslin DR, et al. A polygenic and phenotypic
577 risk prediction for polycystic ovary syndrome evaluated by phenome-wide association studies. J Clin
578 Endocrinol Metab. 2020;105(6):1918–36.

579 29. Carreau AM, Pyle L, Garcia-Reyes Y, Rahat H, Vigers T, Jensen T, et al. Clinical prediction score of
580 nonalcoholic fatty liver disease in adolescent girls with polycystic ovary syndrome (PCOS-HS index).
581 Clin Endocrinol (Oxf). 2019;91(4):544–52.

582 30. Shekhar S, Barrera FJ, Brown E, Rojo A, Obeso J, Plata H, et al. Abstract #1413640: Artificial
583 Intelligence and Machine Learning in the Detection of Polycystic Ovarian Syndrome: A Systematic
584 Review and Meta-Analysis. Endocr Pract. 2023 May 1;29(5):S88–9.

585 31. Azziz R, Woods KS, Reyna R, Key TJ, Knochenhauer ES, Yildiz BO. The Prevalence and Features of
586 the Polycystic Ovary Syndrome in an Unselected Population. J Clin Endocrinol Metab. 2004 Jun
587 1;89(6):2745–9.

588 32. Harton J, Mitra N, Hubbard RA. Informative presence bias in analyses of electronic health records-
589 derived data: a cautionary note. J Am Med Inform Assoc JAMIA. 2022 Jun 14;29(7):1191–9.

590 33. Sisk R, Lin L, Sperrin M, Barrett JK, Tom B, Diaz-Ordaz K, et al. Informative presence and observation
591 in routine health data: A review of methodology for clinical risk prediction. J Am Med Inform Assoc
592 JAMIA. 2020 Nov 9;28(1):155–66.

593

Figure 1. Flow of patients from the BMC CDW into the dataset used by the study.

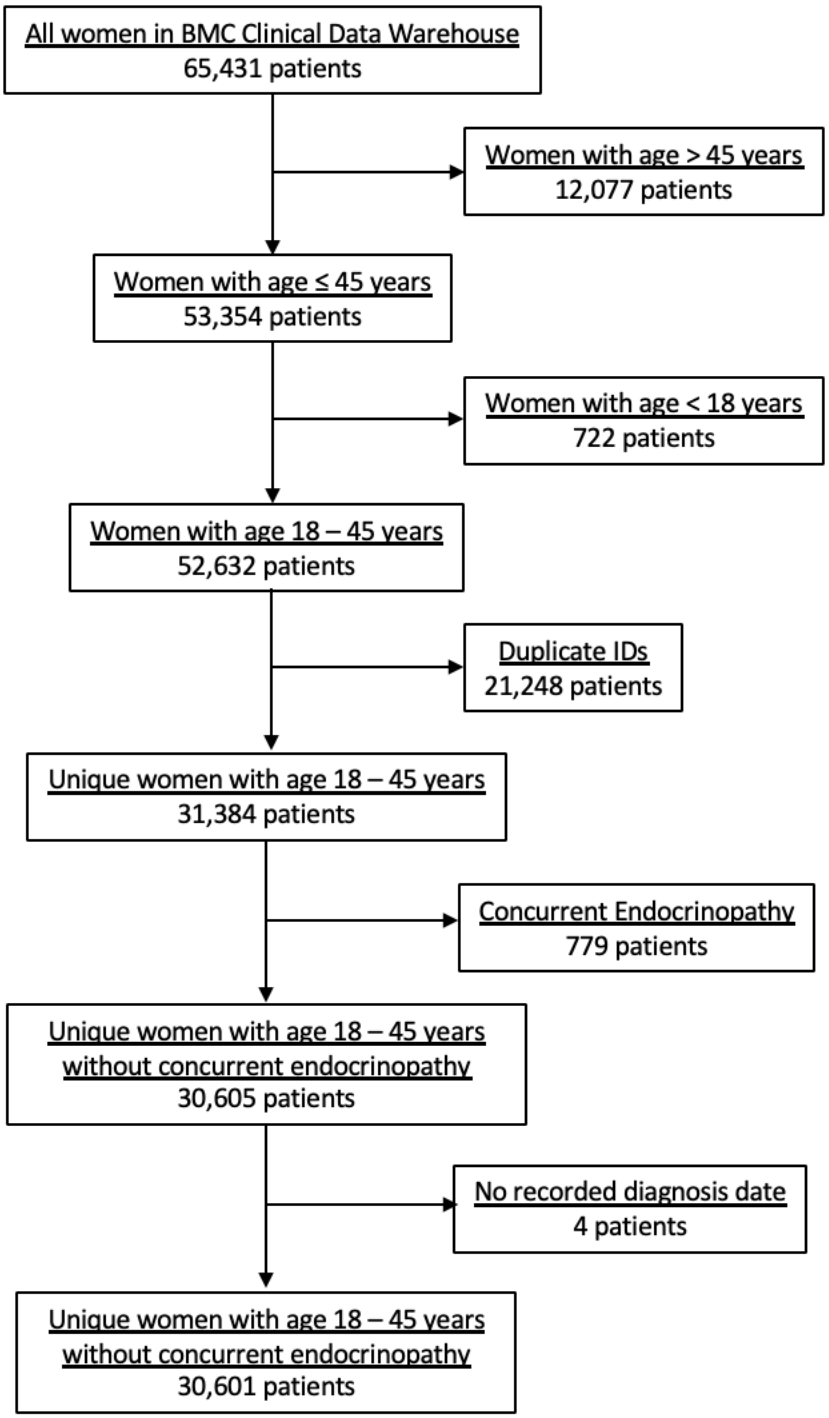

Table 1. Demographic characteristics of the study population and by model.

Variable	**Model I**	**Model II**	**Model III**	**Model IV**
Age, Mean years (SD)	33.6 (6.6)	33.7 (6.6)	33.7 (6.6)	33.6 (6.6)
Race, n (%)				
Black/African American	11881 (40.3)	11824 (40.5)	11861 (40.5)	12395 (40.5)
White/Caucasian	7812 (26.5)	7733 (26.5)	7741 (26.4)	8086 (26.4)
Hispanic/Latina	2858 (9.7)	2837 (9.7)	2841 (9.7)	2929 (9.6)
Asian	1350 (4.6)	1354 (4.6)	1354 (4.6)	1406 (4.6)
Middle Eastern	175 (0.6)	176 (0.6)	176 (0.6)	184 (0.6)
American Indian/Native American	163 (0.6)	162 (0.6)	162 (0.6)	168 (0.5)
Native Hawaiian/Pacific Islander	17 (0.1)	18 (0.1)	18 (0.1)	18 (0.1)
Other	979 (3.3)	966 (3.3)	966 (3.3)	1023 (3.3)
Unknown	4250 (14.41)	4146 (14.19)	4153 (14.19)	4392 (14.4)
Marital Status				
Single	22325 (75.7)	22155 (75.8)	22199 (75.8)	23224 (75.9)
Married	5833 (19.8)	5753 (19.7)	5767 (19.7)	6018 (19.7)
Separated	392 (1.3)	391 (1.3)	392 (1.3)	401 (1.3)
Divorced	388 (1.3)	379 (1.3)	380 (1.3)	397 (1.3)
Widowed	35 (0.1)	35 (0.1)	35 (0.1)	35 (0.1)
Other	502 (1.7)	489 (1.7)	489 (1.7)	516 (1.7)
Unknown	10 (0.03)	10 (0.03)	10 (0.03)	10 (0.03)
Body Mass Index				
Normal (BMI < 25)	7534 (25.6)	7685 (26.3)	7697 (26.3)	7902 (25.8)
Overweight (BMI between 25-30)	5694 (19.3)	5689 (19.5)	5707 (19.5)	5941 (19.4)
Obese (BMI ≥ 30)	7645 (25.9)	7369 (25.2)	7387 (25.2)	7985 (26.1)
Unknown	8612 (29.2)	8469 (29.0)	8481 (29.0)	8,773 (28.7)

Table 2. Most significant variables for PCOS diagnosis prediction in Model I.

Rank	**Variables**	$\boldsymbol{\beta}$	$\boldsymbol{\beta}$ **− %95 CI**	**Y-correlation**	**p-value**	**Y1-mean**	**Y0-mean**	**All-mean**	**All-std**
0	MLP Score	0.71	0.028	0.33	6.80E-197	0.17	0.04	0.05	0.08
1	Intercept	-0.68	-	-	-	-	-	-	-
2	Gravidity	-0.53	0.018	-0.12	4.55E-78	1.28	2.08	2.04	1.39

3	Positive bHCG	-0.5	0.019	-0.09	1.50E-48	0.05	0.23	0.22	0.42
4	Obesity	0.45	0.017	0.11	1.38E-81	0.51	0.27	0.28	0.45
5	Normal BMI	-0.24	0.017	-0.05	3.57E-16	0.15	0.26	0.26	0.44
6	Smoker	-0.18	0.017	-0.03	6.62E-05	0.09	0.14	0.14	0.34
7	Age	-0.16	0.016	-0.08	1.70E-25	31.34	33.79	33.68	6.61
8	Hispanic/Latina Race	-0.1	0.016	-0.02	1.82E-03	0.07	0.10	0.10	0.30

Table 3. Most significant variables for PCOS diagnosis prediction in Model II.

Rank	**Variables**	***β***	***β* - %95 CI**	**Y-correlation**	**p-value**	**Y1-mean**	**Y0-mean**	**All-mean**	**All-std**
0	MLP Score	0.61	0.023	0.26	2.13E-142	0.12	0.04	0.04	0.06
1	Intercept	-0.44	-	-	-	-	-	-	-
2	Age	-0.27	0.015	-0.08	2.26E-31	31.01	33.79	33.69	6.61
3	Gravidity	-0.26	0.016	-0.09	2.35E-63	1.42	2.08	2.06	1.39
4	Obesity	0.21	0.016	0.03	9.60E-06	0.34	0.27	0.27	0.44
5	Positive bHCG	-0.19	0.017	-0.06	4.14E-21	0.10	0.23	0.23	0.42
6	Hispanic/Latina Race	-0.18	0.016	-0.02	2.69E-03	0.06	0.10	0.10	0.30
7	Normal BP	0.16	0.015	0.03	1.37E-07	0.60	0.51	0.51	0.50
8	Normal BMI	0.15	0.016	0.03	8.57E-07	0.34	0.26	0.26	0.44
9	Negative bHCG	0.12	0.015	0.06	1.44E-22	0.37	0.23	0.23	0.42
10	HDL	0.08	0.015	0.01	1.03E-10	52.13	51.59	51.61	7.86

Table 4. Most significant variables for PCOS diagnosis prediction in Model III.

Rank	**Variables**	***β***	***β* - %95 CI**	**Y-correlation**	**p-value**	**Y1-mean**	**Y0-mean**	**All-mean**	**All-std**
0	MLP Score	0.6	0.023	0.26	7.41E-142	0.10	0.04	0.04	0.05
1	Intercept	-	-	-	-	-	-	-	-
2	Age	-0.25	0.015	-0.08	5.91E-30	31.16	33.79	33.69	6.61
3	Gravidity	-0.24	0.016	-0.09	2.47E-63	1.46	2.08	2.06	1.39
4	Positive bHCG	-0.20	0.017	-0.06	3.59E-20	0.11	0.23	0.23	0.42
5	Obesity	0.19	0.016	0.03	2.73E-06	0.34	0.27	0.27	0.44
6	Normal BP	0.17	0.015	0.04	3.94E-08	0.60	0.51	0.51	0.50

7	Hispanic/Latina Race	-0.15	0.016	-0.02	2.00E-03	0.06	0.10	0.10	0.30
8	Normal BMI	0.14	0.016	0.03	6.76E-06	0.33	0.26	0.26	0.44
9	Black/African American Race	0.13	0.015	0.02	2.03E-03	0.46	0.40	0.41	0.49
10	Negative bHCG	0.12	0.015	0.06	2.20E-25	0.37	0.23	0.23	0.42
11	HDL	0.09	0.015	0.01	4.06E-12	52.04	51.59	51.61	7.86

Table 5. Most significant variables for PCOS diagnosis prediction in Model IV.

Rank	**Variables**	***β***	***β* - %95 CI**	**Y-correlation**	**p-value**	**Y1-mean**	**Y0-mean**	**All-mean**	**All-std**
0	MLP Score	0.7	0.024	0.36	0.00E-01	0.20	0.07	0.08	0.10
1	Intercept	-0.44	-	-	-	-	-	-	-
2	Gravidity	-0.37	0.017	-0.14	2.17E-135	1.36	2.08	2.02	1.39
3	Positive bHCG	-0.34	0.017	-0.10	2.23E-65	0.08	0.23	0.22	0.41
4	Obesity	0.31	0.015	0.10	2.86E-66	0.43	0.27	0.28	0.45
5	Age	-0.21	0.015	-0.10	1.91E-52	31.26	33.79	33.59	6.62
6	Hispanic/Latina Race	-0.12	0.015	-0.03	2.34E-06	0.07	0.10	0.10	0.29
7	Smoker	-0.08	0.015	-0.02	3.00E-04	0.11	0.14	0.14	0.34
8	Hypertension	0.07	0.015	0.04	3.63E-12	0.28	0.21	0.22	0.41
9	Education – Some College/Technical/ Vocational School	0.06	0.014	0.03	1.55E-04	0.18	0.15	0.15	0.36
10	Negative bHCG	-0.05	0.015	0.05	2.29E-16	0.31	0.23	0.24	0.42

Table 6. Model performance over the test set, in the format of mean percentage (SD percentage) over 5 repetitions.

	Model I		Model II		Model III		Model IV	
	AUC	F1-weighted	AUC	F1-weighted	AUC	F1-weighted	AUC	F1-weighted
Best full models before SFS	XGBoost (51 features)		XGBoost (51 features)		XGBoost (51 features)		XGBoost (51 features)	
	85.2 (1.8)	94.5 (0.2)	80.6 (0.5)	95.1 (0.2)	80.4 (0.7)	94.8 (0.1)	81.8 (1.4)	91.1 (0.4)
Best full models after SFS	XGBoost (14 features)		XGBoost (16 features)		XGBoost (17 features)		XGBoost (17 features)	
	83.6 (1.7)	94.5 (0.2)	80.5 (0.7)	95.1 (0.2)	79.8 (1.1)	94.8 (0.1)	81.1 (1.3)	90.9 (0.3)

	LR-L1 (14 features)		LR-L1 (16 features)		LR-L1 (17 features)		LR-L1 (17 features)	
	79.2 (1.9)	93.9 (0.2)	71.7 (0.9)	94.7 (0.1)	72.9 (2.1)	94.4 (0.1)	74.8 (1.1)	89.7 (0.3)
	Parsimonious models LR-L2-MLP score (8 features)		Parsimonious models LR-L2-MLP score (10 features)		Parsimonious models LR-L2-MLP score (11 features)		Parsimonious models LR-L2-MLP score (10 features)	
	82.3 (1.7)	94.5 (0.1)	77.6 (1.3)	95.1 (0.1)	77.4 (1.6)	94.9 (0.1)	79.1 (1.1)	90.8 (0.3)