Author’s response to reviews

Title: External validation, update and development of prediction models for pre-eclampsia using an Individual Participant Data (IPD) meta-analysis: the International Prediction of Pregnancy Complication Network (IPPIC Pre-eclampsia) protocol

Authors:

John Allotey (j.allotey@qmul.ac.uk)
Kym Snell (k.snell@keele.ac.uk)
Claire Chan (c.l.chan@qmul.ac.uk)
Richard Hooper (r.l.hooper@qmul.ac.uk)
Julie Dodds (j.dodds@qmul.ac.uk)
Ewelina Rogozinska (e.a.rogozinska@qmul.ac.uk)
Khalid Khan (k.s.khan@qmul.ac.uk)
Lucilla Poston (lucilla.poston@kcl.ac.uk)
Louise Kenny (L.Kenny@ucc.ie)
Jenny Myers (jenny.myers@manchester.ac.uk)
Basky Thilaganathan (basky@pobox.com)
Lucy Chappell (lucy.chappell@kcl.ac.uk)
Ben Mol (ben.mol@adelaide.edu.au)
Peter Von Dadelszen (pvondade@sgul.ac.uk)
Asif Ahmed (asif.ahmed@aston.ac.uk)
Marcus Green (chief.executive@apec.org.uk)
Liona Poon (chiu_yee_liona.poon@kcl.ac.uk)
Asma Khalil (asmakhalil79@googlemail.com)
Karel Moons (K.G.M.Moons@umcutrecht.nl)
Reviewer #1:

1. The manuscript describes the study protocol for a project aiming to identify an optimal approach to the management of risk of eclampsia. I can't honestly find a more "scientific" description of the scope, since the project includes a SR of existing models, their validation, eventual update, and, in case none is found to be "good enough", the derivation and validation of a new prognostic algorithm. The algorithm, furthermore, is expected to predict (as a single tool or as a group of tools) both any, early or late eclampsia.

RESPONSE: The aim of our protocol is not to ‘identify an optimal approach to the management of risk of eclampsia’, but to provide individualised estimates of the risk of pre-eclampsia.

In line with current recommendations by Chalmers et al (2014), Ioannidis et al (2014) and Heneghan et al (2017), our aim is to minimise research waste in this area. Given that there are already 69 published models (majority not externally validated) to predict pre-eclampsia it will be irresponsible of us to start developing a new model, without validating existing ones. This can only be achieved by undertaking a systematic review to identify all published models, followed by their validation. Only if they do not perform well, we will attempt to develop the model. We do not consider this approach to be complex, but logical in its implementation.

We are not planning to develop a single prediction model for any, early or late eclampsia; but up to 12 separate models, according to the availability of the predictors. It will be up to the clinician to choose the optimal model.


2. As a result, the manuscript is very complex, with a lot of recursive parts (anticipation or posticipation of concepts), which prompt to non-linear and heavy reading.
Also, the manuscript, whereas clearly reflective the high methodological standard of the plan, is still very immature in the writing style, consistency, crispness. I understand this is not a final report, but a working protocol, however, I wonder whether it needs additional work to reach the level needed for publication. Many examples are provided below.

RESPONSE: We have taken these criticism on board and have revised the manuscript to ensure ease of reading.

Specific comments:

3. The abstract need to be rewritten better focusing on decision rules; also, value of the network approach and related IPD can be better explained

RESPONSE: We have modified the abstract to highlight the value of the IPD network approach in comparison to aggregate data meta-analysis. (Page 3, Lines 10-13; Page 4, Line 2-4)

Our main aim is to assess the performance of the models in predicting pre-eclampsia in terms of discrimination and calibration.

This project will investigate the performance of existing (and potentially newly developed) prediction models. We are not trying to prescribe how these prediction models are implemented in clinical practice (as for decision rules), for example, specifying a threshold above which to offer preventative treatment. Such decisions should be based on additional information, including health economic evaluation, as well as weighting the benefits and harms of treatment for patients.

4. Page 5, line 27: "and should be started on prophylactic aspirin to reduce adverse outcomes." Is this statement is based on ref 8? If so, please reference it here as well. If not provide a dedicated reference.

RESPONSE: This statement is based on ref 8, which has now been added here as well.

analyses compared with meta-analyses based on aggregate data (Review). Cochrane Database Syst. Rev. 2016;(9):MR000007), the first of which is an example on prognostics.

RESPONSE: We thank the reviewer for this suggestion and have referenced this statement with: Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 2010; 340: c221.

6. Page 7, line 41: "or the relevant preeclampsia outcomes are not studied" - please rephrase, not really clear.

RESPONSE: We have changed this sentence to read as “…or where no such models exist for the relevant pre-eclampsia outcomes.” (Page 7, Line 18-19)

7. Page 7, line 51: which models? The pre-existing, new or old?

RESPONSE: We have now made clear that we are referring to the existing models. (Page 7, Line 22)

8. Page 7, line 58: To study the added role of novel biomarkers on the accuracy of the developed models; what I think is meant is: To study the effect of adding novel biomarkers on the accuracy of the basic models newly developed on clinical predictors only.

RESPONSE: What we meant is that we will study the added value of novel biomarkers by comparing the average external validation performance, and also the heterogeneity in their performance across studies, settings and subgroups, of developed models with and without novel metabolic and micro-RNA based biomarkers.

We have now revised to clarify the above: “To study the effect on accuracy of adding novel metabolic and micro-RNA based biomarkers to the developed model based on clinical, ultrasound and biochemical markers” (Page 8, Lines 1-3)

9. Page 11, 39-44: using PROBAST is acceptable, but being it not yet published, suboptimal. Why not to use any of REMARK, TRIPOD, RIGOR (even if not formal RoB tools)

RESPONSE: REMARK, TRIPOD and RIGOR are reporting guidelines; as the reviewer states these are not risk of bias tools and thus themselves suboptimal. We had also already mentioned
in the Method/Design section (Page 6, Lines 16-18), that we will adhere to TRIPOD reporting guidelines.

We decided to use PROBAST as it is a tool specifically designed to assess both the risk of bias and concerns regarding the applicability of a study that evaluates (either develops and/or validates) a multivariable diagnostic or prognostic prediction model. It is designed to assess primary studies included in a systematic review. It has been through a rigorous development process, and involved experts in the field and has been tested on numerous reviews.

Furthermore, the PROBAST tool has already been used in published research (Ensor J, Riley RD, Moore D, Snell KI, Bayliss S, Fitzmaurice D. Systematic review of prognostic models for recurrent venous thromboembolism (VTE) post-treatment of first unprovoked VTE. BMJ Open 2016; 6(5): e011190.) and has been cited in a recent BMJ publication (Debray TP, Damen JA, Snell KI, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ 2017; 356: i6460.) and presented at a Cochrane Colloquium. Although not yet published in time for this protocol, it is likely that PROBAST will be published by the time we submit the HTA report for the project.

10. Page 18, line 5: "It is likely that the distribution of the C-statistic is not normally distributed as it is a proportion and therefore bounded by the value 1". Recommend rewording as: "It is likely that the distribution of the C-statistic is not normal as it is a proportion and therefore bounded by the value 1" or similar.

RESPONSE: We thank the reviewer for their suggestion and have changed this sentence to read as “It is likely that the distribution of the C-statistic is not normal since it is a proportion and therefore bounded by the value 1” (Page 19, Lines 11-12)

11. Page 18, line 55: A graph showing the observed (O) and expected E probabilities for groups of patient. Please add brackets around E.

RESPONSE: We have added the brackets as suggested, thank you.

12. Page 19, line 44: the C-statistic will be pooled on the logit scale, as the simulation study suggested these to be more appropriate scales for pooling in a meta-analysis. Why "these" plural? Is not the logit scale singular?

RESPONSE: Yes, we thank the reviewer for spotting this. We have updated this sentence to read as “…the C-statistic will be pooled on the logit scale, as the simulation study suggested this to be a more appropriate scale for pooling C-statistics in a meta-analysis” (Page 21, Lines 6-8)
13. Page 20, line 59: measurement of predictor measurement - I understand what is meant, but it is really cacophonic. Can't you tell: characteristics (or modality and precision) of predictor measurements

RESPONSE: We have updated this to read as “…method of measurement of predictor values” (Page 22, Line 16)

14. Page 21, line 10: please define "reasonably well" - it is daunting that in paper with such high level of methodology, a quantity so relevant to the entire process is defines so lously. Maybe you assume you use the same criteria as described below, but then this needs to be specified.

RESPONSE: A ‘good’ model is very difficult to define and depends on how consistently it performs rather than simply the average performance across studies. However, we have now used the same criteria as mentioned later. The sentence now reads, “If a specific model performs reasonably well, say with a C-statistic comparable or greater than that of the other prediction models, and a calibration slope between 0.9 and 1.1 on average across studies, we may interrogate the model performance further within specific subgroups”. (Page 22, Lines 20-23)

15. Page 21, line 41; smaller studies performed in studies with different; maybe: smaller studies performed in populations/cohorts with different

RESPONSE: We thank the reviewer for their comment and have updated the sentence to read as “…smaller studies coming from populations or cohorts with different case-mix variation.” (Page 23, Lines 10-11)

16. Page 21, line 54: have good validation performance; is this "validation results"?

RESPONSE: We mean good model performance when validated in multiple studies. We have amended this sentence for clarity, which now reads, “…that appear to have good predictive performance upon validation (based on the meta-analysis), then we will…” (Page 23, Lines 15-16).

17. Page 22, line 1: If there are an adequate number of studies available; if there is?

RESPONSE: We have amended this to say, “If there are enough studies available”. (Page 23, Lines 18-19)
18. Page 23, line 1: Updating (re-calibrating) existing prediction models; again, not quantitative set of decisions; please provide some stricter or better defined guidance.

RESPONSE: As mentioned previously, it is difficult to say what a ‘good’ prediction model is. For the C-statistic, what is considered ‘good’ can be very different across clinical areas. We have therefore specified that we will look at the discriminative performance compared to the other prediction model, saying “… if a prediction model can be identified which has good average discriminative performance (C-statistic is comparable to, or greater than that of other models), but is miscalibrated (calibration slope not between 0.9 and 1.1) or has large heterogeneity in calibration performance across different validation studies, we will consider recalibration techniques…”. (Page 24, Lines 20-24)

19. Page 23, line 24: idem

RESPONSE: We have addressed this in the response to comment 18 above.

Reviewer #2:

This is a nice and elaborate protocol, that addresses and describes the study plans in appropriate fashion most of the times. I have the following comments, many only ask for clarification but some go a bit further.

20. P6, I do not understand ‘predefine cut-off values for clinical parameters’.

RESPONSE: We have deleted this sentence to avoid misinterpretation.

21. Same page, reference 16 (TRIPOD) is not on IPD MA?

RESPONSE: We thank the reviewer for their comment. We have amended the text to say, “…and by adhering to recent reporting guidelines for prediction models and IPD meta-analysis.” We now include a reference for TRIPOD as well as for PRISMA-IPD. (Page 6, Lines 17-18)

22. P7, first primary objective: is this meant to involve univariable analysis of each marker?

RESPONSE: Primary objective one involves both univariable and multivariable analysis of each marker. As explained further on page 13, "for each of the outcomes and markers of interest, we will perform a two-stage IPD meta-analysis of the prognostic effect, unadjusted and adjusted for particular variables available across studies." Note that the univariable analysis will NOT be used as a screening tool for proceeding to multivariable analysis.

23. P10: I assume that the collaborative network already has some IPD datasets through its members, perhaps this can be described?

RESPONSE: We have added the sentence below for clarification:

“At the time of submission of the protocol, we have access to 74 IPD from 72 researchers. These need further cleaning, quality assessment of the study, data quality checks, and assessment of availability of relevant data to evaluate their inclusion in the analysis.” (Page 11, Line 19-22)

24. Same page: what do you mean with secure cyberspace?

RESPONSE: We have updated the sentence to read as “The data will be obtained in an anonymised format and stored in a secure data repository.” (Page 11, Lines 5-6)

25. Same page, 'the original variable will be replaced with a proxy' sound weird. The advantage of IPD was stated as being able to standardize the definitions of predictors, but this appears to be in contradiction with this sentence? Can you clarify?

RESPONSE: We have updated the sentence to read as “…and where a direct match is not available in the data, a new variable will be created from other information contained within the original dataset if possible, such as calculating BMI from weight and height, or deriving mean pulsatility index by averaging the left and right pulsatility index measurements.” (Page 11, Line 23-25: Page 12, Line 1-2)

26. P13, improving normality is not necessary by default?

RESPONSE: Thank you, yes normality is not necessary by default and a transformation should be applied if the relationship between the variable and outcome (logit(probability of PE) is linear on a different scale. We have amended this sentence to, “…, although suitable transformations (e.g. natural log) will be considered if it improves model fit.” (Page 14, Line 4-5)
27. Same page, why will complete case analysis be employed here?

RESPONSE: We decided to employ complete case analysis for primary objective one because the purpose of this objective is to provide relatively simple descriptives of the prognostic effect of each marker, so multiple imputation was not deemed necessary.

28. P15, you will impute missing outcome values, but will you use them in the analysis (MI vs MID)?

RESPONSE: We will use the imputed outcomes in the analyses, as recommended by Sullivan et al. (2015) when auxiliary variables that may be associated with the outcome are included. We have added a sentence to clarify this and included the reference given below: “Imputed outcomes will be used in the analyses, rather than deleting observations with missing outcomes.” (Page 16, Lines 13-14)


29. MoMs: This is an unfortunate issue. I never really understood the relevance of MoMs. Anyway, you seem to suggest that this may render some predictors useless due to different adjustment/scales. Do you expect this to be a major problem affecting many variables?

RESPONSE: The MoM problem seem to mainly affect biomarkers that are routinely collected in pregnancy such as hCG, AFP and PAPP-A.

Some studies record the raw values for these biomarkers, other studies record them as MoMs, and some record both. Due to the different adjustment variables used in calculating MoMs, we plan to use only the MoM values that are available in the IPD to validate models that include biomarkers using MoMs in the model.

We decided not to calculate MoMs for studies that only recorded the raw values, as it would be unclear exactly what to adjust for (as different laboratories adjust for different variables) and would require us to use median’s calculated within the dataset (sample) rather than the laboratory (population) median values which should be used.

For model development, we will consider the raw values for biomarkers and adjust for other variables within the model if possible (rather than using MoMs, which we agree are a problematic approach engrained in the pregnancy field).
30. P16, how will you adjust for the biomarker assay and platform?

RESPONSE: This is again a big methodology issue, and the scope of the problem is unknown until we obtain the IPD. If possible, we will try to convert from one measurement scale to another, if such an equation exists in the literature. If not, then we will consider the type of assay and the platform as separate variables in the model, and have updated the sentence to clarify “…and will consider these as separate variables in our models.” (Page 17, Line 24-25) However, this will also depend on the number of events per variable that are encountered by taking this approach. It also makes strong assumptions that the prognostic effect of the marker is consistent after we adjust for the assay and platform. The impact of the problem may only be resolved through the IECV approach, where we can see if calibration performance is associated with the assay/platform, and if recalibration is needed upon external validation to address this. Therefore, we anticipate this will be an interactive process, and indeed may lead to some methodology work.

31. P18, the meaning of calibration-in-the-large and calibration slope is incorrectly described. E.g. the calibration slope indicates 'under- or over-prediction' is more a description of calibration in the large. Perhaps the wording was unfortunate?

RESPONSE: We thank the reviewer for their comment and have updated the definitions of both:

“Calibration-in-the-large: This measure indicates the extent that model predictions are systematically too low or too high across the dataset.” (Page 19, Lines 18 – 19)

“Calibration slope: The calibration slope indicates whether there is agreement between observed outcomes and predictions across the range of predicted risks.” (Page 20, Lines 2-4)

32. P20, with which instrument will risk of bias be assessed?

We have previously detailed on page 12, Lines 22-25, that we will use the PROBAST risk of bias assessment tool. We have also added this in brackets here for clarity (Page 21, Line 23).

33. Meta-regression: I think this can only be planned as a descriptive analysis.

RESPONSE: We agree that meta-regression is problematic, due to examining covariates across studies (thus potential study-level confounding) and low power. We therefore now label this as an exploratory analysis. This section now begins by saying, “If there are enough studies in the
analysis (10 or more studies), we will consider meta-regression models as an exploratory analysis to investigate if there are any differences in the performance measures due to the following pre-defined study-level factors: …” (Page 22, Lines 12-14)

34. Why wasn't DCA also mentioned in the section on external validation?
RESPONSE: Decision curve analysis is already mentioned as part of the external validation of existing models under the heading ‘Comparison of the performance of different models’ (Page 23, line 13). Here we say, “Decision curve analysis will be used to show the net benefit of the pre-eclampsia prediction models being externally validated, again using the subset of studies for which a direct comparison of the most promising models is possible.” (Page 24, Lines 11-13)

35. I do not understand how using study-specific intercepts will help to update a model with good overall discrimination and calibration performance but with heterogeneity in calibration. This will mainly be helpful for included studies/centres, but not for other centres?
RESPONSE: For the centres included in the analysis, study-specific intercepts could improve overall performance of the model for each centre. That is, we could reduce between-study heterogeneity in calibration performance. But we agree with the reviewer that for new centres, this is problematic. However, if new centres could estimate their own intercept, for example based on outcome prevalence or some other characteristics, they might be able to select an appropriate intercept from those that have been estimated, and thus improve model performance in new centres as shown elsewhere (Debray et al., 2015). Of course, if an intercept cannot be estimated for the new centre, using the average intercept is an option but may result in poor calibration externally. Therefore, we still maintain that providing centre-specific intercepts is helpful, but that also providing the average centre intercept will be needed for new centres that cannot estimate their own intercept.


36. Why will you not consider recalibration of models with good discrimination but poor overall calibration?
RESPONSE: This is a good point and one we agree with. We have amended this based on both reviewers’ comments now to say, “…if a prediction model can be identified which has good average discriminative performance (C-statistic is comparable to, or greater than that of other
models), but is mis-calibrated (calibration slope not between 0.9 and 1.1) or has large heterogeneity in calibration performance across different validation studies, we will consider recalibration techniques such as using study-specific intercepts, in an attempt to improve model performance”. (Page 24, Lines 20-25)

37. You should make a statement about events per variable (EPV) in the section on the development of new models.

RESPONSE: We thank the reviewer for their comment and have added the sentence below to the manuscript:

“Early onset pre-eclampsia is the rarest of the three outcomes (0.5% of all pregnancies). As a rule of thumb when developing a prediction model, we need at least 10 events for each candidate predictor variable to reduce the potential for large overfitting. If necessary, we will limit the number of candidate predictors considered to achieve this. However, we are likely to have an adequate number of events per variable using IPD from multiple studies.” (Page 25, Line 8–14)

38. New models will be developed using a separate intercept for each study. Do you mean that a random intercept model will be used?

RESPONSE: We will consider both options, a stratified intercept and a random intercept. A stratified intercept may improve performance within particular studies if a suitable intercept can be selected (as mentioned for a previous comment), but an average intercept may be necessary if a suitable intercept is not known (i.e. if a new study/center does not have data in which to estimate an appropriate intercept or if different to the studies included). We will look at the model performance given both approaches.

39. New models will also check for heterogeneity in predictor effects: Will you check this using random slopes? This would impact on EPV because you need to estimate another parameter. Further, how will this be reconciled with MFP? Will you first drop predictors with too heterogeneous effects, and then perform MFP?

RESPONSE: This is certainly an interesting and challenging methodology issue. We will examine heterogeneity in a two-stage approach and visually using forest plots, and especially look for situations where the effect appears null (or opposite direction) in some studies but strong in others. This will be done before consideration of MFP (i.e. assuming a linear relationship), though further work of this methodology issue is warranted.
40. You frequently discuss study-specific intercepts, but what if one study involves data on multiple centres? Isn't it better to use centre-specific intercepts?

RESPONSE: Possibly, but we are unlikely to have that information. Also, pre-eclampsia is quite a rare event, so there may not be events in some centres, especially if some centres are small. For these reasons, we do not plan to evaluate multiple centres within studies.

41. IECV: if you use a random intercept model, and then apply that model to one study that was left out, you probably have to use the average (fixed) intercept. This is likely to have an impact on calibration, so is this method fully valid?

RESPONSE: The IECV approach is used to mimic the concept of applying a developed model to a new study for external validation. In that sense, in each cycle, we would use the average intercept from the developed model (strategy 1) or choose one of the intercepts from the developed model from a study that is ‘similar’ (e.g. in prevalence) to an omitted study (strategy 2). As discussed above, strategy 1 and 2 can both be evaluated to examine their performance, including calibration.