BMC Pregnancy and Childbirth

RESEARCH

Open Access

A comprehensive and bias-free machine learning approach for risk prediction of preeclampsia with severe features in a nulliparous study cohort

Yun C. Lin[1*], Daniel Mallia[2], Andrea O. Clark-Sevilla[1], Adam Catto[2], Alisa Leshchenko[2], Qi Yan[3], David M. Haas[4], Ronald Wapner[3], Itsik Pe'er[1], Anita Raja[2] and Ansaf Salleb-Aouissi[1]

Abstract

Preeclampsia is one of the leading causes of maternal morbidity, with consequences during and after pregnancy. Because of its diverse clinical presentation, preeclampsia is an adverse pregnancy outcome that is uniquely challenging to predict and manage. In this paper, we developed racial bias-free machine learning models that predict the onset of preeclampsia with severe features or eclampsia at discrete time points in a nulliparous pregnant study cohort. To focus on those most at risk, we selected probands with severe PE (sPE). Those with mild preeclampsia, superimposed preeclampsia, and new onset hypertension were excluded.

The prospective study cohort to which we applied machine learning is the Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-be (nuMoM2b) study, which contains information from eight clinical sites across the US. Maternal serum samples were collected for 1,857 individuals between the first and second trimesters. These patients with serum samples collected are selected as the final cohort.

Our prediction models achieved an AUROC of 0.72 (95% CI, 0.69–0.76), 0.75 (95% CI, 0.71–0.79), and 0.77 (95% CI, 0.74–0.80), respectively, for the three visits. Our initial models were biased toward non-Hispanic black participants with a high predictive equality ratio of 1.31. We corrected this bias and reduced this ratio to 1.14. This lowers the rate of false positives in our predictive model for the non-Hispanic black participants. The exact cause of the bias is still under investigation, but previous studies have recognized PLGF as a potential bias-inducing factor. However, since our model includes various factors that exhibit a positive correlation with PLGF, such as blood pressure measurements and BMI, we have employed an algorithmic approach to disentangle this bias from the model.

The top features of our built model stress the importance of using several tests, particularly for biomarkers (BMI and blood pressure measurements) and ultrasound measurements. Placental analytes (PLGF and Endoglin) were strong predictors for screening for the early onset of preeclampsia with severe features in the first two trimesters.

Keywords Preeclampsia, Machine learning, PlGF, Fairness in machine learning, Preeclampsia with severe features, Ensemble model

*Correspondence:
Yun C. Lin
ycl2112@columbia.edu
Full list of author information is available at the end of the article

Introduction

Preeclampsia (PE) is one of the leading causes of maternal morbidity, with consequences during and after pregnancy [1]. Ensuring optimal patient outcomes requires robust prediction models for PE risk, emphasizing early detection. However, PE poses significant diagnostic and prognostic difficulties given its variable presentations in terms of clinical indications, speed of development, and timing, as well as its unknown causes. PE might evolve slowly and remain mild or quickly present severe complications leading to what is known as PE with severe features (sPE) [1]. Moreover, there are two sub-categories: early onset PE requiring delivery before 34 weeks and late onset after that. While the early onset of PE is associated with a higher incidence of adverse pregnancy outcomes, understanding the relationship between the early and late onset of PE has proven challenging [2, 3]. Some researchers treat them as distinct, but work by Poon et al. [2] treats the condition as a spectrum, best represented by a survival time model. Beyond this, the presence of seizures that cannot be attributed to any other underlying condition in a patient diagnosed with PE would be categorized as Eclampsia (E) [1].

Though a complete understanding of PE still needs to be discovered, rich literature exists on risk factors for and indicators of PE. Biochemical and biophysical markers can have an added benefit for screening for PE when combined with clinical characteristics taken from medical history, demographics, clinical measurements, etc. [2, 4–7]. Research [2, 8–10] has suggested placental growth factor (PlGF), soluble Flt-1 (sFlt-1), pregnancy-associated plasma protein A (PAPP-A), and ultrasound measurements as clinical factors that are significant in signaling an increase in the risk of PE.

Applying this significant volume of knowledge to prediction is pertinent. This study aims to build bias-free machine learning classifiers at various discrete points in pregnancy that combine well-known risk factors for and indicators of sPE and E, which can help screen for cases early in pregnancy in a nulliparous study cohort. While many other studies have focused on predicting preeclampsia in a general population, our study focuses solely on nulliparous patients, making the prediction tasks much more difficult since no prior obstetrical history information is available. Additionally, our study distinguishes itself by using machine learning specifically to build prediction models for severe manifestations.

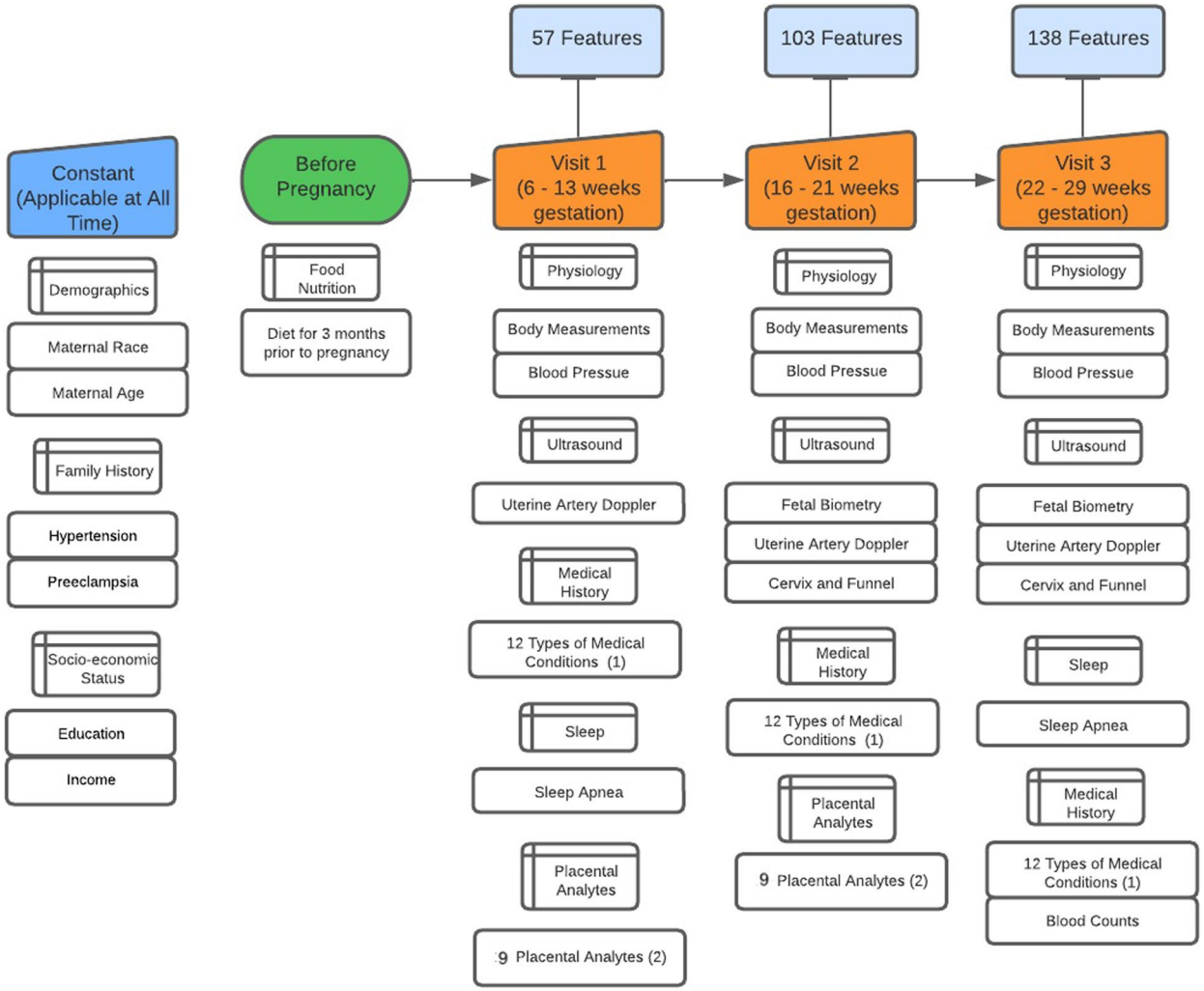
Materials and methods

Study population

The prospective cohort we considered is the Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-be (nuMoM2b) [11], which contains information from eight clinical sites across the US between October 2010 and May 2014. Participants gave written informed consent, and institutional review board approval was obtained at all sites. Maternal race was self-reported by participants. Option for self-reported race include: Non-Hispanic White, Non-Hispanic Black, Hispanic, America Indian, Asian, Native Hawaiian, Other and Multiracial. The study contains a wide array of information collected for nulliparous participants across four visits, $6^0 - 13^6$ (V1), $16^0 - 21^6$ (V2), $22^0 - 29^6$ (V3) weeks gestation and the delivery visit (V4). The predictive model for preeclampsia is constructed to encompass V1 through V3 but excludes V4, as forecasting is deemed less relevant post-delivery.

At V1 and V2, maternal serum was collected, enabling a limited follow-up nuMoM2b sub-study to understand the relationship between placental analytes and a set of adverse pregnancy outcomes (APOs). The sub-study defines all patients with one or more of the following adverse pregnancy outcomes (APOs) as cases: 1) delivery prior to 37 weeks' gestation (PTB); 2) pregnancy complicated by preeclampsia or eclampsia (PE); 3) birth weight for gestational age < 5th percentile (SGA); or, 4) delivery of a stillborn baby (SB). Controls were defined as women with full term live births not noted as complicated by PE or SGA (including missing status for PE and/or SGA). After selecting the cases, a random selection of control patients was conducted to obtain a total of 4,500 serum samples, as allowed by the study budget. In 2016, the maternal sample serum was sent to 2 laboratories. The first lab used 0.5 mL aliquot of the serum to measure four analytes: ADAM-12, endoglin, sFLT-1, VEGF while the second used 0.5 mL aliquots to measure five analytes: AFP, fbHCG, Inhibin A, PAPP-A, PLGF. The aliquots were collected at V1 and V2 of the pregnancy and store at − 70C within two hours following collection. They were measured using lanthanide-based Time Resolved Fluorometry (TRF) and were run using the AutoDELFIA system. The multiple of median (MoM) values of the placental analytes were calculated and used as an input to our model. MoM values are calculated based on the visit that was collected either V1 or V2 This is to ensure no leakages in the model prediction. In the prediction model, every patient is represented as a sample, and various measurements of placental analytes from the same patient are considered distinct features. Figure 1 describes in detail the number and categories of features selected, and Fig. 2 contains a flowchart of the final study cohort selection process.

For the specific features included in our prediction model, please refer to supplement Tables 1–5. Information from the prior visits is also incorporated into the V2 and V3 prediction models. Therefore, the prediction model for V2 was trained on information from V1 and



[1] Specific medical condition can be found in Table Supplement 5

[2] Specific placental analytes can be found in Table Supplement 3

Fig. 1 Data process timeline, This figure shows the gestational weeks at each visit. For each visit, the number of features at that visit is listed and the category of new feature included is also shown

V2. The prediction model for V3 was trained on data collected from V1, V2, and V3. For V1, 57 features were used to train the model, 103 for V2, and 138 for V3.

To focus on those most at risk, we selected probands with severe PE (sPE). Those with mild preeclampsia, superimposed preeclampsia, and new onset hypertension were excluded. There are no cases of fetal demise at < 20 weeks in the final study cohort. We preserved 36 instances of stillbirth, all of which belonged to the no pregnancy-related hypertension (NPH) category. Same cohort is used for each prediction model (V1-V3). In this study cohort, there are 329 sPE cases out of which 5 cases are eclampsia, 71 cases are early-onset sPE and 253 cases are late-onset. All cases of preeclampsia occur after V3, so all predictions are prognostics. Since, multiple measurement are taken for the placental analytes,

Study outcome

The labeling of sPE (severe preeclampsia) was according to the labeling in the nuMoM2b study. Supplement Fig. 1 contains a flowchart indicating the study diagnostic criteria for the primary outcome of this study sPE. sPE is defined as a diagnosis of preeclampsia plus at least one of the criteria in the list: Thrombocytopenia, Pulmonary

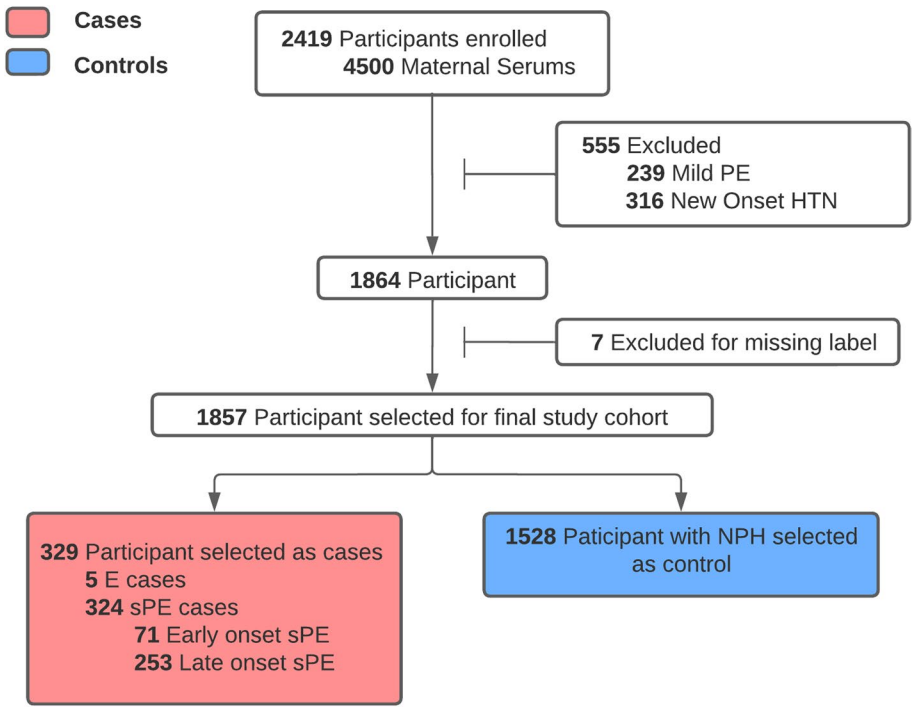


Fig. 2 Final study cohort selection process. Out of the participants from the placental analytes sub-study, we excluded participants with conditions such as new onset hypertension, mild preeclampsia, and missing label for preeclampsia to focus on the participants that are most at risk

edema, Severe hypertension, Proteinuria ≥ 5,000 mg/ 24 h, Oliguria (urine output < 500 mL/ 24 h), Severe headache, Epigastric pain and Fetal growth restriction. The nuMoM2b dataset also contained labels in accordance with the ACOG criteria published in 2013. Initial testing of the proposed pipeline with this ACOG labeling indicated results very similar to that achieved with the nuMoM2b criteria.

PEPrML pipeline

Our PreEclampsia Predictor with Machine Learning (PEPrML) pipeline produces machine learning-capable models that are explainable and trustworthy. Classifiers to predict sPE + E versus NPH and early sPE versus late sPE + E were modeled for every visit. Categorical features were one-hot encoded. The mean imputation was used for continuous features with missingness. We used a cross-validation strategy, a popular approach in machine learning. It consists in splitting the training set into folds, usually 5 folds, learning from 4 folds and validating on the fifth. The process is repeated five times and the average validation error is used to select the best parameters (gridsearch of the possible setting of the parameters, which is method-dependent parameters). Those are used to build the final model that is then applied to the out of sample, that is the test set. After each trial the dataset is shuffled and split in to train and test set. The same process is repeated 100 times resulting in 100 different train and test splits. The results of the test sets were reported. We balanced the ratio of control versus cases by undersampling in the training and test sets, as this introduces less overfitting, leads to a faster training time, and avoids an over-inflated Area Under the ROC curve (AUC). One argument against undersampling is the possibility of removing important examples from the majority class, to mitigate this risk, we conducted one hundred trials to mitigate this issue with different undersampling of the negative majority class, ensuring that important examples were retained [12]. Undersample data contain 50% positive labels and 50% negative labels. 0.5 was selected as the test positivity cut-off for calculating sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). This process is described in detail in Fig. 3.

We experimented with logistic regression (LR), support vector machines (SVM), random forest (RF), and eXtreme Gradient Boosting (XGBoost) [13]. The model build process is repeated each visit from V1 to V3 with the same cohort. LR, SVM, and RF are trained using implementation from Scikit-learn, while XGBoost was trained using implementation from Distributed (Deep) Machine Learning Community [13, 14]. Gridsearch is implemented using Scikit-learn and undersample is implemented using Imbalance-Learn [15]. For RF and

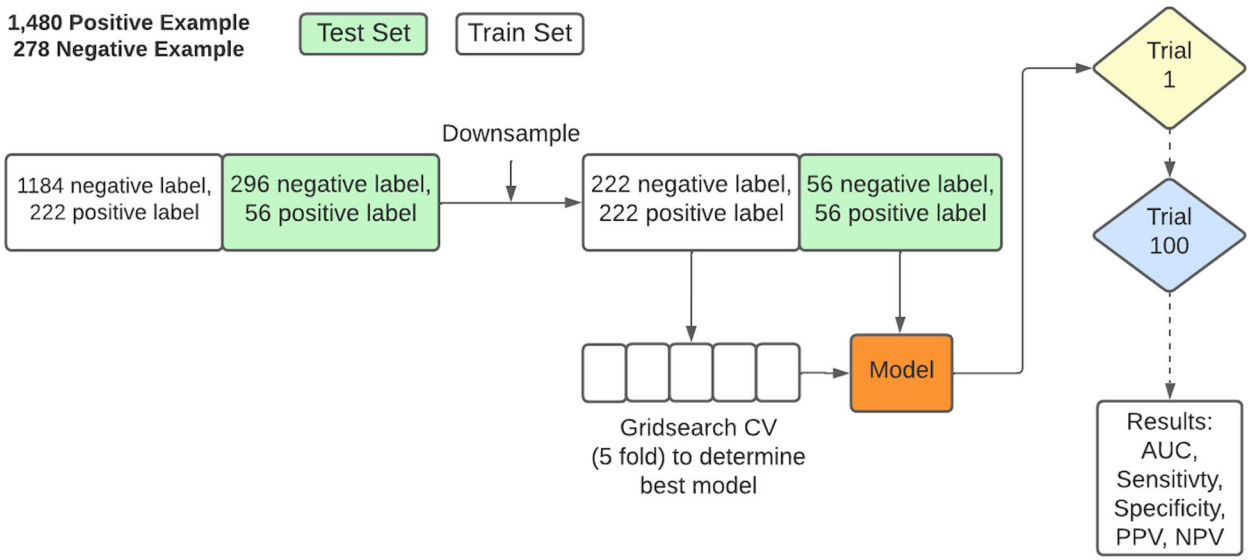


Fig. 3 The training process of PEPrML pipeline. Samples were balanced for train and test sets. fivefold grid search cross-validation was used to select the hyperparameters for each trial. We repeated 100 trials and recorded the results

XGBoost, we extracted the interpretable feature importance rankings, identifying the top factors to generate partial dependence plots (PDPs) [16]. Two ensemble methods (RF and XGBoost) were chosen as classifiers specifically because they are more robust to noise and overfitting, exhibiting a double descent risk curve [17]. In the supplement material, we provide a detailed analysis of this phenomenon. We use Partial Dependence Plots (PDPs) to display the marginal effect that features of interest have on the predicted outcome of a given model, allowing us to advance our understanding of the outcome. For the RF model, we calculated the Equal Opportunity Ratio (EOR) [18], Predictive Parity Ratio (PPR) [19], Predictive Equality Ratio (PER) [18], Accuracy Equality Ratio (AER) [20], and Statistical Parity Ratio (SPR) [20]. We mitigated the race-based biases using Ceteris Paribus Cutoff Plots.

Comparison to other preeclampsia models

It is not possible to directly compare our machine learning model to other works, as most of them do not focus on a nulliparous cohort with severe cases. Therefore, when comparing our model to other cohorts, we need to adapt their models to our study cohort. Both Poon et al. and Akolerkar et al. employed Bayes' theorem to construct their models, a methodology we implemented using the Scikit-survival package [21]. All models were assessed at V1. Poon et al. utilized features derived from maternal risk factors, MAP, PlGF, uterine artery pulsatility index, and PAPP-A. In contrast, Akolerkar et al. incorporated all the features used by Poon et al., along with additional placental analyte features.

Fairness metrics

To determine the fairness of our model, we identified and calculated a set of metric ratios as discussed below to determine the subgroup(s) for which the threshold of the four fifths rule was violated [22]. We then plotted a *ceteris paribus* cutoff plot for the subgroup affected/impacted by the bias and adjusted the classification threshold accordingly. Fairness metrics create a framework for measuring discrimination based on characteristic attributes such as race, gender, and ethnicity. To better understand fairness metrics, we assume a source distribution (Y, X, A), where Y is the target label, X is the set of available features and $A \in \{0, 1\}$ is the characteristic attribute. For the sake of simplicity, we define A as a binary attribute, but A can have many labels, such as race. A predictor of Y is defined as $\widehat{Y} = f(X, A)$. For a model to be deemed fair, it should be the case that the learned predictor does not discriminate with respect to A [18]. In a practical setting, ratios of fairness metrics are calculated for protected groups and the privileged group. A model is considered unfair when such a ratio crosses a threshold specified by the four fifths rule. supplement Table 7 summarizes the fairness metrics and their respective mapping to the evaluation metrics used in our analysis.

In our work, we selected the Predictive Equality Ratio (PER) as the primary fairness metrics. PER requires $\widehat{Y}$ to have equal false positive rate across the two subgroups $A = 1$ and $A = 0$:

$$Pr\{\widehat{Y} = 1|A = 0, Y = 0\} = Pr\{\widehat{Y} = 1|A = 1, Y = 0\}$$

The fairness evaluation is conducted using the Python package Dalex: Responsible Machine Learning in Python [23]. Fairness metric ratios were calculated over multiple trials and averaged. In our analysis, the focus is on the characteristic attribute of race. For further evaluation, we will denote $A \in \{Asian, Black, Hispanic, OtherWhite\}$ as the race attribute. We denote the Non-Hispanic White participants ($A = White$) as the privileged group, while all other races are considered as protected groups.

For i$\in A \setminus \{White\}$, given FPR_i is the false positive rate of race i and FPR_{White} is the fa;se positive rate of privileged subgroup, the predictive equality ratio of race i (PER_i) is calculated by:

$$PER_i = \frac{FPR_i}{FPR_{White}}$$

Ceteris Paribus Cutoff Plot

We wanted to measure of how biased our model is toward a certain race, but also a measure that summarizes the bias across different races. This can be done through calculating the parity loss. Parity loss of PER can be calculated by:

$$Parityloss\ of\ PER = \sum_{i \in A \setminus \{White\}} \left| log \frac{FPR_i}{FPR_{Whjte}} \right| = \sum_{i \in A \setminus \{White\}} \left| log PER_i \right|$$

To mitigate the bias identified through parity loss, a post-processing step using the *ceteris paribus* cutoff plotting was used. This is a model-agnostic algorithm, which works for models with different structures, such as neural networks, random forests, boosting models, and linear models [23]. Similar to the process of constructing a ROC curve, ceteris paribus cutoff plotting works by directly altering the classification threshold, but instead of visualizing the impact on TPR and FPR, the parity loss is used, and alteration of the classification threshold is only performed on a specific subgroup. Typically, we select the threshold by identifying the value that minimizes the parity loss.

Software packages

We developed our pipeline in Python 3. Instructions about how to run the experiments are provided in the Github repository. We also conducted bias mitigation experiments using the Dalex package [23]. Dataset balancing was done using the imbalanced-learn package. The model used to generate our results was trained using the XGBoost package.

The underlying code for this study is available in PRAISE-Lab repository and can be accessed via this link: https://github.com/PRAISE-Lab-Repository/PEPrML.git

We added synthetic data with similar structure (features and visits) to mimic nuMoM2b. While the synthetic data features are indeed fictitious, the feature values are sampled from their respective distributions in nuMoM2b according to the labels assigned to the synthetic data. Models can be built for the synthetic data by following the instructions from GitHub.

Result

Study population characteristics

1,857 participants were selected as the final study cohort. Among these, 5 developed E and 324 developed sPE, of which 71 (~22%) were early onset (< 34 weeks), and 253 (~78%) were late onset. The remaining 1,528 patients were NPH (Fig. 2). Participants had a median age of 27 and IQR of 9; 3.3% were Asian, 17.6% were Hispanic, 57.5% non-Hispanic white, 15.9% non-Hispanic black, and 5.7% were of other races or multiracial.

Significant characteristics ($P < 0.001$) among the sPE + E participants versus NPH include a higher mean value for body mass index (BMI) (27 kg/m2 vs. 24 kg/m2), systolic blood pressure (SBP) (112 mmHg vs. 108 mmHg) at V1, diastolic blood pressure (70 mmHg vs. 66 mmHg) at V1, a lower mean value for PlGF (0.92 vs. 1.00) and PAPP-A (0.86 vs. 1.00) at V1. Other significant placental analytes include Endoglin at V2, Inhibin A at V2, VEGF at V1, PLGF at V2, and the sFLT1-to-PLGF ratio at V2. Systolic and diastolic blood pressure are significant for all visits. Weight, neck circumference, and waist circumference information are collected only for V1 and are also significant. Significant ultrasound information includes Early Diastolic Notch, Resistance Index, Pulsatility Index, and Systolic/Diastolic Ratio for both right and left uterine arteries for V2 and V3. Significant medical conditions include diabetes for V1 and V2, hypertension (not including chronic hypertension) for all three visits, and chronic hypertension for V1.

Some significant characteristics ($P < 0.001$) among the early sPE participants versus late sPE + E include a lower mean value for PlGF (0.73 vs. 0.98). Other important placental analytes include Inhibin A for V2, PLGF for V2, and the sFLT1-to-PLGF ratio for V2. Income is also significant. Significant ultrasound information includes Early Diastolic Notch, Resistance Index, Pulsatility Index,

and Systolic/Diastolic Ratio for both right and left uterine arteries for V2 and V3. The measure from the ultrasound of baby's abdominal circumference is also significant for V3. Diastolic blood pressure is significant for V3. The only significant medical condition is kidney disease for V2. For a detailed summary of statistics of all features, please refer to supplement Tables 1—5.

Model performance

A summary of performance results for sPE+E versus NPH can be found in Fig. 4. Results in Fig. 4a indicate that predictive capabilities increase as additional information is added at V2 and V3. RF models achieved an AUC of 0.72 (95% CI, 0.69–0.76) at V1, 0.75 (95% CI, 0.71–0.79) at V2, and 0.77 (95% CI, 0.74–0.80) at V3. Welch's t-test was conducted for each pair of classifiers. RF model performance is significantly different (< 0.001) for all visits compared to LR and SVM, while only significantly different to XGboost at V1. Detailed measures for RF and other comparison methods can be found in Table 1. Further performance breakdown is offered in supplement Table 6 and supplement Table 7, which

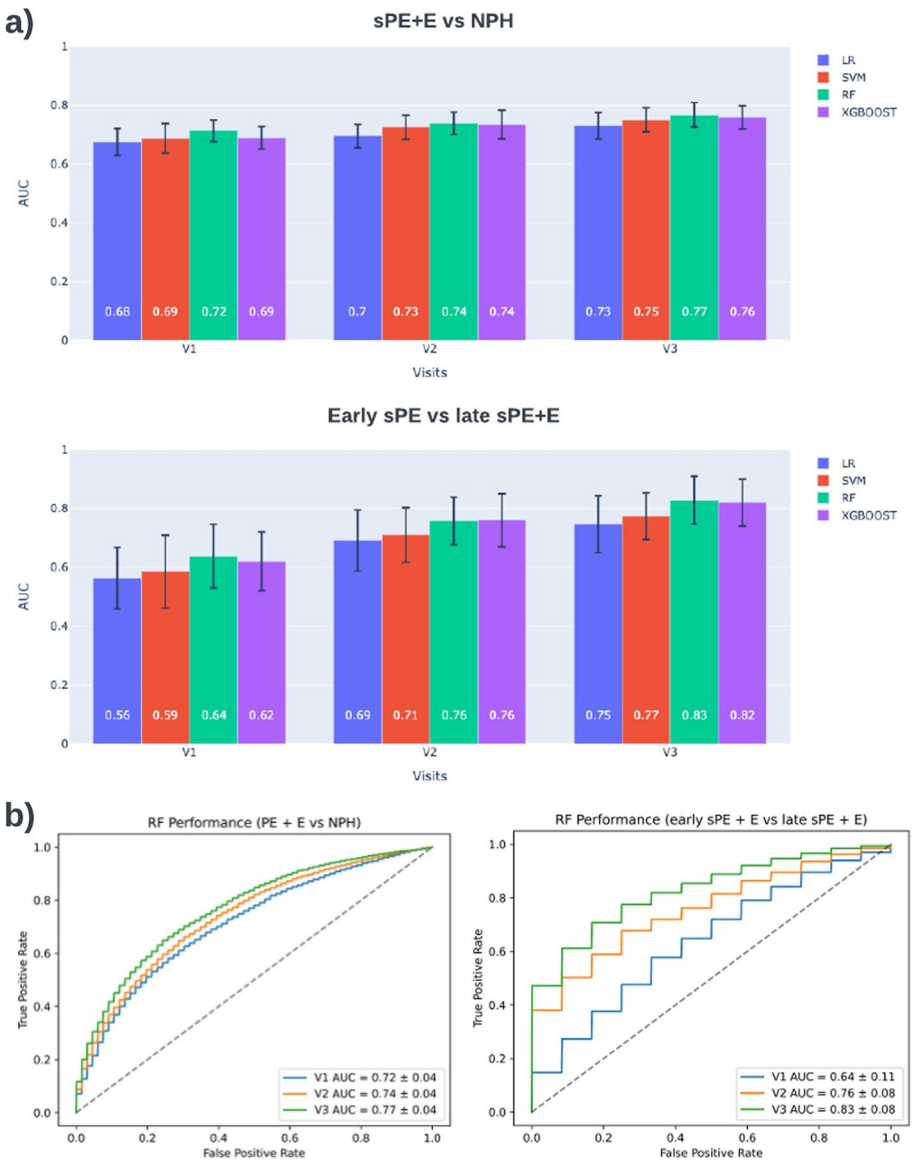


Fig. 4 sPE + E vs NPH and early sPE vs late sPE + E model performance. **a** Average AUC for 100 trials per visit for 4 classifiers. **b** RF classifier has best performance across visits for both comparisons. The ROC curve demonstrated the tradeoff between the true positive rate versus false positive rate. This summarizes the results for 100 trials

Table 1 Detailed summary of sPE + E vs NPH model performance per visit for four classifiers

Model	Visits	AUC	Sensitivity (TP/TP + FN)	Specificity (TN/FP + TN)	PPV (TP/TP + FP)	NPV (TN/FN + TN)
LR	V1	0.68 ± 0.05	0.59 ± 0.06	0.67 ± 0.06	0.64 ± 0.05	0.62 ± 0.04
	V2	0.70 ± 0.04	0.63 ± 0.05	0.67 ± 0.05	0.65 ± 0.04	0.64 ± 0.03
	V3	0.73 ± 0.04	0.65 ± 0.07	0.69 ± 0.05	0.68 ± 0.04	0.67 ± 0.05
SVM	V1	0.69 ± 0.05	0.57 ± 0.06	0.70 ± 0.06	0.66 ± 0.05	0.62 ± 0.04
	V2	0.73 ± 0.04	0.59 ± 0.06	0.74 ± 0.05	0.70 ± 0.05	0.65 ± 0.04
	V3	0.75 ± 0.04	0.60 ± 0.06	0.77 ± 0.05	0.72 ± 0.05	0.66 ± 0.03
RF	V1	0.72 ± 0.04	0.64 ± 0.06	0.68 ± 0.06	0.67 ± 0.04	0.65 ± 0.04
	V2	0.74 ± 0.04	0.65 ± 0.06	0.70 ± 0.05	0.68 ± 0.04	0.67 ± 0.04
	V3	0.77 ± 0.04	0.69 ± 0.06	0.71 ± 0.06	0.70 ± 0.05	0.69 ± 0.05
XGBoost	V1	0.69 ± 0.04	0.62 ± 0.05	0.66 ± 0.06	0.65 ± 0.04	0.63 ± 0.04
	V2	0.74 ± 0.05	0.66 ± 0.07	0.69 ± 0.06	0.68 ± 0.05	0.67 ± 0.05
	V3	0.76 ± 0.04	0.67 ± 0.07	0.71 ± 0.05	0.70 ± 0.04	0.69 ± 0.04

For V1, 57 features were used to train the model, 103 for V2, and 138 for V3. Detail of features used can be seen in supplement Table 1–5

summarize results for predicting early and late onset, versus NPH, respectively.

The model's predictive power for early onset preeclampsia is higher than for late onset, as demonstrated by the two tables. Across the board, all metrics have higher values, but the variance is also higher for these values, most likely due to the smaller set of cases with early onset sPE. We modeled classifiers to directly predict early sPE vs. late sPE + E to understand better what enabled this performance. A summary of performance results for early sPE vs. late sPE + E can be found in Fig. 4. Again, performance increased with gestational age, and RF models performed the best, obtaining an AUC of 0.64 (95% CI, 0.53 – 0.75) at V1, 0.76 (95% CI, 0.68 – 0.82) at V2, and 0.83 (95% CI, 0.75 – 0.91) at V3. Detailed performance measures for RF and other comparison methods can be found in supplement Table 8.

Interpreting sPE + E vs NPH model

The feature importance lists for V1, V2, and V3, where the prediction task is prognosis, are given in supplement Fig. 2, Fig. 5a, and supplement Fig. 3, respectively, enabling a better understanding of the key features that contribute to the RF and XGBoost decision processes. For V1, the top 5 features are BMI, mean arterial pressure (MAP), SBP, waist circumference, and endoglin. For V2, the top five features are BMI, PlGF (V2), MAP (V2, V1), and SBP (V2, V1). For V3, the top five features are MAP (V3, V1), SBP (V2), BMI, and PLGF(V2).

The PDP for BMI shown in Fig. 5c indicates a risk increase in sPE + E at around 22.41 kg/m^2 and at the peaks at 35 kg/m^2. We see a substantial increase in the risk of sPE + E with a systolic reading of 110 mmHg or higher, and by Visit 2, this number drops to 102 mmHg (supplement Fig. 4.a). The diastolic reading did not exhibit such a pronounced increase in the risk of sPE + E, but we did observe a slight increase above 78 mmHg. Looking at the MAP at Visit 1, supplement Fig. 4.b, we see an increase in risk at 82.67 mmHg. There is a sharp increase in the predicted risk for sPE + E observed in the PDP for PlGF at Visit 1 for MoM measurements less than 1.5.

Racial fairness in sPE + E vs NPH model

Our model for predicting sPE + E vs. NPH is biased mainly against non-Hispanic Black participants. Using the White race as the reference race, we identified that the predictive equality ratio for non-Hispanic Black participants (1.31) is high, according to the four-fifths rule.

To address this problem, we created a *ceteris paribus* Cutoff plot of the parity loss for the non-Hispanic Black sub-population to determine the optimal confidence threshold for prediction. Adjusting the threshold from 0.5 to 0.54 accordingly mitigated the over-prediction of PE occurrence by our model for non-Hispanic Black participants, reducing the predictive equality ratio for non-Hispanic Black participants from 1.31 to 1.14 (Fig. 6).

Discussion

The results presented here demonstrate that it is possible to learn RF models with superior, well-rounded performance for early prediction of preeclampsia at multiple time points throughout pregnancy, with minimal preprocessing of data, feature engineering, or feature selection. Exhibiting a relatively balanced score for PPV and Sensitivity, RF increases performance by all metrics at each new visit as more information becomes available. The feature importance plots confirm existing knowledge

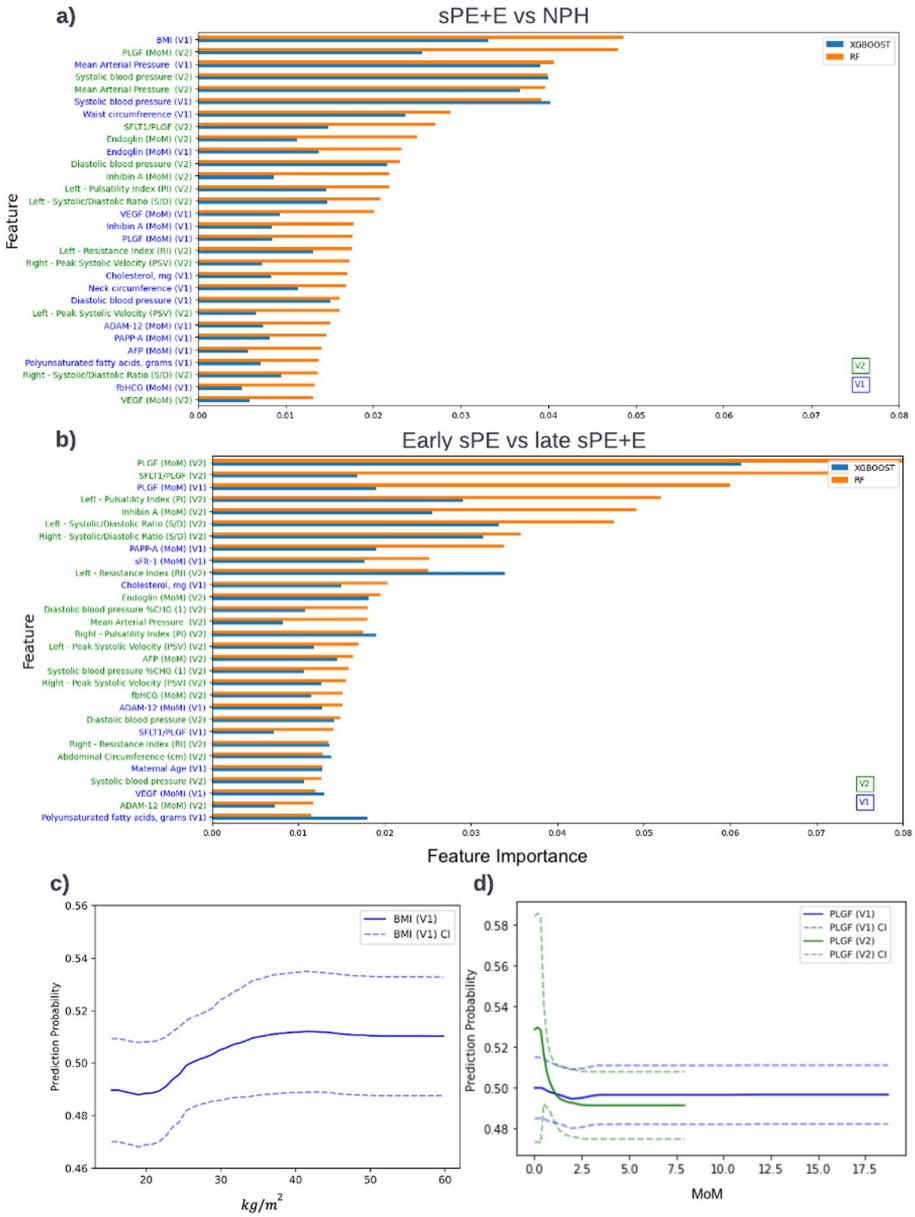


Fig. 5 Interpreting machine learning model for sPE + E vs NPH and early sPE vs late sPE + E. **a** V2 model features importance for sPE + E vs NPH model, **b** V2 model feature importance for early sPE vs late sPE + E, **c**—**d** PDP for BMI and PlGF based on model build for sPE + E vs NPH. V2 model includes both features from V1 and V2

about known predictive features such as blood pressure, uterine artery blood flow, and placental analytes and identify features not commonly referenced in the prediction literature, such as Endoglin, Cholesterol, and Inhibin A. Inclusion of the placental analytes are a great contribution to the prediction of preeclampsia, especially in the early phases of pregnancy. An analysis of model performance without the placental analytes and just traditional risk factors shows a decrease in performance of up to 7% for the V1 model (supplement Table 12). Review of RF fairness metrics indicated a correctable bias against non-Hispanic Black participants.

Our study confirmed that blood pressure and placental analytes were significant in predicting PE across study visits [24–26]. The results of our statistical tests deviate from other works [2, 10, 27] in that risk factors such as maternal age, race, sleep apnea, and family history of PE were not significant. Socio-economic status did not contribute to the prediction of preeclampsia in our study cohort, as suggested by other studies such as Arechvo

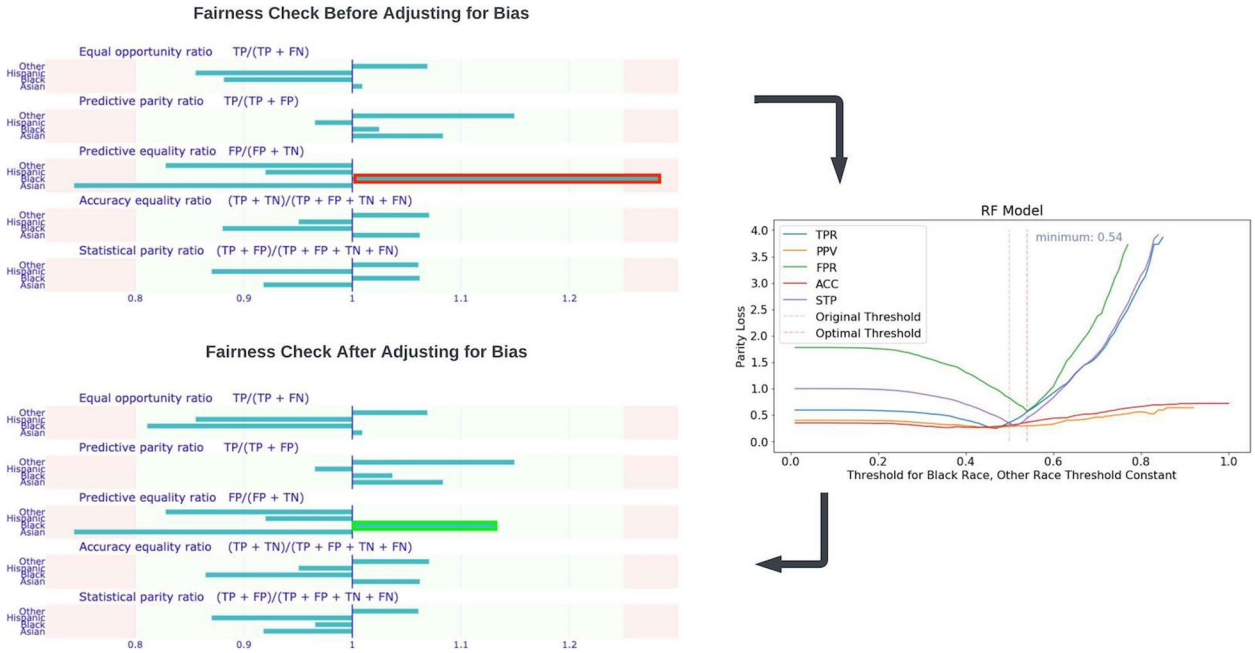


Fig. 6 Fairness check for sPE + E vs NPH mode. The threshold set based on the four-fifth rule is 0.8 and 1.25. A *Ceribus Paribus* plot was used to adjust the prediction threshold for the Black population

et al. [24]. Thus, care must be taken in comparing the model performance presented here for the nuMoM2b dataset with other studies, given that the nuMoM2b dataset characterizes demographically diverse nulliparous mothers with unknown risk for PE at the time of first prediction while the target label is strictly focused on sPE + E criteria.

Our selected predictors in the first trimester of pregnancy are like those used by previously published competing risk models from Akolekar et al., Poon et al., and O'Gorman et al. [28–30], but our study contains more features and focuses solely on a nulliparous study cohort. To compare our results to these two prior studies, we reconstructed their experiment using our nulliparous cohort and features from V1. We found that our model yielded better outcomes across the board. In Table 2, our model performance, on average, has a 3–4% higher AUC for V1 and reaches to 12% for V3. While Poon et al. [29] report a 91% AUC for preterm PE and 78% AUC for predicting term PE just by utilizing features such as maternal risk factors, MAP, PlGF, uterine artery pulsatility index, and PAPP-A, we did not observe this high AUC in our prediction model. This might be attributed to the fact that our prediction task focuses on PE with severe features for nulliparous women only, which makes the prediction tasks much more difficult.

Ensemble methods, specifically RF and XGBoost [31], are the top performers in our study. Other studies have shown ensemble methods to have a strong predictive power for preeclampsia [32–34]. This may be due to the

Table 2 Our model versus other models

	AUC	Sensitivity (TP/TP + FN)	Specificity (TN/FP + TN)	PPV (TP/TP + FP)	NPV (TN/FN + TN)
Poon et al	0.68 ± 0.04	0.62 ± 0.05	0.67 ± 0.07	0.65 ± 0.05	0.64 ± 0.04
Akolerkar et al	0.69 ± 0.05	0.63 ± 0.07	0.67 ± 0.06	0.66 ± 0.04	0.65 ± 0.04
PEPrML V1 (Our Model)	0.72 ± 0.04	0.64 ± 0.06	0.68 ± 0.06	0.67 ± 0.04	0.65 ± 0.04
PEPrML V2 (Our Model)	0.74 ± 0.04	0.65 ± 0.06	0.70 ± 0.05	0.68 ± 0.04	0.67 ± 0.04
PEPrML V3 (Our Model)	0.77 ± 0.04	0.69 ± 0.06	0.71 ± 0.06	0.70 ± 0.05	0.69 ± 0.05

All models were evaluated at V1. Poon et al. utilized features derived from maternal risk factors, MAP, PlGF, uterine artery pulsatility index, and PAPP-A. On the other hand, Akolerkar et al. incorporated all the features used by Poon et al. as well as additional placental analytes features

ensemble nature and the ability of the underlying model, decision trees, to capture some of the subtle distinctions between the varied and poorly understood subgroups of preeclampsia patients [35]. Our hypothesis on why RF models are better than XGBoost is that RF is more suited for datasets that are smaller and XGBoost shines on larger datasets, since boosting work by iteratively improving on the prediction from the prior build tree, while RF involves a majority voting. The PDP for BMI, a well-known risk factor for PE, shown in Fig. 5c indicates a risk increase in PE around 22.41 kg/m^2 and at the peaks at 35 kg/m^2. One possible rationale is that the effect of magnesium circulation is reduced when the BMI is at 35 kg/m^2, since a good magnesium circulation can significantly reduce the risk of eclampsia or convulsions [36]. Furthermore, PDPs for various placental analytes indicate that a decreased level of PlGF during the first and second trimesters precede the onset of PE [2, 37, 38]. Agrawal et al. [39] found that the predictive value was highest for PlGF levels between 80 and 120 pg/mL, which coincides with the sharp increase in the predictive risk for PE observed in the PDP for PlGF at Visit 1 for measurements less than 100 pg/mL (1.5 MoM). MacDonald et al. [40] suggested a sFlt-1:PlGF ratio > 33.4 which agrees with our PDP in supplement Fig. 5. Levine et al. [41] found that endoglin levels at 25 through 28 weeks of gestation were significantly higher (8.5 ng/mL) in term PE patients. We observe this same cut-off value in the PDP in supplement Fig. 4.c, which shows a pronounced increase in the risk of PE at around 9 ng/mL at V1, albeit occurring much earlier, at 6–13 weeks of gestation. Analytes such as PlGF, unlike blood pressure, were consistently important across the sPE + E vs. NPH model and the early vs. late model (Fig. 5), indicating their predictive power, particularly their ability to *rule out* early onset [4, 31].

Implications

This study demonstrates the utility of early and multiple time points screening for PE. It shows that early blood pressure measurement can be a proxy for the risk of high blood pressure later in pregnancy. Also, information about placental analytes, which can be gathered at a reasonable cost tradeoff between assessment and hospitalization [4], allows predictions that enormously surpass the accuracy of a model based only on ACOG guidelines [42]. Further validation is required for the proposed separate models for multiple time points to ensure prediction consistency: a patient identified as high risk early in pregnancy should not be deemed low risk later without sufficient explanation. Also, identifying women at increased risk in the first trimester allows for timely prophylaxis with low-dose aspirin, which is highly effective in preventing preterm disease [43]. In order to evaluate the practicality of our model, we conducted experiments that involved testing our model on a general cohort that included previously excluded patients with mild preeclampsia and gestational hypertension. In comparison to the sPE vs NPH model, for RF the AUC decreased as follows: from 0.72 to 0.66 at V1, from 0.74 to 0.68 at V2, and from 0.77 to 0.70 at V3. This decrease can mainly be attributed to the inclusion of patients with mild preeclampsia, which is evident from the recall of the random forest (RF) model, but the recall still is on par with the sPE vs NPH model. However, it is noteworthy that the recall still remains comparable to that of the sPE vs NPH model for the sPE cases. Our random forest model for all PE cases achieved an AUC of 0.70 at V3, indicating a reasonably good level of prediction. Furthermore, when evaluating the recall metric, we observed that the model performed better in predicting severe cases of preeclampsia than predictions for mild preeclampsia. This finding highlights the model's ability to effectively identify and differentiate severe cases, which is particularly important in clinical decision-making.

Fairness metrics and analysis of causes for biases should become standard practice in model validation. We hypothesize that the limited sample size may have caused the bias against the non-Hispanic Black participants, given cohort skew towards White participants and the potentially inappropriate higher representation of the non-Hispanic Black population among the sPE + E class than the NPH class (20.9% vs. 13.8%, respectively). However, after correcting for this imbalance, the bias still persisted. We then hypothesize that this bias might come from a difference in the distribution of values for the top placental analytes, as suggested in another study [44]. We did observe significant differences in the distribution of top predictive features ($P < 0.001$), such as BMI and PLGF (V1, V2). Due to the correlation between some top features, we cannot simply normalize each by race. Therefore, adjusting the predictive threshold for the Black population is still an efficient way to reduce bias. While the cost of a false negative diagnosis for maternal and fetal health is very high, the stress, fees, and possibly inappropriate treatment of a false positive should not be ignored.

Distinguishing between sPE + E and NPH is critical, but the binary labels pose a challenge. The former group undoubtedly contains different subgroups and phenotypes of preeclampsia, and learning to make these distinctions will have the dual benefit of enhancing our understanding of preeclampsia and allowing for better predictive performance. Thus, moving beyond the initial literature-inspired feature set to a broader set of features will be the target of future work. Furthermore, temporal features capturing change between clinical

measurements at different visits will be investigated, as this may enhance prediction quality at the second and third time points [32]. This would enable more timely monitoring and treatment of late onset preeclampsia.

A more significant departure will involve re-framing the prediction task. Compelling arguments have been made that preeclampsia is best interpreted as a syndrome rather than a disease [31, 45]. Label difficulties have led at least one study of short term preeclampsia screening to focus on a label that consists of the presence, or not, of at least one of multiple adverse maternal or fetal outcomes [31].

Limitations

A set of features identified in the related medical literature was employed for this initial study, but this can be expanded without issue. Using the nuMoM2b data represents an exciting opportunity to learn from a sizable sample of U.S. mothers that is more diverse than other similar studies and that has been captured in a longitudinal study with a considerable number of features [3, 31, 46]. The occurrence rate of PE in this study was consistent with reported rates [4, 47]. However, this meant that even with such a sizable sample, the analysis was limited to more than a couple of hundred sPE + E cases. The substudy also had limitations: analytes were only available for V1 and V2. Our study only applies to the nulliparous population within the US. Therefore, our models do not take previous obstetric history into account.

One noticeable limitation of the study is the limited cases of existing medical conditions in participants of the placental analytes sub-study. This low presence can cause the model to attribute less importance to these risk factors, while these could be crucial in clinical practice. Lastly, our study only focuses on comparing patients with sPE + E and NPH, without addressing those patients who developed PE with mild features, or only hypertension. The scope of the paper is focused on those with severe conditions, as these cases are of greater clinical concern for practitioners involved in treatment decisions. However, to provide insight into the generalizability of our model for broader preeclampsia conditions, we evaluated all cases of preeclampsia versus those without, as reflected in the supplement Table 13.

nuMoM2b is a comprehensive cohort comprising an extensive array of features available for a nulliparous population. This dataset has a substantial number of patients across the United States. We believe that the AUC achieved through our model demonstrates the limits of predictive power when combining a multitude of factors previously acknowledged but always analyzed individually. However, we posit that enhancing the predictive model could be achieved by increasing the sample size and incorporating data related to metagenomics and the microbiome, potentially bolstering its predictive capacity. A possible source for future work would be looking at the CDC Natality cohort, with births ranging from 1968 up until 2021. This data was previously analyzed for trends related to preterm birth (PTB), but adjusting the target outcome to sPE + E could be an appropriate line of research to explore in the future, given the significant sample size [48].

Conclusions

Our analysis suggests that it is important and possible to create screening models to predict the participants at risk of developing preeclampsia with severe features and eclampsia for a nulliparous study cohort. The top features stress the importance of using several tests, in particular tests for biomarkers and ultrasound measurements. The models could potentially be used as a screening tool as early as 6–13 weeks gestation to help clinicians screen for and identify participants who may subsequently develop preeclampsia, confirming the cases they suspect or identifying unsuspected cases. The proposed approach is easily adaptable to address any adverse pregnancy outcome with fairness.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12884-024-06988-w.

Supplementary Material 1.

Acknowledgements

The nuMoM2b Study (Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-Be) was supported by grant funding from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) grant number R01LM013327. The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript. This work also was awarded innovation and health disparities prizes in NICHD's Decoding Maternal Morbidity Data Challenge.https://www.nichd.nih.gov/research/supported/challenges/decoding-maternal-morbidity.

Authors' contribution

The numom2b dataset was analyzed and interpreted by YCL, who also constructed the experiments for the prediction models. DM played a significant role in writing and editing the manuscript, while ACS conducted the analysis on the PDP plot. All authors contributed to the experimental design and reviewed and approved the final manuscript.

Data availability

The data that support the findings of this study are available from NIH Data and Specimen Hub, but restrictions apply to the availability of these data, which were used under licence for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of NIH Data and Specimen Hub. Please contact the corresponding author (Yun Chao Lin) for data involved in this study via email at adam.lin@columbia.edu. https://dash.nichd.nih.gov/explore/study?q=numom2b

Declarations

Ethics approval and consent to participate

Human subjects approval for this study, titled "SCH: Prediction of Preterm Birth in Nulliparous Women", was obtained following review by Columbia University Human Subjects Institutional Review Board, and the City University of New York CUNY Institutional Review Board.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Computer Science, Columbia University, 1214 Amsterdam Ave, 721 Schapiro CEPSR, New York, NY 10027, USA. [2]Department of Computer Science, CUNY Hunter College, New York, NY, USA. [3]Department of Obstetrics and Gynecology, Columbia University, New York, NY, USA. [4]Department of Obstetrics and Gynecology, Indiana University School of Medicine, Indianapolis, IN, USA.

Received: 28 February 2024 Accepted: 15 November 2024
Published online: 24 December 2024

References

1. Lockwood CJ, Moore T, Copel J, Silver RM, Resnik R, Dugoff L, Louis J. Creasy and Resnik's Maternal-Fetal Medicine: Principles and Practice. Philadelphia: Elsevier. 2023;45:826–54.
2. Poon LC, Nicolaides KH. Early prediction of preeclampsia. Obstet Gynecol Int. 2014;2014:297397.
3. Wójtowicz A, Zembala-Szczerba M, Babczyk D, Kołodziejczyk-Pietruszka M, Lewaczyńska O, Huras H. Early- and late-onset preeclampsia: a comprehensive cohort study of laboratory and clinical findings according to the New ISHHP Criteria. Int J Hypertens. 2019;2019:1–9.
4. Sroka D, Verlohren S. Short Term Prediction of Preeclampsia, 2021.
5. Facco FL, Lappen J, Lim C, Zee PC, Grobman WA. Preeclampsia and sleep-disordered breathing: a case-control study. Pregnancy Hypertens. 2013;3:133–9.
6. Eskild A, Vatten LJ. Abnormal bleeding associated with preeclampsia: a population study of 315,085 pregnancies. Acta Obstet Gynecol Scand. 2009;88:154–8.
7. Conde-Agudelo A, Villar J, Lindheimer M. Maternal infection and risk of preeclampsia: systematic review and metaanalysis. Am J Obstet Gynecol. 2008;198:7–22.
8. Fox R, Kitt J, Leeson P, Aye CYL, Lewandowski AJ. Preeclampsia: risk factors, diagnosis, management, and the cardiovascular impact on the offspring. J Clin Med. 2019;8:1625.
9. Karumanchi SA, Epstein FH. Placental ischemia and soluble fms-like tyrosine kinase 1: cause or consequence of preeclampsia? Kidney Int. 2007;71(10):959–61.
10. Verlohren S, Herraiz I, Lapaire O, et al. New gestational phase-specific cutoff values for the use of the soluble fms-like tyrosine kinase-1/placental growth factor ratio as a diagnostic test for preeclampsia. Hypertension. 2014;63:346–52.
11. Haas DM, Parker CB, Wing DA, et al. A description of the methods of the nulliparous pregnancy outcomes study: monitoring mothers-to-be (nuMoM2b). Am J Obstet Gynecol. 2015;212:539-e1.
12. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016;5(4):221–32.
13. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Paper presented at: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016.
14. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
15. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res. 2017;18(1):559–63.
16. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat. 2015;24(1):44–65.
17. Belkin M, Hsu D, Ma S, Mandal S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proc Natl Acad Sci. 2019;116(32):15849–54.
18. Hardt M, Price E, Price E, Srebro N. Equality of Opportunity in Supervised Learning. Vol. 29, Neural Information Processing Systems. Curran Associates, Inc.; 2016. Available from: https://papers.nips.cc/paper_files/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html.
19. Chouldechova A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data. 2017;5:153–63.
20. Verma S, Rubin J. Fairness definitions explained. In Proceedings of the international workshop on software fairness. 2018:1–7.
21. Pölsterl S. scikit-survival: a library for time-to-event analysis built on top of scikit-learn. J Mach Learn Res. 2020;21(1):8747–52.
22. Bobko P, Roth PL. The four-fifths rule for assessing adverse impact: An arithmetic, intuitive, and logical analysis of the rule and implications for future research and practice: Research in personnel and human resources management. Leeds, UK: Emerald Group Publishing Limited; 2004.
23. Baniecki H, Kretowicz W, Piatyszek P, Wisniewski J, Biecek P. dalex: Responsible Machine Learning with Interactive Explainability and Fairness in Python. J Mach Learn Res. 2021;22:1–7.
24. Sibai BM, Ewell M, Levine RJ, et al. Risk factors associated with preeclampsia in healthy nulliparous women. Am J Obstet Gynecol. 1997;177:1003–10.
25. Smith GCS, Stenhouse EJ, Crossley JA, Aitken DA, Cameron AD, Connor JM. Early pregnancy levels of pregnancy-associated plasma protein a and the risk of intrauterine growth restriction, premature birth, preeclampsia, and stillbirth. J Clin Endocrinol Metab. 2002;87(4):1762–7.
26. McLaughlin K, Snelgrove JW, Audette MC, et al. PlGF (Placental Growth Factor) Testing in Clinical Practice: Evidence From a Canadian Tertiary Maternity Referral Center. Hypertension. 2021;77(6):2057–65.
27. Phan K, Pamidi S, Gomez YH, et al. Sleep-disordered breathing in high-risk pregnancies is associated with elevated arterial stiffness and increased risk for preeclampsia. Am J Obstet Gynecol. 2022;226(6):833.e1-833.e20. https://doi.org/10.1016/j.ajog.2021.11.1366.
28. Akolekar R, Syngelaki A, Sarquis R, Zvanca M, Nicolaides KH. Prediction of early, intermediate and late pre-eclampsia from maternal factors, biophysical and biochemical markers at 11–13 weeks. Prenat Diagn. 2011;31(1):66–74.
29. Poon LC, Shennan A, Hyett JA, Kapur A, Hadar E, Divakar H, McAuliffe F, da Silva CF, von Dadelszen P, McIntyre HD, Kihara AB. The International Federation of Gynecology and Obstetrics (FIGO) initiative on preeclampsia (PE): a pragmatic guide for first trimester screening and prevention. Int J Gynaecol Obstet. 2019;145(Suppl 1):1.
30. 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(1):103-e1.
31. Couronné R, Probst P, Boulesteix AL. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics. 2018;19:1–14.
32. Jhee JH, Lee S, Park Y, et al. Prediction model development of late-onset preeclampsia using machine learning-based methods. PLoS One. 2019;14:e0221202.
33. Marić I, Tsur A, Aghaeepour N, et al. Early prediction of preeclampsia via machine learning. Am J Obstet Gynecol MFM. 2020;2:100100.
34. Schmidt MLJ, Rieger MO, Neznansky MM, et al. A machine-learning based algorithm improves prediction of preeclampsia-associated adverse outcomes. Am J Obstet Gynecol. 2022;227:77.e1-77.e30.
35. Myatt L. The prediction of preeclampsia: the way forward. Am J Obstet Gynecol. 2022;226:S1102-S1107.e8.
36. Gestational Hypertension and Preeclampsia. ACOG Practice Bulletin, Number 222. Obstet Gynecol. 2020;135(6):e237–60. https://doi.org/10.1097/AOG.0000000000003891.
37. Su Y, Lee CN, Cheng WF, Shau WY, Chow SN, Hsieh FJ. Decreased maternal serum placenta growth factor in early second trimester and preeclampsia. Obstet Gynecol. 2001;97:898–904.

38. Tidwell S, Ho HN, Chiu WH, Torry RJ, Torry DS. Low maternal serum levels of placenta growth factor as an antecedent of clinical preeclampsia. Am J Obstet Gynecol. 2001;184(6):1267–72.
39. Agrawal S, Shinar S, Cerdeira AS, Redman C, Vatish M. Predictive Performance of PlGF (Placental Growth Factor) for Screening Preeclampsia in Asymptomatic Women. Hypertension. 2019;74(5):1124–35.
40. MacDonald TM, Tran CH, Kaitu'u-Lino TJ, et al. Assessing the sensitivity of placental growth factor and soluble fms-like tyrosine kinase 1 at 36 weeks' gestation to predict small-for-gestational-age infants or late-onset preeclampsia: a prospective nested case-control study. BMC Pregnancy Childbirth. 2018;18:354.
41. Levine RJ, Lam C, Qian C, et al. Soluble endoglin and other circulating antiangiogenic factors in preeclampsia. N Engl J Med. 2006;355(10):992–1005.
42. Zhang J, Klebanoff MA, Roberts JMD. Prediction of adverse outcomes by common definitions of hypertension in pregnancy. Obstet Gynecol. 2001;97:261–7.
43. Story L, Nelson-Piercy C. Aspirin versus placebo in pregnancies at high risk for preterm pre-eclampsia. Obstetric Medicine. 2018;11(2):90–1.
44. Wright A, von Dadelszen P, Magee LA, Syngelaki A, Akolekar R, Wright D, Nicolaides KH. Effect of race on the measurement of angiogenic factors for prediction and diagnosis of pre-eclampsia. BJOG. 2023;130:78–87.
45. Roberts JM. Preeclampsia: new approaches but the same old problems. Am J Obstet Gynecol. 2008;199:443–4.
46. Serra B, Mendoza M, Scazzocchio E, et al. A new model for screening for early-onset preeclampsia. Am J Obstet Gynecol. 2020;222:608.e1-608.e18.
47. Rana S, Lemoine E, Granger JP, Ananth KS. Preeclampsia. Circ Res. 2019;124(7):1094–112.
48. Clark-Sevilla AO, Lin YC, Saxena A, Yan Q, Wapner R, Raja A, Pe'er, I., & Salleb-Aouissi, A. Diving into CDC pregnancy data in the United States: Longitudinal study and interactive application. JAMIA Open. 2024;7(1):ooae024. https://doi.org/10.1093/jamiaopen/ooae024.

Publisher's Note