**Comments from the Editor**

Thank you for submitting this manuscript to BMC pregnancy and Childbirth. It has undergone external peer review and, despite the conflicting recommendations of the referees, we anticipate being able to revise manuscript for publication if you can address satisfactorily the points raised and make the necessary modifications. In general, the manuscript is important as it can address the question of whether what models of angiogenic factors is the best for predicting preeclampsia within 7 days. Some issues about statistics section and construction of the prediction of models are not clear enough. I anticipate that you can address the points raised and make the necessary modifications.

To ensure the Editor and Reviewers will be able to recommend that your revised manuscript is accepted, please pay careful attention to each of the comments that have been pasted underneath this email. This way we can avoid future rounds of clarifications and revisions, moving swiftly to a decision.

*Authors’ response: Thank you very much for this opportunity to clarify the questions raised during the review and re-submit our manuscript. The reviewers’ comments were very helpful and we believe the manuscript is much better for the suggested changes. We have responded point-by-point to the comments raised during this round of review and all new changes and edits made in the manuscript are text highlighted in yellow.*

**Reviewer Comments**:

Reviewer 3

In your manuscript, you describe well the mechanism of pre-eclampsia itself and the two leading biomarkers used in the diagnosis of pre-eclampsia today. In addition to s-Flt-1 and PIGF, please add an explanation of endoglin and its contribution.

*Authors’ Response: Thank you for this comment. As you clearly pointed out, the manuscript has focused on the mechanism of interaction between s-Flt-1 and PIGF for the prediction of preeclampsia. The primary focus of this study is to assess the performance of different modelling approaches (continuous vs widely used cut-off based approach) using s-Flt-1 and PIGF as biomarkers for the prediction of preeclampsia. We also like to clarify that this is a secondary analysis of trial data from the Interventional Study on Prediction of Preeclampsia/Eclampsia in Women With Suspected Preeclampsia (INSPIRE) that aimed to evaluate the use of sFlt-1/PlGF ratio in women presenting with suspected preeclampsia. We therefore do not have further information on other biomarkers for preeclampsia and was not the scope of our current work.*

Since you link the occurrence of pre-eclampsia to disturbances in lipid metabolism and lactate dehydrogenase and calcium supplementation, the mechanism by which you seek to explain the reason for the different concentrations in the control group and in the milder and more severe forms of pre-eclampsia is under-represented.

*Authors’ Response: Thank you for this comment. The mechanisms we referred to are those currently used to explain the relationship of the biomarkers (sFlt-1/PlGF ratio ) with preeclampsia. The current work demonstrates the potential misclassification and implications for using sFlt-1/PlGF cut off based values instead of continuous values of sFlt-1, PlGF, or sFlt-1/PlGF ratio. We have shown that using continuous values of biomarkers provide a better prediction for preeclampsia within 7 days.*

In addition, I suggest that the text be supplemented with data on the impact of increased body mass index (BMI), especially the proportion of increase during pregnancy. While incorporating the above additions, please also explain all this in the discussion.

*Authors’ Response: Thank you for raising this question. We have provided the distribution of BMI (median and IQR) in the two groups in Table 1. Whilst we acknowledge this is an important risk factor, we had limited statistical power to formally include this in our regression model. Since the aim of the study was to explore the prognostic utility rather than identifying risk factors, using the existing sample size guidelines for prognostic model development, we could only fit a model with a maximum of four regression coefficients (intercept, drug arm, biomarkers and potential fractional polynomial transformation for biomarkers) in our model using the available data. This limitation has been explained in the methodology section under “Sample size assessment for the development of a prognostic model” (lines 118-123) as below.*

“The adequacy of sample size was assessed using the ***pmsampsize*** library in R program as recommended by Riley et al. 22. The minimum events per parameter required for reliably developing a new model that achieved the desired shrinkage factor, R2, and margin of difference was ~10 events per variable. Therefore, with 42 outcome events, ~3-4 variables could be used for model development.”

*Also to note that the data is from a randomized trial and the median (IQR) BMI in one arm was 26.7 (23.1-31.7) and 28.3(24.3-32.4) in the other arm. We have adjusted for the trial arm in all analyses to account for any differences between the two arms even though this would be of minimal impact.*

**Reviewer 4**

Dear Colleagues

Your research work is pretty interesting. Angiogenic factors have demonstrated their utility for the prediction and management of placental insufficiency, especially PE. As you mention, few studies have assessed the capacity of angiogenic factors as a continuous variable.

However, some issues should be addressed.

INTRODUCTION

-First, I think the objective of the study should be better defined " for predicting PE within 7 days from the suspicion". The same happens with the conclusion.

*Authors’ Response: Thank you for pointing this out. We have now included this detail in the Introduction section, last paragraph. The following revised statement appears on lines 81-84:*

“Specifically, we aimed to compare the prognostic utility of models using the continuous values of sFlt-1, PlGF, or sFlt-1/PlGF ratio for predicting PE within 7 days of screening among those with suspected PE compared to the recommended cut-off-based value of the ratio of sFlt-1/PlGF of 38.”

-Some more information regarding the evidence of AF for the suspicion of PE should be added

-I recommend to reformulate the last sentence of the introduction to make it clearer

*Authors’ Response: Thank you for this suggestion. We have now included this detail in the Introduction section, second paragraph. The last sentence (lines 81-84) has now been revised as stated above.*

METHODOLOGY

- Some more information regarding the multivariable logistic regression models should be described. Which variables are included? -I recognise my limitation to assess the statistics

*Authors’ Response: Thank you for raising this suggestion. The Methodology section under multivariable model building has details on variables included and variable transformations. This has further been clarified by adding a new table that clearly describes the model equations for each of the four models. Please see the revised Table 2.*

RESULTS:

I would recommend to add cut-off>85 on the table. It would be interesting to describe whether the PE were early or late, if some HELLP syndrome was diagnosed and whether they were associated with FGR.

*Authors’ Response: Thank you for this important comment. We acknowledge that cutoffs of 85 has been proposed in the literature. However, as mentioned in our discussion, these additional cut-offs of 85 and those that are in common use in the literature will suffer the same limitations of loss of power as using a cut-off of 38. Therefore, we have decided not to include them in further analysis beyond descriptive analyses. A general limitation of the cut-offs have been explained in the lines 74-75 on introduction:*

*“*Binary thresholds, whilst simpler for clinical application 19, may lead to a loss in statistical power and lead to models with poor predictive performance 20”

DISCUSSION

It's surprising you have excellent results for the prediction of PE at one week with a cut-off>38. These results should be expected for >85. You could make some explanation regarding this fact.

*Authors’ Response: Please see our response above on the limitations of usage of cut-offs for prediction. A general limitation of the cut-offs have been explained in the lines 74-75 on introduction:*

I would recommend to comment the multivariable logistic model and to compare with other published models before.

*Authors’ Response: We had limited statistical power to construct an elaborate multivariable model that includes other risk factors that have reported in the literature to enable direct and fair comparisons with other published models. However, in the introduction, we have tried to make reference to existing multivariable models that have explored the prognostic utility of these biomarkers. For example, we have referred to the existing logistic models as highlighted in introduction and the discussion sections (please see lines 67-72 and lines 225-227):*

Lines 67-72: Prognostic models based solely on the classical clinical risk factors (high blood pressure and proteinuria) have been reported to have poorer predictive ability compared to the models developed using sFlt-1 and PIGF biomarkers 10 11. However, most of the prognostic models built using sFlt-1, PIGF, or their ratio are based on dichotomising the continuous measurements of these markers using threshold values 4 5 12–15. Moreover, only a few studies have modelled or compared sFlt-1, PIGF, or their ratio on a continuous scale 16–18.

Lines 225-227: “This finding was in line with the fact that biomarkers perform better as continuous variables than as dichotomous cut-offs as it is more biologically plausible because the disease is a continuum 17”

Would recommend to review English vocabulary

*Authors’ Response: Thank you for this suggestion. We have carefully reviewed the MS and revised the manuscript accordingly to the best of our ability.*

**Reviewer 5**

General Comment

I did not realize that this paper was a revised manuscript as the email inviting me to review this paper did not indicate this fact before I accepted the invitation to Review.

Summary:

This is a retrospective study in which the authors have developed 4 binomial logistic regression models to predict PE within 7 days using 4 independent predictors documented in 384 women previously reported in the INSPIRE trial.

The authors developed one prognostic/predictive model per independent predictor assessed ie. 1) sFlt-1 Only; 2) PlGF Only; 3) sFlt-1/PlGF ratio only and 4) dichotomized model based on sFlt-1/PlGF ratio <38 . The authors then compared each models performance and concluded that probabilistic models derived from the sFlt-1/PlGF ratio or

The authors conclusion is that models derived from continuous measurements specifically sFlt-1 and sFlt-1/PlGF ratio have a higher predictive performance as compared to PlGF and sFlt-1/PlGF ratio <38.

Comments

1) It is well known that dichotomizing continuous variables eg sFlt-1/PlGF ratio causes several problems namely loss of information reducing association with outcomes especially if research studies are small, increase the risk of a positive result being a false positive, underestimate variance between outcome groups, arbitrary polarization about the cut-point.

2) The authors have constructed a binomial model from a single independent predictor which is already dichotomized eg. sFlt-1/PlGF ratio <38. I can understand if they merged the dichotomized ratio with other pregnancy specific related information but to just use sFlt-1/PlGF ratio <38 on its own would mean assessed pregnancies could only have 1 of 2 values as shown fig 1 P-P plot for sFlt-1/PlGF.

*Authors’ Response: Thank you for this important comment. We constructed the binomial models based on sFlt-1/PlGF ratio values without categorizing them in any cut off values but as original continuous values. The fourth model did use sFlt-1/PlGF ratio <38 or >38 but it was for the sole purpose of comparing its performance with the other three continuous predictor based models. In all the models, we have adjusted for the trial arm. However, due to a lack of statistical power, we were not able to adjust for other pregnancy-related information. This has been acknowledged as a limitation and we have addressed this in the methodology section under “Sample size assessment for the development of a prognostic model” (lines 118-123):*

“The adequacy of sample size was assessed using the ***pmsampsize*** library in R program as recommended by Riley et al. 22. The minimum events per parameter required for reliably developing a new model that achieved the desired shrinkage factor, R2, and margin of difference was ~10 events per variable. Therefore, with 42 outcome events, ~3-4 variables could be used for model development.”

3) By performing a binomial regression the authors have created a logit function (Y= Intercept + Beta coefficient \*X) for each model. This would mean that the have converted a measurement from one unit eg pg/ml to another probability. All this means is that they have simply rescaled the original measurement from one scale to another. This would explain why calibration slopes for sFlt-1, PlGF and their ratio are 1 as illustrated in Table 2 and Fig 1.

*Authors’ Response:* Calibration slope is a well-established measure for the assessment of prognostic models. As suggested by the reviewer, since the probability is a function of the beta coefficients for X, naturally, prognostic models developed on the development data (typically referred to as “training data” in statistical literature) will have a high calibration metric compared to its performance on an external data (typically referred as “validation dataset”). Therefore, we have corrected for this “over-performance” (i.e. optimistic result obtained on our developmental data) and reported “optimism-corrected” performance measures derived using bootstrap resampling. While the optimism corrected calibration slopes were close to 1 for models based on biomarkers kept as continuous (as pointed out by the reviewer), the metric for model with biomarkers as ratio value suffers from poor calibration (value of 0.84 i.e. this model leads to overfitting). Overall, these performance measures strongly suggest that models based on cut-offs are less likely to be accurate when applied on a new (unseen) data compared to the models based on continuous values of biomarkers.

4) The authors use Bayesian information criterion (BIC) and presumably have ranked their models on the basis of increasing BIC. A downside of BIC is that in small datasets using BIC can result in selection of models which are too simple. BIC penalizes complexity of the model when model building minimizing the chance that a model is selected which is over fitted when constructing hierarchical models in a stepwise manner. The authors should add a statement in methods that optimum model selection form amongst the models was that with the lowest BIC if that is what they wish to use.

*Authors’ Response: Thank you for this clarification. We have now added in Table footnote how BIC values are used in model interpretation i.e., the lowest BIC values indicate better model performance. Also of note is that various metrics for model assessment were used in addition to BIC and detailed in Table 3. For example, the c-index and* R2 *values were consistent with BIC values leading to the same conclusions regarding optimum selection of models/preference amongst the models fitted.*

5) As this paper is entitled “The prognostic utility ….” then I would expect to see data related to performance ie AUC, use of delong test to compare between AUCs etc.

*Authors’ Response: Thank you for this important comment. Whilst the AUCs aren’t directly presented in the results section, they are presented as Harrell’s C-index in Table 3. The C-index is equivalent to the area under the ROC curve when the outcome is binary.*

In reply to Reviewer’s comments they mention that their sample size is too small. If that is so then would a sample size of 384 be equally too small to construct a representative model? An assessment of prognostic/predictive performance given that the paper includes this phrase in its title would be expected.

*Authors’ Response: Thank you for this important comment. The overall aim of our work is not to develop a new prognostic algorithm to identify Pre-eclampsia (our study is under-powered for that purpose) BUT our aim is to contribute to the literature in comparing the prognostic utility of the biomarkers as categorical variable (i.e. as a ratio above or below 38) which is widely used against the use of biomarkers as continuous. Therefore, we have been very cautious and have tried to present this rationale in the methodology section where we clearly demonstrated that the maximum number of parameters that we can have in our model is four. With only four parameters that can be reliably estimated, this precludes the possibility of developing a new prognostic algorithm. However, this doesn’t affect the assessment of the prognostic utility of using the biomarkers as categorical vs continuous. Therefore, our manuscript uses the phrase “prognostic utility” rather than “prognostic modelling”.*

6) The authors in their introduction mention that dichotomizing, often for clinical ease of decision making and management, results in loss of predictive performance as per my point (1). The authors therefore need to state in their methods how they would use there constructed models. They have univariate models for a binomial outcome with each model giving a probability (‘p’ ranging from 0 to 1) of the event of interest. For clinical utility to be assessed the authors would have to dichotomize ‘p’ either to achieve a given sensitivity, specificity, FRP, NPV or PPV or a value for ‘p’ which is optimal for both sensitivity and specificity (Youdens). The authors discuss predictive performance in their discussion, yet do not give or provide the any of the generally and commonly accepted measures of predictive performance in their results. At a like for like level is there a difference?

*Authors’ Response:* Thank you for raising this important point. The performance measures that we have reported include calibration measures (Calibration slope and Calibration in the large) and the discrimination measure of C-index. The C-index is equivalent to AUC for a binary outcome. The AUC is a composite metric of Sensitivity and Specificity (or the Youden’s Index) and hence, the presentation of the C-index will convey all these underlying metrics. While sensitivity, specificity, and the PPVs and NPVs can be computed for testing the accuracy of a specific cut-off, our aim was not to identify another “cut-off”. Therefore, we have reported the commonly used calibration and discrimination measures for assessing the prognostic utility and importantly, a comparison and assessment based on a cut-offs based approach of biomarkers vs. using continuous values of the biomarkers.

g) The authors gestational age ranges from 24 to 37 weeks. Over this period sFlt-1, PlGF levels and their ratio changes in a quadratic and manner. sFlt-1 one also reduces as gestation advances after reaching it peak level at around 30-32 weeks. The authors model assumes that all three markers are independent of gestation when in fact that is not so. There are several studies which report median values for gestational range which the authors could have used to make their measurement independent of gestation. Another obvious independent factor is maternal weight which also increases with gestational age. There is thus a hemodynamic effect on sFlt-1 and PlGF. Both the effect of gestation and weight would need to be corrected for.

*Authors’ Response: This is a very good point raised by the reviewer. We agree that further work is required to evaluate the impact of GA and maternal weight on the values of the biomarkers and subsequently, PE. As we have said in response to a previous comment by the reviewer, the overall aim of our work is not to develop a new prognostic algorithm to identify Pre-eclampsia (our study is underpowered for that purpose) BUT our aim is to contribute to the literature in comparing the prognostic utility of the biomarkers as categorical variable (i.e. as a ratio above or below 38) which is widely used against the use of biomarkers as continuous.*

*To evaluate and account for the impact of GA and maternal weight, we would have to include them as covariates in our model. Unfortunately, as clearly demonstrated in the methodology section, the maximum number of parameters that we can have in our model is four. We have acknowledged this limitation in the MS.*

*The suggestion from the reviewer regarding the use of median values of the biomarkers from other studies is reasonable. However, this would only be appropriate if these studies were specifically aimed at establishing normative values for these biomarkers to be used as a standard for making direct comparisons with other studies as is commonly done in the field of child growth standards. To the best of our knowledge, such standards do not exist for sFlt-1 and PlGF biomarkers.*

Authors Response to Original Review

The authors reply does not fully answer the questions raised. Specifically those related to assessment of predictive performance, why measurements were not MoM in order to correct impact of confounders

*Authors’ response: The INSPIRE trial is a randomized study that aimed to compare the incidence of PE in two trial arms. Since this is a randomized study, there is no baseline confounding as the patient allocation was done at random and any differences in baseline distribution of variables are attributed to play of chance. However, we appreciate that there is a possibility of post-baseline confounding influencing the results; this is a topic which remains beyond the scope of this work.*