PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (http://bmjopen.bmj.com/site/about/resources/checklist.pdf) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

| TITLE (PROVISIONAL) | Cohort profile Colombian Cohort for the Early Prediction of Preterm Birth (COLPRET). Early prediction of preterm birth based on personal medical history, clinical characteristics, vaginal microbiome, biophysical characteristics of the cervix, and maternal serum biochemical markers. |
|---------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| AUTHORS             | Becerra-Mojica, Carlos; Parra-Saavedra, Miguel; Díaz-Martinez, Luis; Martinez-Portilla, Raigam; Rincón Orozco, Bladimiro                                                                                                                                         |

VERSION 1 – REVIEW

| REVIEWER          | Tarca, Adi L Wayne State University                                                                                       |
|-------------------|-----------------------------------------------------------------------------------------------------------------------------|
| REVIEW RETURNED   | 07-Feb-2022                                                                                                                 |
| GENERAL COMMENTS  | The authors propose a cohort study to develop predictive strategies for preterm birth by combining maternal characteristics and obstetrical history with cervical length, vaginal microbiome and maternal serum biochemical markers. The authors are to be commended for putting forward a protocol and an analysis plan before gathering the data.                              |

Major:
1) The main limitation of the study is that very small sample size. Of the 244 recruited patients, even if the preterm birth is say 10%, there is very limited power to discover biomarker and evaluate multivariate models, especially using omics platforms. This is regardless what strategy is used for analysis.
2) There are no statistical power considerations to provide support for the ability to discover biomarkers for preterm birth based on the proposed sample size and the number of biomarkers that will be tested.
3) The authors state: “A multiple model will be created using a nested logistic approach in which the base model will comprise all possible maternal characteristics found statistically significant at the univariate analysis.” Such an approach will lead to overfitting because if one starts with randomly generated data for covariates there will always be a few that will seem significant (false positives), and if combined, they will lead to a model with significant prediction but no meaningful insight. The authors should use a cross-validation strategy, in which variable selection and model fitting are performed, by rotation, on a training set, and then the model applied to a test set.

Minor:
4) Line 12: “Cervical length is the most accepted biomarker, but its
performance is low in pregnant women without a history of PTB.”

I am not sure what is the basis for this statement. While I do agree that cervical length is not accurate enough, i.e. has low sensitivity at meaningful false positive rates (e.g. 10%), cervical length is a predictor of preterm birth irrespective of the history of PTB. See for instance this study in nulliparous women: https://pubmed.ncbi.nlm.nih.gov/33940643/

and this study in a combined set of nulliparous and parous pregnancies: https://pubmed.ncbi.nlm.nih.gov/32918893/

5) Line 24: “Pregnant women in the first trimester of a single pregnancy” do the authors mean singleton pregnancies?

REVIEWER
Lee, Joo
Gachon University

REVIEW RETURNED
08-Feb-2022

GENERAL COMMENTS
Overall, there are no problems with the research process, but there are some part that need correction.
1. In the abstract, it is mentioned that “Eleven participants (11%) had a spontaneous preterm birth or premature rupture of membranes”, but there is no mention of PRM in the main text. Please define exactly what the outcome is.
2. Please check the last part of Figure3. Four out of 100 cases, their contacts were lost, but it is not presented on the figure. And the total number does not match.
3. In research ethics, it is necessary to explain that there is no penalty for not participating in the study among invited pregnant women.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Dr. Adi L Tarca, Wayne State University
Comments to the Author:

The authors propose a cohort study to develop predictive strategies for preterm birth by combining maternal characteristics and obstetrical history with cervical length, vaginal microbiome and maternal serum biochemical markers. The authors are to be commended for putting forward a protocol and an analysis plan before gathering the data.

Major:

1) The main limitation of the study is that very small sample size. Of the 244 recruited patients, even if the preterm birth is say 10%, there is very limited power to discover biomarker and evaluate multivariate models, especially using omics platforms. This is regardless of what strategy is used for analysis.

Answer: We agree with the reviewer regarding the importance of the minimum sample size for the purpose of the study, we clarify that the number we present in the findings to date, represents only the participants recruited at the time of writing this manuscript; and the complete analysis is presented in the next answer paragraph.

2) There are no statistical power considerations to provide support for the ability to discover biomarkers for preterm birth based on the proposed sample size and the number of biomarkers that
will be tested.

**Answer:** Sample size calculation: calculations will be based on the main outcome, which is preterm birth before 37 weeks. However, this is a diagnostic test analysis on which we will calculate the needed number of patients according to the expected increase in sensitivity (Se) for a fixed 15% false positive rate (FPR) using the current best predictive model for preterm birth as the baseline Se. According to Celik et al. from the Fetal Medicine Foundation Model, the Se at a 15% FPR for a comprehensive model using cervical length, obstetric history, and maternal characteristics for the prediction of preterm birth below 37 weeks is 34.7%. We expect to increase the prediction to 50% for the same 15% FPR. For a delta of 15.30% of the expected increase using an 80% probability of finding a statistically significant difference for a threshold of 0.05 and an expected 15% attrition, we would need a total of 384 patients in the cohort.

In addition, we plan to increase the cohort as required by the results of the exploratory omics studies.

Celik, E., To, M., Gajewska, K., Smith, G. C., Nicolaides, K. H., & Fetal Medicine Foundation Second Trimester Screening Group (2008). Cervical length and obstetric history predict spontaneous preterm birth: development and validation of a model to provide individualized risk assessment. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology, 31(5), 549–554. https://doi.org/10.1002/uog.5333

3) The author's state: “A multiple model will be created using a nested logistic approach in which the base model will comprise all possible maternal characteristics found statistically significant at the univariate analysis.” Such an approach will lead to overfitting because if one starts with randomly generated data for covariates there will always be a few that will seem significant (false positives), and if combined, they will lead to a model with significant prediction but no meaningful insight. The authors should use a cross-validation strategy, in which variable selection and model fitting are performed, by rotation, on a training set, and then the model applied to a test set.

**Answer:** We decided to take a classical approach for a multivariate logistic model and add a Machine learning approach with automatic variable selection and penalization of non-significant variables to avoid overfitting and lack of parsimoniousy and robustness for the model.

A classical and Machine learning approach will be used for the construction of predictive models for preterm birth. First, a training and testing dataset will be constructed in a 50%/50% ratio. The training dataset will be used for the construction of the models and the testing dataset for testing and calibration of the models.

In the training dataset, an AI elastic net model will be used as the primary approach for model creation. The elastic net uses a ridge regression that penalizes non-statistically significant variables and reduces their coefficients to zero to avoid a non-robust model. Elastic net also uses the lasso regression, which performs an automatic selection of the variables that best predicts the outcome (preterm birth). The final model will be selected based on the lowest lambda which includes a 10-fold cross-validation automatically chosen by the lowest Akaike information criterion (AIC) and Bayesian information criteria (BIC). Because of automatic selection of predictors achieved by L1-penalty, no previous subset selection, which typically has been used in previous methods, need to be performed, thereby reducing the variance and instability of the prediction model. Automatic selection of predictors performed in elastic net results in a simpler, sparse model that includes only a subset of variables, thereby allowing for better interpretation of the model.

For the classic approach, we will use the training dataset to create a multivariate nested model using forward and backwards stepwise regression to assess the association between several predictors and the main outcome. A verification inflation factor (VIF) analysis will be performed to establish possible multicollinearity among variables in the model, where a VIF greater than 10 will be considered as highly multicollinear and will be excluded from the model, and a VIF of 4 or more will be explored. A multiple model will be created using a nested logistic approach in which the base model will comprise all possible maternal characteristics found statistically significant at the univariate analysis, then a second model will be added to the first model using biomarkers such as CL, CCI, to evaluate the added value of the addition of these markers for the prediction of preterm birth. A Naeguelkerke R² by X² analysis will be used to calculate is there a statistically significant difference among models. And as a third step, a third model will be created using the first two models and adding new markers for preterm birth such as microbiome.
The testing dataset will be used for validation of both models using a ROC curve analysis (compared by DeLong method), sensitivity (Se), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV) and accuracy as discrimination methods. Calibration of the models will be performed by plotting the observed vs predicted using the testing dataset and compared using the Hosmer-Lameshow test for goodness-of-fit.

Finally, a Kaplan Meier survival analysis will be used to establish the difference in time among women selected as high and low risk using the previous model approach.

Data will be analyzed using STATA 17 for Mac and R statistics.

Minor:

4) Line 12: "Cervical length is the most accepted biomarker, but its performance is low in pregnant women without a history of PTB."

I am not sure what is the basis for this statement. While I do agree that cervical length is not accurate enough, i.e. has low sensitivity at meaningful false positive rates (e.g. 10%), cervical length is a predictor of preterm birth irrespective of the history of PTB. See for instance this study in nulliparous women:

https://nam12.safelinks.protection.outlook.com/?url=https%3A%2F%2Fpubmed.ncbi.nlm.nih.gov%2F33940643%2F&data=04%7C01%7C%7C6b171a5324d247401be408da082122ab%7C84d9e7f9640abf435a88aaaaaaaa%7C1%7C%7C637831235688409387%7CUunknown%7CTWFpbGZsb3d8eyJwIjoiMC4wLjAwMDAiLCJQIjoiV2luUGFwcHJ1ZS10YWJsZS12aWV3LWZpZi12MC5ldmVudC8yMDE2OS5hZGRtcG93Lmh0bWw%3D%26amp%3Ddata=S2FtcG93Lmh0bWw%3D&data=RegkI2PipCzjoSvvzD8tlyOHcNxt3DhuEl2XZ1fnqM%3D&amp;amp;reserved=0

and this study in a combined set of nulliparous and parous pregnancies:

https://nam12.safelinks.protection.outlook.com/?url=https%3A%2F%2Fpubmed.ncbi.nlm.nih.gov%2F32918893%2F&data=04%7C01%7C%7C6b171a5324d247401be408da082122ab%7C84d9e7f9640abf435a88aaaaaaaa%7C1%7C%7C637831235688409387%7CUunknown%7CTWFpbGZsb3d8eyJwIjoiMC4wLjAwMDAiLCJQIjoiV2luUGFwcHJ1ZS10YWJsZS12aWV3LWZpZi12MC5ldmVudC8yMDE2OS5hZGRtcG93Lmh0bWw%3D%26amp%3Ddata=S2FtcG93Lmh0bWw%3D&data=eccHwg2B5q3rlC48YzFosy%2FHpc7wQWhC%2BacgCnyLJe4%3D&amp;amp;reserved=0

Answer: We changed the statement in line 12 to include the results from the research the reviewer cited, and in line with this change, we modified the first paragraph on page 6.

5) Line 24: "Pregnant women in the first trimester of a single pregnancy" do the authors mean singleton pregnancies?

Answer: thanks; we changed the word to "singleton" pregnancies

Reviewer: 2
Dr. Joo Lee, Gachon University
Comments to the Author:
Overall, there are no problems with the research process, but there are some part that need correction.

1. In the abstract, it is mentioned that "Eleven participants (11%) had a spontaneous preterm birth or premature rupture of membranes", but there is no mention of PRM in the main text. Please define exactly what the outcome is.

Answer: The outcome is spontaneous preterm birth; we removed the term PRM from the sentence considering the PPROM is included in the outcome definition.

2. Please check the last part of Figure3.
Four out of 100 cases, their contacts were lost, but it is not presented on the figure. And the total number does not match.

**Answer:** Thanks; we arranged the numbers and modified the flow chart.

3. In research ethics, it is necessary to explain that there is no penalty for not participating in the study among invited pregnant women.

**Answer:** In the informed consent, we declare that if the patient decides not to participate in the study, such a decision does not affect the clinical management; we added this consideration in the main document.

---

**VERSION 2 – REVIEW**

| REVIEWER         | Tarca, Adi L                  |
|------------------|-------------------------------|
| Wayne State University |                     |
| REVIEW RETURNED  | 19-Apr-2022                   |

**GENERAL COMMENTS**

The authors have addressed my comments on the original submission.