Ultrasound Obstet Gynecol 2022; **60**: 739–745
Published online in Wiley Online Library (wileyonlinelibrary.com). **DOI:** 10.1002/uog.26105

Machine-learning-based prediction of pre-eclampsia using first-trimester maternal characteristics and biomarkers

Z. ANSBACHER-FELDMAN[1], A. SYNGELAKI[2], H. MEIRI[3], R. CIRKIN[1], K. H. NICOLAIDES[2] and Y. LOUZOUN[1]

[1]*Department of Mathematics, Bar Ilan University, Ramat Gan, Israel;* [2]*Fetal Medicine Research Institute, King's College Hospital, London, UK;* [3]*The ASPRE Consortium and TeleMarpe, Tel Aviv, Israel*

KEYWORDS: artificial intelligence; first trimester; machine learning; mean arterial pressure; neural network; placental growth factor; posterior risk; pre-eclampsia; prior risk; uterine artery Doppler

CONTRIBUTION

What are the novel findings of this work?
Non-linear machine-learning classifiers can be used in combination with maternal risk factors and non-normalized first-trimester biomarkers to predict preterm pre-eclampsia (PE) with high accuracy. The incidence of PE and maternal first-trimester characteristics are race-dependent, and excluding the race information from the model significantly reduces the prediction accuracy in general and especially in non-white populations when only maternal factors are considered. We find that the values of mean arterial blood pressure, uterine artery pulsatility index and placental growth factor are crucial to reach accurate prediction, whereas pregnancy-associated plasma protein-A has a limited contribution.

What are the clinical implications of this work?
This work allows the prediction of PE using raw biomarker data without the need to convert them into multiples of the median, which is currently the standard approach to PE screening. This should facilitate wider implementation of first-trimester preterm PE prediction.

ABSTRACT

Objective *To evaluate the accuracy of predicting the risk of developing pre-eclampsia (PE) according to first-trimester maternal demographic characteristics, medical history and biomarkers using artificial-intelligence and machine-learning methods.*

Methods *The data were derived from prospective non-interventional screening for PE at 11–13 weeks' gestation at two maternity hospitals in the UK. The data were divided into three subsets. The first set, including 30 437 subjects, was used to develop the training process, the second set of 10 000 subjects was utilized to optimize the machine-learning hyperparameters and the third set of 20 352 subjects was coded and used for model validation. An artificial neural network was used to predict from the demographic characteristics and medical history the prior risk that was then combined with biomarker values to determine the risk of PE and preterm PE with delivery at < 37 weeks' gestation. An additional network was trained without including race as input. Biomarkers included uterine artery pulsatility index (UtA-PI), mean arterial blood pressure (MAP), placental growth factor (PlGF) and pregnancy-associated plasma protein-A. All markers were entered using raw values without conversion into standardized multiples of the median. The prediction accuracy was estimated using the area under the receiver-operating-characteristics curve (AUC). We further computed the detection rate at 10%, 20% and 40% false-positive rates (FPR). The impact of taking aspirin was also added. Shapley values were calculated to evaluate the contribution of each parameter to the prediction of risk. We used a non-parametric test to compare the expected AUC with the one obtained when we randomly scrambled the labels and kept the predictions. For the general prediction, we performed 10 000 permutations of the labels. When the AUC was higher than the one obtained in all 10 000 permutations, we reported a* P*-value of < 0.0001. For the race-specific analysis, we performed 1000 permutations. When the AUC was higher than the AUC in permutations, we reported a* P*-value of < 0.001.*

Results *The detection rate for preterm PE vs no PE, at a 10% FPR, was 53.3% when screening by maternal factors only, and the corresponding AUC was 0.816; these increased to 75.3% and 0.909, respectively, with the addition of biomarkers into the model. Information*

Correspondence to: Prof. K. H. Nicolaides, Fetal Medicine Research Institute, King's College Hospital, 16–20 Windsor Walk, London SE5 8BB, UK (e-mail: kypros@fetalmedicine.com)

Accepted: 3 August 2022

on race was important for the prediction accuracy; when race was not used to train the model, at a 10% FPR, the detection rate of preterm PE vs no PE decreased to 34.5–45.5% (for different races) when screening by maternal factors only and to 55.0–62.1% when biomarkers were added. The major predictors of PE were high MAP and UtA-PI, and low PlGF. The accuracy of prediction of all PE cases was lower than that for preterm PE. Aspirin use was recommended for cases who were at high risk of preterm PE. The AUC of all PE vs no PE was 0.770 when screening by maternal factors and 0.817 when the biomarkers were added; the respective detection rates, at a 10% FPR, were 41.3% and 52.9%.

Conclusions *Screening for PE using a non-linear machine-learning-based approach does not require a population-based normalization, and its performance is similar to that of logistic regression. Removing race information from the model reduces its prediction accuracy, especially for the non-white populations when only maternal factors are considered. © 2022 International Society of Ultrasound in Obstetrics and Gynecology.*

INTRODUCTION

Pre-eclampsia (PE) is a major cause of maternal and fetal morbidity and mortality[1,2]. First-trimester screening for PE by a combination of maternal characteristics and medical history with the measurements of mean arterial pressure (MAP), uterine artery pulsatility index (UtA-PI), serum placental growth factor (PlGF) and serum pregnancy-associated plasma protein-A (PAPP-A) can predict about 75% of preterm PE cases with delivery at < 37 weeks' gestation and 40–45% of term PE cases, at a 10% false-positive rate (FPR)[3–5]. Treatment of the high-risk group with aspirin (150 mg/day) from 12 to 36 weeks' gestation reduces the rate of preterm PE by approximately 60%[6].

The competing-risks approach, which is a method of screening for PE developed by the Fetal Medicine Foundation, assumes that every woman has a personalized distribution of gestational age at delivery with PE; whether a woman experiences PE or not before a specified gestational age depends on competition between delivery before or after the development of PE[5]. The distribution of biomarkers is specified based on gestational age at delivery with PE. The values of UtA-PI, MAP, PlGF and PAPP-A are expressed as multiples of the median (MoM) after adjustment for gestational age and various maternal factors that have been found to have a substantial effect on the $\log_{10}$ transformed values of the biomarkers[7–10]. However, MoM-based methods require detailed information on the distribution of measures in a sufficiently large cohort to allow prediction, which is often lacking in many populations. In addition, when applied to biochemical markers, the conversion to MoM has to be updated repeatedly for each new dataset in order to adjust it to different batches, manufacturers and analyzers.

Recently, artificial-intelligence, machine-learning and deep-learning methods have attracted strong interest around the world. These methods have already been tested in the diagnosis and prediction of many prenatal complications, such as Down syndrome, various structural anomalies identified by ultrasound and autism spectrum disorders[11–14]. In these studies, learning from dataset patterns enabled artificial-intelligence and machine-learning methods to identify interactions between variables and outcomes that are not accessible by linear methods[14,15].

The objective of this study was to examine the potential value of neural networks for the prediction of PE by a combination of maternal factors and biomarkers obtained at 11–13 weeks' gestation without converting raw data into MoMs.

METHODS

Study population

The data were derived from prospective screening for adverse obstetric outcome in women attending for their routine first-trimester hospital visit at King's College Hospital, London, and Medway Maritime Hospital, Gillingham, UK, between March 2006 and March 2017. This visit was held at 11 + 0 to 13 + 6 weeks' gestation and included, first, recording of maternal characteristics and medical history[3]; second, transabdominal ultrasound for measurement of left and right UtA-PI using color Doppler and calculation of the mean UtA-PI[16]; third, measurement of MAP by a validated automated device according to a standardized protocol[17]; and fourth, measurement of serum concentration of PlGF and PAPP-A using the DELFIA Xpress system (PerkinElmer Life and Analytical Sciences, Waltham, MA, USA) or Cobas e411 system (Roche Diagnostics, Penzberg, Germany). MAP, UtA-PI, PlGF and PAPP-A were measured on the day of the visit. The women gave written informed consent to participate in the study, which was approved by the NHS research ethics committee.

The inclusion criteria for this study were a singleton pregnancy undergoing first-trimester combined screening for aneuploidy and subsequently delivering a phenotypically normal live birth or stillbirth at ≥ 24 weeks' gestation. Pregnancies with aneuploidy or major fetal abnormality and those resulting in termination, miscarriage, or fetal death before 24 weeks were excluded.

Outcome measures were preterm PE with delivery at < 37 weeks' gestation and term PE with delivery at ≥ 37 weeks. Data on pregnancy outcomes were collected from the hospital maternity records or the general medical practitioners of the women. The obstetric records of all women with pre-existing or pregnancy-associated hypertension were examined to determine if the condition was PE, as defined by the American College of Obstetricians and Gynecologists[2]. According to this definition, diagnosis of PE requires the presence of new-onset hypertension (systolic blood pressure ≥ 140 mmHg or diastolic blood pressure ≥ 90 mmHg) at ≥ 20 weeks' gestation or chronic hypertension and either proteinuria (≥ 300 mg/24 h or protein-to-creatinine ratio ≥ 30 mg/mmol or ≥ 2+ on dipstick testing) or evidence of renal dysfunction (serum creatinine > 97 μmol/L), hepatic dysfunction (transaminases ≥ 65 IU/L) or hematological dysfunction (platelet count < 100 000/μL).

Machine learning

The input data were converted into *Z*-scores. Categorical parameters were represented using one-hot encoding and were not normalized. For the prediction, we used a feed-forward neural network with two hidden layers. The activation function, which is commonly used in neural networks, was a rectified linear unit; this is defined as $y = \max(0, x)$. Dropout was applied to the second layer. An Adam optimizer algorithm was used[18]. The loss function was binary cross-entropy with logits (weighted). Machine learning was performed using PyTorch[19]. Grid search was implemented via NNI (https://github.com/microsoft/nni), an automatic tool for hyperparameter tuning, which optimizes machine-learning performance. The following hyperparameters were tuned: batch size, learning rate, dropout rate, size (number of neurons) of each hidden layer, activation function and weight decay. The tuning was done separately for the prior and posterior risk-based predictions (the prior model was based on maternal factors only, while the posterior model also used biomarkers). The dataset was split into a training set prepared from the data of 30 437 subjects, an internal validation set of 10 000 subjects and a test set of 20 352 subjects. The tuning was done on the internal validation set, and the results were reported based on the test set, which was not available at the time of tuning. While the data on the outcome of the training set were disclosed, for the test dataset, the outcome data were coded and unknown to the team in Israel that conducted the machine-learning analysis. The tuning was performed on the internal validation set, using the area under the receiver-operating-characteristics curve (AUC), a combined measure of sensitivity and specificity.

Experimental setup

Multiple tests were performed, and in all tests, the same training/validation and test division were used. The prediction was performed either including or excluding the race input. When the race input was ignored, the training was performed on the entire dataset, but the test was performed separately for each race group.

In all cases, the following combinations were tested independently: (a) PE *vs* no PE, (b) preterm PE *vs* no PE and (c) preterm PE *vs* no PE plus term PE (everything else). When preterm PE was compared with no PE, term PE cases were ignored in both the training and the test sets.

Shapley values

Data Shapley values[20] reached fairness by considering all subsets of subjects in the training set and calculating a weighted sum of the individual contributions. The computational effort for the exact calculation of Data Shapley values grows exponentially with the number of subjects (*n*) because a set of *n*-elements contains $2n - 1$ non-empty subsets. However, there are effective possibilities to estimate Data Shapley values. In this work, Truncated Monte Carlo Shapley was used[21]. The Truncated Monte Carlo algorithm starts with a random permutation of the training set. First, the performance of a random model was calculated. In this work, the AUC for the predefined validation dataset was used as the performance score. Afterward, the randomly permuted subjects were added successively to the training dataset, and machine-learning models were trained. The contribution of each added subject was calculated by subtracting the previously achieved validation performance from the validation performance of the new model. This procedure was repeated until the addition of a new subject achieved only marginal improvement. Afterward, the procedure was repeated with a new permutation. One contribution is thus calculated for each permutation and each subject.

Statistical analysis

Two methods were used for evaluation. The prediction accuracy was estimated using the AUC. The detection rate (the recall) was also computed as a function of the proportion of women who screened positive for PE. Since the total proportion of PE cases in the population was low, our goal was to minimize the proportion of women who screened positive for PE but maximize the recall. The *P*-values reported are the probability that the results are random. A non-parametric permutation test was used to compare the observed AUC with the one obtained when we randomly scrambled the labels of each sample but kept its predicted score. For the general prediction, 10 000 permutations of the labels were performed. When the AUC was higher than the AUC obtained in all 10 000 permutations, a *P*-value of < 0.0001 was reported. For the race-specific analysis, 1000 permutations were performed. When the AUC was higher than the AUC obtained in permutations, a *P*-value of < 0.001 was reported.

RESULTS

Characteristics of study population

The study population of 60 789 pregnancies included 1722 (2.8%) subjects that developed PE. The characteristics of the study population are summarized in Table 1. In the PE group, compared with the non-PE group, there were higher body mass index, interpregnancy interval, proportion of black women and rates of chronic hypertension, diabetes mellitus, systemic lupus erythematosus or antiphospholipid syndrome, family and personal history of PE and conception via assisted fertility techniques, and lower incidence of smoking.

Table 1 Characteristics of study population according to pre-eclampsia (PE) status

Characteristic	*Non-PE* (n = *59 067*)	*PE* (n = *1722*)	P
Maternal age (y)	31.0 (26.6–34.8)	31.2 (26.7–35.2)	0.112
Maternal weight (kg)	67.0 (59.2–78.0)	74.0 (63.9–87.2)	< 0.0001
Maternal height (cm)	165 (160–169)	164 (159–168)	< 0.0001
GA at screening (days)	89.2 (85.1–93.3)	89.0 (85.0–93.0)	0.024
Race			< 0.0001
White	43 963 (74.3)	993 (57.2)	
Black	9790 (16.6)	599 (34.5)	
South Asian	2641 (4.5)	83 (4.8)	
East Asian	1230 (2.1)	24 (1.4)	
Mixed	1515 (2.6)	37 (2.1)	
Medical history			
Chronic hypertension	630 (1.1)	215 (12.4)	< 0.0001
DM Type I	228 (0.4)	12 (0.7)	< 0.0001
DM Type II	294 (0.5)	26 (1.5)	< 0.0001
SLE/APS	113 (0.2)	9 (0.5)	0.006
Smoker	5667 (9.6)	101 (5.8)	< 0.0001
Family history of PE	2257 (3.8)	136 (7.8)	< 0.0001
Method of conception			< 0.0001
Spontaneous	57 258 (96.8)	1644 (94.7)	
In-vitro fertilization	1408 (2.4)	72 (4.1)	
Use of ovulation drugs	473 (0.8)	20 (1.2)	
Parity			< 0.0001
Nulliparous	27 303 (46.2)	1008 (58.1)	
Parous, no previous PE	30 179 (51.0)	494 (28.5)	
Parous, previous PE	1657 (2.8)	234 (13.5)	
Interpregnancy interval (y)	3.0 (2.0–4.9)	3.9 (2.3–6.7)	< 0.0001
Aspirin	1111 (1.9)	98 (5.6)	< 0.0001
Biomarker			
MAP (mmHg)	86.3 (81.1–91.8)	93.8 (87.8–99.8)	< 0.0001
UtA-PI	1.7 (1.3–2.0)	1.9 (1.5–2.4)	< 0.0001
PlGF (pg/mL)	35.3 (25.9–49.6)	28.1 (20.4–40.7)	< 0.0001
PAPP-A (IU/L)	2.7 (1.7–4.2)	2.3 (1.4–3.8)	< 0.0001

Data are given as median (interquartile range) or *n* (%). APS, antiphospholipid syndrome; DM, diabetes mellitus; GA, gestational age; MAP, mean arterial pressure; PAPP-A, pregnancy-associated plasma protein-A; PlGF, placental growth factor; SLE, systemic lupus erythematosus; UtA-PI, uterine artery pulsatility index; y, years.

Performance of screening for pre-eclampsia

The data were separated into training, internal validation and external test sets (Figure 1). The training set was the input of an artificial neural network to predict three independent tasks: PE *vs* no PE, preterm PE *vs* no PE (term PE cases omitted) and preterm PE *vs* everything else (no PE plus term PE). The internal validation set was used for tuning hyperparameters to maximize the AUC of the internal validation. The trained model was then applied to the test set, and the results are reported below.

The detection rates, at FPR of 10%, 20% and 40%, on the test set are given in Table 2, and the ROCs are presented in Figure 2. This prediction was performed separately based on the prior-risk data (demographic characteristics and medical and pregnancy history data) and posterior-risk data (all the characteristics plus biomarkers PAPP-A, PlGF, MAP and UtA-PI). The accuracy increased consistently with the addition of biomarkers. Yet, even without using the biomarkers, the artificial neural network successfully predicted PE. Specifically, the AUC for preterm PE *vs* everything else increased from 0.808 to 0.904 when posterior information was added, and the detection rate, at a 10% FPR,

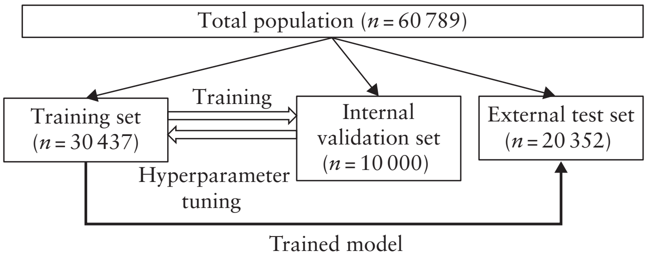
Total population (*n* = 60 789)
Training set (*n* = 30 437)
Training
Internal validation set (*n* = 10 000)
External test set (*n* = 20 352)
Hyperparameter tuning
Trained model

Figure 1 Flowchart showing division of data, from 60 789 pregnancies which underwent first-trimester screening for pre-eclampsia, into training, validation and test sets. The training dataset was used for the training process of the algorithm, and the validation dataset was used to check the algorithm's performance. Different combinations of hyperparameters were checked in this process, and the parameters that optimized the performance on the validation set were used in the final model. The trained model was then applied to the test set.

Table 2 Performance of screening for pre-eclampsia (PE) on test set

Screening method	*AUC*	P	*Detection rate (%) at:* 10% FPR	20% FPR	40% FPR
All PE *vs* no PE (*n* = 60 789)					
MF	0.770 (0.747–0.793)	< 0.0001	41.3 (35.2–47.4)	58.7 (53.6–63.8)	76.6 (72.5–80.7)
MF, PAPP-A, PlGF, MAP, UtA-PI	0.817 (0.797–0.837)	< 0.0001	52.9 (48.2–57.6)	66.1 (61.2–71.0)	822 (79.9–84.5)
Preterm PE *vs* no PE (*n* = 59 551)					
MF	0.816 (0.769–0.863)	< 0.0001	53.3 (44.3–62.3)	67.1 (59.1–75.1)	81.5 (70.6–92.4)
MF, PAPP-A, PlGF, MAP, UtA-PI	0.909 (0.895–0.923)	< 0.0001	75.3 (68.9–81.7)	87.8 (83.6–92.0)	93.4 (89.5–97.3)
Preterm PE *vs* everything else (*n* = 60 789)					
MF	0.808 (0.759–0.857)	< 0.0001	47.3 (40.1–54.5)	64.2 (53.3–75.1)	79.5 (67.6–91.4)
MF, PAPP-A, PlGF, MAP, UtA-PI	0.904 (0.890–0.918)	< 0.0001	75.2 (68.7–81.7)	86.7 (81.2–92.2)	92.2 (86.7–97.7)

P-values reported are probability that results are random and were calculated by 10 000 random permutations. *P*-value < 0.0001 means that the reported area under the receiver-operating-characteristics curve (AUC) is higher than the AUC in permutations. FPR, false-positive rate; MAP, mean arterial pressure; MF, maternal factors; PAPP-A, pregnancy-associated plasma protein-A; PlGF, placental growth factor; UtA-PI, uterine artery pulsatility index.

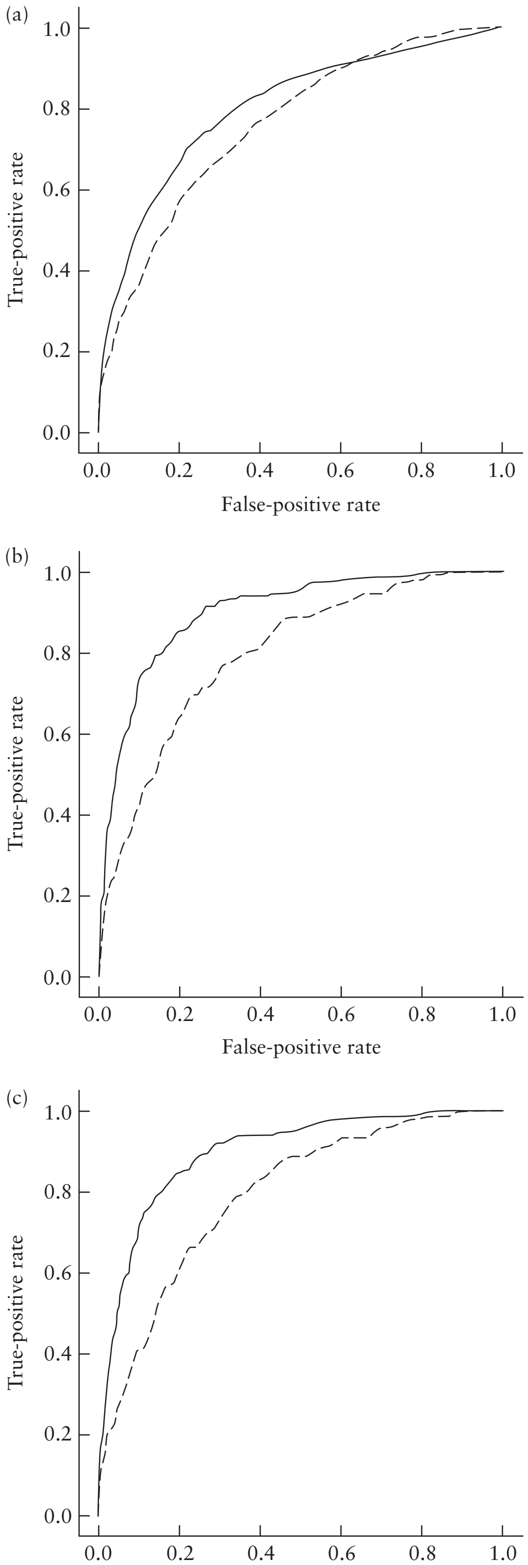

Figure 2 Receiver-operating-characteristics (ROC) curves demonstrating performance of screening based on posterior-risk data (biomarkers and maternal factors; ——) or prior-risk data (maternal factors only; – – –) in predicting: (a) pre-eclampsia (PE) *vs* no PE (area under the ROC curve (AUC), 0.817 (posterior risk) and 0.770 (prior risk)), (b) preterm PE *vs* no PE (AUC, 0.909 (posterior risk) and 0.816 (prior risk)) and (c) preterm PE *vs* everything else (i.e. no PE plus term PE) (AUC, 0.904 (posterior risk) and 0.808 (prior risk)).

increased from 47% to 75%. When screening for preterm PE *vs* no PE based on posterior-risk data, the results were slightly better (AUC of 0.909 and detection rate of 75%, at a 10% FPR). However, in that analysis, some of the samples were ignored (term PE data).

The results of the analysis on the influence of the mother's race on test accuracy are shown in Table 3. The study population included white, black, South Asian, East Asian and mixed races. However, the numbers of PE-positive cases of South Asian, East Asian and mixed races were too low for robust analysis and the AUCs for these groups were thus not reported. Excluding race from the analysis was consistently associated with a reduction in the accuracy of all predictors. For example, the AUC of screening for preterm PE *vs* everything else based on the prior-risk information decreased from 0.808 to 0.750 and the detection rate, at a 10% FPR, decreased from 47% to 38%. When comparing populations, the accuracy was typically higher for the white population than for the black population when considering only maternal factors.

Contribution of biomarkers to prediction accuracy

To test the impact of different factors on the prediction accuracy, we computed the Shapley values of the different features used for prediction. The Shapley values represent the average contribution to the score of each input feature when computed with different combinations of the other features. The highest contribution was provided by MAP and UtA-PI, for which a high value led to a high risk for preterm PE, followed by PlGF, for which a low value was associated with a high risk for preterm PE (Figure 3). This was followed by race, with a higher risk of preterm PE associated with the black race, and a lower risk associated with the white race. Other markers, such as PAPP-A, had a very limited contribution.

Effect of aspirin

To determine the effect of aspirin treatment and the relationship between our current risk prediction and treatment, we first checked whether our risk prediction is associated with aspirin treatment. We computed the proportion of women receiving aspirin as a function of their risk percentile and found that most, but not all, women receiving aspirin were from the group at high risk for preterm PE. We then tested the efficacy of aspirin by comparing the proportion of women with PE according to the predicted risk percentile in aspirin and non-aspirin groups. While, on lower risk percentiles, the aspirin group had a low proportion of women with any PE and preterm PE, on higher risk percentiles, women taking aspirin actually had a very high proportion of cases positive for PE and preterm PE. This is likely to be the consequence of selecting women to receive treatment because their risk for PE is very high.

Table 3 Performance of screening model for pre-eclampsia (PE) on test set, trained without taking into account maternal race, overall and according to race (analysis was not presented separately for other races due to insufficient data)

			Detection rate (%) at:		
Screening method	*AUC*	P	*10% FPR*	*20% FPR*	*40% FPR*
All PE *vs* no PE					
MF					
All (*n* = 60 789)	0.742 (0.722–0.762)	< 0.001	33.1 (31.9–34.3)	50.4 (47.3–53.5)	74.6 (69.5–79.7)
White (*n* = 44 899)	0.740 (0.717–0.763)	< 0.001	31.6 (28.5–34.7)	49.8 (45.3–54.3)	73.9 (67.8–80.0)
Black (*n* = 10 364)	0.725 (0.694–0.756)	< 0.001	34.8 (29.7–39.9)	45.2 (39.7–50.7)	70.3 (66.0–74.6)
MF, PAPP-A, PlGF, MAP, UtA-PI					
All (*n* = 60 789)	0.790 (0.778–0.802)	< 0.001	43.8 (41.8–45.8)	60.1 (56.0–64.2)	81.3 (79.0–83.6)
White (*n* = 44 899)	0.730 (0.718–0.742)	< 0.001	40.4 (36.9–43.9)	57.1 (55.0–59.2)	77.7 (74.0–81.4)
Black (*n* = 10 364)	0.820 (0.787–0.853)	< 0.001	43.9 (35.7–52.1)	62.6 (54.4–70.8)	85.6 (81.1–90.1)
Preterm PE *vs* no PE					
MF					
All (*n* = 59 551)	0.750 (0.723–0.777)	< 0.001	37.6 (31.5–43.7)	54.1 (48.6–59.6)	77.4 (71.1–83.7)
White (*n* = 44 150)	0.760 (0.697–0.823)	< 0.001	45.5 (34.6–56.4)	57.3 (43.2–71.4)	76.5 (65.9–87.1)
Black (*n* = 9971)	0.680 (0.602–0.758)	< 0.001	34.5 (26.3–42.7)	41.4 (29.3–53.5)	69.0 (56.7–81.3)
MF, PAPP-A, PlGF, MAP, UtA-PI					
All (*n* = 59 551)	0.880 (0.857–0.903)	< 0.001	62.1 (57.2–67.0)	81.2 (74.6–87.8)	93.3 (89.4–97.2)
White (*n* = 44 150)	0.870 (0.806–0.934)	< 0.001	59.0 (50.4–67.6)	80.2 (66.3–94.1)	90.7 (89.6–91.8)
Black (*n* = 9971)	0.880 (0.853–0.907)	< 0.001	55.0 (43.1–66.9)	76.7 (66.2–87.2)	95.1 (90.6–99.6)
Preterm PE *vs* everything else					
MF					
All (*n* = 60 789)	0.750 (0.727–0.773)	< 0.001	38.1 (32.0–44.2)	54.8 (48.1–61.5)	75.7 (71.0–80.4)
White (*n* = 44 899)	0.760 (0.705–0.815)	< 0.001	44.2 (34.1–54.3)	58.8 (44.3–73.3)	77.1 (68.5–85.7)
Black (*n* = 10 364)	0.680 (0.609–0.751)	< 0.001	35.2 (28.1–42.3)	38.6 (29.6–47.6)	66.1 (53.4–78.8)
MF, PAPP-A, PlGF, MAP, UtA-PI					
All (*n* = 60 789)	0.880 (0.857–0.903)	< 0.001	60.8 (55.3–66.3)	79.9 (74.0–85.8)	92.4 (88.7–96.1)
White (*n* = 44 899)	0.860 (0.794–0.926)	< 0.001	62.9 (50.6–75.2)	80.2 (66.3–94.1)	90.7 (89.6–91.8)
Black (*n* = 10 364)	0.870 (0.843–0.897)	< 0.001	55.0 (43.1–66.9)	74.3 (65.7–82.9)	95.1 (90.6–99.6)

P-values were calculated here by 1000 random permutations. AUC, area under receiver-operating-characteristics curve; FPR, false-positive rate; MAP, mean arterial pressure; MF, maternal factors; PAPP-A, pregnancy-associated plasma protein-A; PlGF, placental growth factor; UtA-PI, uterine artery pulsatility index.

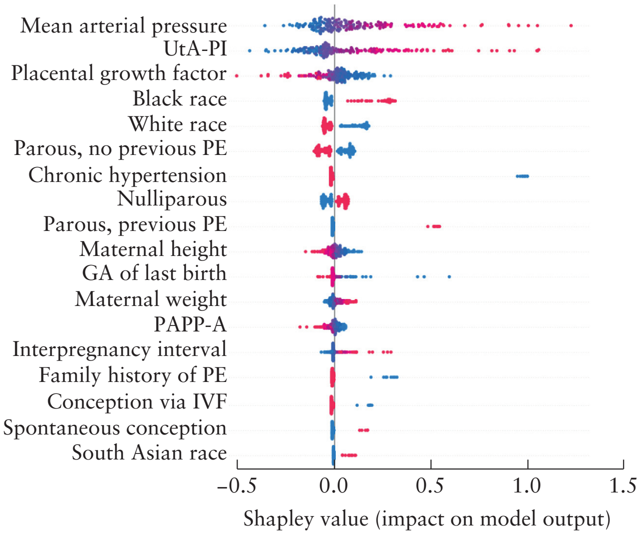

Figure 3 Shapley values of predictors of preterm pre-eclampsia (PE) *vs* no PE. Red circles represent high input of the variable, while blue circles represent low input. Values to the right are associated with higher risk for preterm PE, while values to the left are associated with lower risk. Input variables are ordered by their average contribution to the model output (the higher variables have a stronger effect on the model). GA, gestational age; IVF, *in-vitro* fertilization; PAPP-A, pregnancy-associated plasma protein-A; UtA-PI, uterine artery pulsatility index.

DISCUSSION

Main findings

In this first-trimester screening study, artificial intelligence and machine learning with the assistance of neural-network algorithms were used for predicting the risk of subsequent development of PE. There were two main findings: first, at a 10% FPR, the prediction of preterm PE *vs* no PE plus term PE was 47% when screening based on maternal characteristics and medical history, increasing to 75% after the addition of biomarkers and, second, the inclusion of race in the prediction algorithm was important because, when race was not included, the detection rate of preterm PE *vs* no PE plus term PE, at a 10% FPR, of combined screening decreased from 75% to 55–63%.

Comparison with previous studies and implications for clinical practice

First-trimester prediction of preterm PE is important because treatment of the high-risk group with aspirin (150 mg/day from 12 to 36 weeks' gestation) reduces the rate of early PE with delivery < 32 weeks by about 90% and preterm PE by about 60%[4,6,22]. Consequently, early prediction and prevention of PE has been adopted by the guidelines of the International Society for the Study of Hypertension in Pregnancy[1] and the International Federation of Gynecology and Obstetrics[23].

The predictive performance for preterm PE using artificial-intelligence and machine-learning methods was similar to that achieved by the competing-risks model[3–5,24,25]. The advantage of the machine-learning approach is the use of raw biomarker data without the need for conversion into MoMs, which would simplify the implementation of screening. Additionally, calculators from the machine-learning approach can be introduced easily and rapidly in an automated way with the help of cloud-based or other online tools. We now have an online prediction tool based on the model at: https://pepred.math.biu.ac.il/Home.

The use of Shapley-value analysis[20,21] performed in our study showed a very high contribution of MAP, UtA-PI, PlGF and race to the prediction of PE risk. There were a few women with chronic medical conditions, including chronic hypertension, diabetes mellitus and autoimmune disease, and among those who had these conditions, the impact of chronic conditions was high. PAPP-A had a low contribution to the prediction of PE.

Strengths and limitations

The main strengths of the study are the large population derived from prospective screening for PE, recording all the important demographic and medical factors known to be associated with PE, measurement of MAP and UtA-PI using standardized protocols and by appropriately trained practitioners, and measurement of PlGF within 30 min of collection using an automated machine that was calibrated on a daily basis.

The limitations of the study are the lack of testing of the prediction algorithm in other populations. For example, our finding of the large influence of race on the accuracy of PE prediction demonstrates a limitation of our study, as the race element introduces bias towards the most prevalent race in the study population. Consequently, it is anticipated that it will be necessary to make adjustments to the algorithm when it is applied to other populations.

Conclusion

Our study demonstrates the utility and accuracy of a novel automated machine-learning approach in first-trimester prediction of preterm PE. The study also demonstrates the importance of taking into account race in the prediction of PE.

ACKNOWLEDGMENTS

Z.A.-F. was sponsored by Bar Ilan University Graduate program. The screening study for pre-eclampsia was supported by grants from the Fetal Medicine Foundation (UK Charity No: 1037116).

REFERENCES

1. Magee LA, Brown MA, Hall DR, Gupte S, Hennessy A, Karumanchi SA, Kenny LC, McCarthy F, Myers J, Poon LC, Rana S, Saito S, Staff AC, Tsigas E, von Dadelszen P. The 2021 International Society for the Study of Hypertension in Pregnancy classification, diagnosis & management recommendations for international practice. *Pregnancy Hypertens* 2021; **27**: 148–169.
2. American College of Obstetricians and Gynecologists. Gestational Hypertension and Preeclampsia: ACOG Practice Bulletin, Number 222. *Obstet Gynecol* 2020; **135**: e237–260.
3. Wright D, Syngelaki A, Akolekar R, Poon LC, Nicolaides KH. Competing risks model in screening for preeclampsia by maternal characteristics and medical history. *Am J Obstet Gynecol* 2015; **213**: 62.e1–10.
4. O'Gorman N, Wright D, Syngelaki A, Akolekar R, Wright A, Poon LC, Nicolaides KH. Competing risks model in screening for preeclampsia by maternal factors and biomarkers at 11-13 weeks gestation. *Am J Obstet Gynecol* 2016; **214**: 103.e1–12.
5. Wright D, Wright A, Nicolaides KH. The competing risk approach for prediction of preeclampsia. *Am J Obstet Gynecol* 2020; **223**: 12–23.e7.
6. Rolnik DL, Wright D, Poon LC, O'Gorman N, Syngelaki A, de Paco Matallana C, Akolekar R, Cicero S, Janga D, Singh M, Molina FS, Persico N, Jani JC, Plasencia W, Papaioannou G, Tenenbaum-Gavish K, Meiri H, Gizurarson S, Maclagan K, Nicolaides KH. Aspirin versus Placebo in Pregnancies at High Risk for Preterm Preeclampsia. *N Engl J Med* 2017; **377**: 613–622.
7. Tayyar A, Guerra L, Wright A, Wright D, Nicolaides KH. Uterine artery pulsatility index in the three trimesters of pregnancy: effects of maternal characteristics and medical history. *Ultrasound Obstet Gynecol* 2015; **45**: 689–697.
8. Wright A, Wright D, Ispas CA, Poon LC, Nicolaides KH. Mean arterial pressure in the three trimesters of pregnancy: effects of maternal characteristics and medical history. *Ultrasound Obstet Gynecol* 2015; **45**: 698–706.
9. Wright D, Silva M, Papadopoulos S, Wright A, Nicolaides KH. Serum pregnancy-associated plasma protein-A in the three trimesters of pregnancy: effects of maternal characteristics and medical history. *Ultrasound Obstet Gynecol* 2015; **46**: 42–50.
10. Tsiakkas A, Duvdevani N, Wright A, Wright D, Nicolaides KH. Serum placental growth factor in the three trimesters of pregnancy: effects of maternal characteristics and medical history. *Ultrasound Obstet Gynecol* 2015; **45**: 591–598.
11. Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H. State-of-the-art in artificial neural network applications: A survey. *Heliyon* 2018; **4**: e00938.
12. He F, Lin B, Mou K, Jin L, Liu J. A machine learning model for the prediction of down syndrome in second trimester antenatal screening. *Clin Chim Acta* 2021; **521**: 206–211.
13. Matthew J, Skelton E, Day TG, Zimmer VA, Gomez A, Wheeler G, Toussaint N, Liu T, Budd S, Lloyd K, Wright R, Deng S, Ghavami N, Sinclair M, Meng Q, Kainz B, Schnabel JA, Rueckert D, Razavi R, Simpson J, Hajnal J. Exploring a new paradigm for the fetal anomaly ultrasound scan: Artificial intelligence in real time. *Prenat Diagn* 2022; **42**: 49–59.
14. Caly H, Rabiei H, Coste-Mazeau P, Hantz S, Alain S, Eyraud JL, Chianea T, Caly C, Makowski D, Hadjikhani N, Lemonnier E, Ben-Ari Y. Machine learning analysis of pregnancy data enables early identification of a subpopulation of newborns with ASD. *Sci Rep* 2021; **11**: 6877.
15. Zhang H, Mo J, Jiang H, Li Z, Hu W, Zhang C, Wang Y, Wang X, Liu C, Zhao B, Zhang J, Zhang K. Deep learning model for the automated detection and histopathological prediction of meningioma. *Neuroinformatics* 2021; **19**: 393–402.
16. Plasencia W, Maiz N, Bonino S, Kaihura C, Nicolaides KH. Uterine artery Doppler at 11 + 0 to 13 + 6 weeks in the prediction of pre-eclampsia. *Ultrasound Obstet Gynecol* 2007; **30**: 742–749.
17. Poon LC, Zymeri NA, Zamprakou A, Syngelaki A, Nicolaides KH. Protocol for measurement of mean arterial pressure at 11-13 weeks' gestation. *Fetal Diagn Ther* 2012; **31**: 42–48.
18. Zhang, Zijun. Improved adam optimizer for deep neural networks. IEEE/ACM 26th International Symposium on Quality of Service (IWQoS), 2018.
19. Deo RC. Machine learning in medicine. *Circulation* 2015; **132**: 1920–1930.
20. Shapley LS. A value for n-person games. In *Contributions to the Theory of Games*, Kuhn HW, Tucker AW (eds). Princeton University Press: Princeton, NJ, 1953; 307–18.
21. Ghorbani A, Zou J. Data Shapley: equitable valuation of data for machine learning. Proceedings of the 36th International Conference on Machine Learning (ICML 2019), Long Beach, CA, USA, 2019: 2242–51.
22. Roberge S, Bujold E, Nicolaides KH. Aspirin for the prevention of preterm and term preeclampsia: systematic review and metaanalysis. *Am J Obstet Gynecol* 2018; **218**: 287–293.e1.
23. Poon LC, Shennan A, Hyett JA, Kapur A, Hadar E, Divakar H, McAuliffe F, da Silva Costa F, von Dadelszen P, McIntyre HD, Kihara AB, Di Renzo GC, Romero R, D'Alton M, Berghella V, Nicolaides KH, Hod M. The International Federation of Gynecology and Obstetrics (FIGO) initiative on pre-eclampsia: A pragmatic guide for first-trimester screening and prevention. *Int J Gynaecol Obstet* 2019; **145**: 1–33.
24. Tan MY, Syngelaki A, Poon LC, Rolnik DL, O'Gorman N, Delgado JL, Akolekar R, Konstantinidou L, Tsavdaridou M, Galeva S, Ajdacka U, Molina FS, Persico N, Jani JC, Plasencia W, Greco E, Papaioannou G, Wright A, Wright D, Nicolaides KH. Screening for pre-eclampsia by maternal factors and biomarkers at 11–13 weeks' gestation. *Ultrasound Obstet Gynecol* 2018; **52**: 186–195.
25. Wright D, Tan MY, O'Gorman N, Poon LC, Syngelaki A, Wright A, Nicolaides KH. Predictive performance of the competing risk model in screening for preeclampsia. *Am J Obstet Gynecol* 2019; **220**: 199.e1–13.

Ultrasound Obstet Gynecol 2022; **60**: 739–745
Published online in Wiley Online Library (wileyonlinelibrary.com). **DOI:** 10.1002/uog.26105

Predicción de la preeclampsia basada en aprendizaje automático a partir de características maternas y biomarcadores del primer trimestre

RESUMEN

Objetivo. Evaluar la precisión de la predicción del riesgo de desarrollo de preeclampsia (PE) en función de las características demográficas maternas del primer trimestre, los antecedentes médicos y biomarcadores mediante métodos de inteligencia artificial y aprendizaje automático.

Métodos. Los datos proceden de un cribado prospectivo no intervencionista de la PE a las 11–13 semanas de gestación en dos maternidades del Reino Unido. Los datos se dividieron en tres subconjuntos. El primer subconjunto, que incluía 30 437 sujetos, se utilizó para desarrollar el proceso de entrenamiento, el segundo subconjunto de 10 000 sujetos se utilizó para optimizar los hiperparámetros de aprendizaje automático y el tercer subconjunto de 20 352 sujetos se codificó y se utilizó para la validación del modelo. Se utilizó una red neuronal artificial para predecir el riesgo previo a partir de las características demográficas y los antecedentes clínicos, que fue combinado a continuación con los valores de los biomarcadores para determinar el riesgo de PE y PE pretérmino con parto a <37 semanas de gestación. Se entrenó una red adicional que no incluyó la raza como datos de entrada. Entre los biomarcadores se incluyeron el índice de pulsatilidad de la arteria uterina (UtA-PI), la presión arterial media (PAM), el factor de crecimiento placentario (FCPI) y la proteína plasmática A asociada al embarazo. Todos los marcadores se introdujeron utilizando valores brutos sin conversión a múltiplos estandarizados de la mediana. La precisión de la predicción se estimó mediante el área bajo la curva (ABC) de características operativas del receptor. Además, se calculó la tasa de detección con tasas de falsos positivos (TFP) del 10%, el 20% y el 40%. También se incluyó el impacto de tomar aspirina. Se calcularon los valores de Shapley para evaluar la contribución de cada parámetro a la predicción del riesgo. Se utilizó una prueba no paramétrica para comparar el ABC esperada con la obtenida cuando se mezclaron aleatoriamente las etiquetas y se mantuvieron las predicciones. Para la predicción general se realizaron 10 000 permutaciones de las etiquetas. Cuando el ABC fue superior a la obtenida en las 10 000 permutaciones, se notificó un valor P de <0,0001. Para el análisis específico por raza se realizaron 1000 permutaciones. Cuando el ABC fue mayor que el ABC de las permutaciones, se notificó un valor P de <0,001.

Resultados. La tasa de detección de la PE pretérmino frente a no PE, con una TFP del 10%, fue del 53,3% cuando el cribado se realizó sólo por factores maternos, y el ABC correspondiente fue de 0,816. Estos valores aumentaron respectivamente al 75,3% y 0,909 con la adición de biomarcadores al modelo. La información sobre la raza fue importante para la precisión de la predicción. Cuando no se utilizó la raza para entrenar el modelo, con una TFP del 10%, la tasa de detección de PE pretérmino frente a la no PE disminuyó hasta el 34,5–45,5% (para diferentes razas) cuando el cribado estuvo basado únicamente en factores maternos y hasta el 55,0–62,1% cuando se añadieron biomarcadores. Los principales factores predictivos de la PE fueron una PAM y un UtA-PI elevados, y un FCPI bajo. La precisión de la predicción de todos los casos de PE fue inferior a la de la PE pretérmino. Se recomendó el uso de aspirina en los casos de alto riesgo de PE pretérmino. El ABC de todas las PE frente a no PE fue de 0,770 cuando el cribado se realizó respecto a factores maternos y de 0,817 cuando se añadieron los biomarcadores. Las tasas de detección respectivas, con una TFP del 10%, fueron del 41,3% y del 52,9%.

Conclusiones. El cribado de la PE mediante un enfoque no lineal basado en el aprendizaje automático no requiere una normalización basada en la población, y su rendimiento es similar al de la regresión logística. Eliminar la información sobre la raza del modelo reduce su precisión de predicción, especialmente en el caso de las poblaciones no caucásicas cuando sólo se tienen en cuenta los factores maternos.

采用基于机器学习的方法使用孕早期母体特征和生物标志物预测子痫前期

摘要

目的 使用人工智能和机器学习方法，根据母体孕早期人口统计学特征、病史和生物标志物，评价预测发生子痫前期（PE）风险的准确性。

方法 数据来自于英国两家妇产医院在孕妇妊娠11-13周时进行的前瞻性非干预性PE筛查。这些数据分为三组。第一组包括30,437名受试者，用于开发训练过程；第二组包括10,000名受试者，用于优化机器学习的超参数；第三组包括20,352名受试者，经编码用于模型验证。使用一个人工神经网络根据人口统计学特征和病史预测基础风险，然后结合生物标志物的值，确定PE风险以及在妊娠不足37周分娩时的早产PE风险。另外一个网络的训练则不包括种族数据。生物标志物包括子宫动脉搏动指数（UtA-PI）、平均动脉压（MAP）、胎盘生长因子（PlGF）和妊娠相关的血浆蛋白-A。所有标志物按原始值输入，没有转换为标准化的中位数倍数。预测的准确性根据接受者操作特征曲线下的面积（AUC）进行估算。我们进一步计算了10%、20%和40%假阳性率（FPR）下的检测率。同时还加入了服用阿司匹林的影响。我们计算了夏普利值以评估每个参数对风险预测的贡献。我们进行了一项非参数检验来比较预期AUC和我们随机打乱标签并保留预测结果时得到的AUC。对于一般的预测，我们对标签进行了10,000次排列。当AUC高于所有10,000次排列中得到的AUC时，我们报告P值<0.0001。对于特定种族的分析，我们进行了1,000次排列。当AUC高于排列中的AUC时，我们报告P值<0.001。

结果 在10%的FPR下，仅按母体因素进行筛查时，早产PE与无PE的检出率为53.3%，相应的AUC为0.816；将生物标志物加入模型后，两项数据分别提高到75.3%和0.909。种族信息对预测的准确性很重要；不使用种族信息来训练模型时，在10%的FPR下，仅按母体因素筛查时早产PE与无PE的检出率下降至34.5-45.5%（针对不同种族），而加入生物标志物后则下降至55.0-62.1%。PE的主要预测因子是高MAP和UtA-PI以及低PlGF。所有PE病例的预测准确率低于早产PE的预测准确率。建议有早产PE高风险的病例使用阿司匹林。按母体因素筛查时所有PE与无PE的AUC为0.770，加入生物标志物后为0.817；在10%的FPR下，二者的检出率分别为41.3%和52.9%。

结论 使用基于机器学习的非线性方法筛查PE，不需要基于人口的标准化，其表现与逻辑回归相似。从模型中去除种族信息会降低其预测准确性，在只考虑母体因素的情况下对非白种人的预测尤其如此。