PEER REVIEW HISTORY

BMJ Medicine publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

This paper was submitted to a another journal from BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Medicine. The paper was subsequently accepted for publication at BMJ Medicine.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Predicting birthweight: development and validation of prognostic model using individual participant data meta-analysis
AUTHORS	Allotey, John; Archer, Lucinda; Snell, Kym; Coomar, Dyuti; Massé, Jacques; Sletner, Line; Wolf, Hans; Daskalakis, George; Saito, Shigeru; Ganzevoort, Wessel; Ohkuchi, Akihide; Mistry, Hema; Farrar, Diane; Mone, Fionnuala; Zhang, Jun; Seed, Paul; Teede, Helena; Da Silva Costa, Fabricio; Souka, Athena P.; Smuk, Melanie; Ferrazzani, Sergio; Salvi, Silvia; Prefumo, Federico; Gabbay-Benziv, Rinat; Nagata, Chie; Takeda, Satoru; Sequeira, Evan; Lapaire, Olav; Cecatti, Jose; Morris, Rachel; Baschat, Ahmet; Salvesen, Kjell; Smits, Luc; Anggraini, Dewi; Rumbold, Alice; van Gelder, Marleen; Coomarasamy, Arri; Kingdom, John; Heinonen, Seppo; Khalil, Asma; Goffinet, François; Haqnawaz, Sadia; Zamora, Javier; Riley, Richard; Thangaratinam, Shakila; Collaborative Network, IPPIC

VERSION 1 - REVIEW

REVIEWER 1	Schlembach, Dietmar; Vivantes Clinicum Berlin-Neukoelln, Clinic for Obstetric Medicine. Competing Interest: Study site for the German multicenter RCT on PETN (DFG funded [https://gepris.dfg.de/gepris/projekt/286545667], Trial registration: DRKS00011374 registered at September 29th, 2017 and NCT03669185 , registered September 13th, 2018)
REVIEW RETURNED	29-Aug-2023

GENERAL COMMENTS	I read with great interest the manuscript on prediction of birthweight by "simpel" markers in early pregnancy. estimates were 1.00 (95% CI 0.78 to 1.23; calibration slope), 9.7g (95% CI -154.3g to 173.8g; calibration-in-the-large) and 1.00 (95% CI 0.94 to 1.07; observed vs expected ratio). By inclusion of assumed gestational age at delivery, maternal weight at first antenatal visit, height, age, parity, smoking status, ethnicity, medical history including chronic hypertension, diabetes, assisted conception, and any previous history of pre-eclampsia, stillbirth, and small-for-gestational-age baby. - the quite extensive statistic part (although necessary) is for a clinician not easy to understand. Maybe some of the statistics could be supplementary material?

REVIEWER 2	Reviewer 2. Competing Interest: None
REVIEW RETURNED	29-Aug-2023

GENERAL COMMENTS

RE: Predicting birthweight: development and validation of prognostic model using individual participant data meta-analysis

The authors seek to develop and validate a model for the prediction of birthweight at various potential gestational ages of delivery using information collected at the time of first antenatal visit. This is a very important and interesting study, and the results will be beneficial to clinicians and researchers working within this field. The manuscript is overall well written and clear. I have some comments and feedback which I have listed below.

Major Comments:
1) One main concern I have is surrounding the timing of the collection of the predictor information. The paper refers to information 'readily available at the first antenatal visit' but the exact timing of this visit is unclear and inevitably different across the studies, thus likely to impact the performance of the model and its generalisability to new settings where the timing of the visit may differ. This may not be visible in the results presented as the largest study is included in every round of internal-external cross validation and thus we never see the performance of the model when this study is omitted.
2) While I understand why the authors have not completed the last round of internal-external cross validation where the NICHD 2018 study is omitted as it is the largest study, I would like to see these results where it is omitted, even if there is an argument to not include the statistics from this round in the calculation of the pooled statistics.
3) Page 6 Line 54-59, the authors state, "Furthermore, the predictive performance of previously published models have not been validated externally, and so none can currently be recommended for use in routine clinical practice." Given the flooding of the literature with new models without first seeking to validate existing models, I urge the authors to include a short validation of the previous available models in this paper to demonstrate its performance. That is, of course, unless the previously developed model has methodological flaws or requires predictors which are not readily available at the time of first antenatal visit or in this data resource.

Minor Comments (please note the page numbers referred to below correspond to those on the bottom of the pages - these don't appear to be consistent with the page numbers on the top of the pages):
1) Abstract (Methods) Page 3 Line 43 – Should this be 95% prediction intervals or confidence intervals?
2) Abstract (Results) Page 3 Line 46 – Some description of the analysis sample would be good here to allow the reader to understand the population the model could potentially be applied in. For example, geographical location of the 4 included cohorts, average age of mothers, range of gestational ages at delivery in analysis sample, average birthweight of babies in the sample…etc
3) Abstract (Results) Page 3 Line 57 - specify the (0.90 to 1.07) are ranges across the three cohorts or instead just list the three

	slopes as there are only three and two are already given. Same applies for the CITL and O:E ratio. Alternatively, just present the pooled estimates and mention the level of heterogeneity across studies. 4) Methods Page 8 Line 39 – Mention the geographical location of the four studies here in the methods section 5) Methods Page 9 Line 6 – “To calculate the minimum sample size required, we assumed a lower bound of 0.5 for the anticipated adjusted R2”. Why was such a low R2 value used to calculate the minimum sample size? Was this based on performance of previous models from the literature? If not, consider using the R2 of the previous models to predict birthweight. 6) Methods Page 10 - While the R2 and calibration statistics are crucial, I would also like to see some measure of the individual level error in the prediction such as the RMSE used for linear prediction models. This will allow you to really ascertain the individual level accuracy of the predictions from the model. 7) Results Page 13 Line 5-12 – “The ratio of mean observed to mean predicted birthweight in each cohort was near perfect (range from 1.01 to 1.04), with confidence intervals showing that the model does not statistically significantly predict birthweight more or less than the birthweight observed (Table 3).” Consider rephrasing this sentence. 8) Results Page 14 Line 27-33 – “Visual inspection of the calibration plots also showed near perfect calibration on average for all cycles of internal-external cross validation, although some miscalibration can be seen for individual observations, especially at higher predicted birthweights (Figure 1).” There appears to be quite significant levels of miscalibration from the plots. 9) Discussion Page 18 Line 33-38 – “Our IPD meta-analysis combines data from multiple studies to develop a mathematical model, providing a more robust estimate of the association between predictors and birthweight.” This statement needs rephrasing. Firstly, one of the studies was very large compared to the other three and this likely to have influenced the results and performance (especially as the results from the internal-external validation without this study are omitted). Secondly, the model may provide you with a more robust estimate of birthweight using the combined information from the included predictors, but it does not provide a more robust estimate of the association between predictors and the outcome. 10) Reference 53 appears to have a formatting issue 11) Box 1 – Please add a note that these have been superimposed over current growth charts

VERSION 1 – AUTHOR RESPONSE

Reviewer #1 Comments

1. I read with great interest the manuscript on prediction of birthweight by "simple" markers in early pregnancy. Estimates were 1.00 (95% CI 0.78 to 1.23; calibration slope), 9.7g (95% CI -154.3g to 173.8g; calibration-in-the-large) and 1.00 (95% CI 0.94 to 1.07; observed vs expected ratio). By inclusion of assumed gestational age at delivery, maternal weight at first antenatal visit, height, age, parity, smoking status, ethnicity, medical history including chronic hypertension, diabetes, assisted conception, and any previous history of pre-eclampsia, stillbirth, and small-for-gestational-age baby.

The quite extensive statistic part (although necessary) is for a clinician not easy to understand. Maybe some of the statistics could be supplementary material?

The TRIPOD reporting guideline emphasises the importance of clear and full reporting of statistical methods.

Reviewer #2 Comments

2. The authors seek to develop and validate a model for the prediction of birthweight at various potential gestational ages of delivery using information collected at the time of first antenatal visit. This is a very important and interesting study, and the results will be beneficial to clinicians and researchers working within this field. The manuscript is overall well written and clear. I have some comments and feedback which I have listed below.

Thank you for your positive comments.

Major Comments:

3. One main concern I have is surrounding the timing of the collection of the predictor information. The paper refers to information 'readily available at the first antenatal visit' but the exact timing of this visit is unclear and inevitably different across the studies, thus likely to impact the performance of the model and its generalisability to new settings where the timing of the visit may differ. This may not be visible in the results presented as the largest study is included in every round of internal-external cross validation and thus we never see the performance of the model when this study is omitted.

Model predictors were maternal weight at first antenatal visit, maternal height, maternal age, parity, smoking status, ethnicity (White, Black, South Asian, Hispanic, Mixed or Other), history of chronic hypertension, history of diabetes, mode of conception, and any previous history of pre-eclampsia, stillbirth, or small-for-gestational-age baby. Except for maternal weight, the information from other predictors will remain unchanged throughout pregnancy once measured. While the exact timing of the first antenatal visit and maternal weight measurement may vary between and within datasets, the IPPIC variable for maternal weight at first antenatal visit was standardised across datasets for weight measured in first trimester or pre-pregnancy weight. This dataset harmonisation approach ensures that the predictor information is as consistent as possible across datasets, enhancing the validity of our analysis. Any modest variation in the timing of measurement of the maternal weight variable will not substantially impact model performance or generalisability.

We have added the below to our methods to clarify with reference to other IPPIC publications:

"Maternal weight at first antenatal visit was standardised in the IPPIC dataset to include pre-pregnancy and first trimester weight." (Pg 8, Line 7-8)

4. While I understand why the authors have not completed the last round of internal-external cross validation where the NICHD 2018 study is omitted as it is the largest study, I would like to see these results where it is omitted, even if there is an argument to not include the statistics from this round in the calculation of the pooled statistics.

We believe keeping the NICHD 2018 study in all rounds of the internal-external cross-validation (IECV) approach is a methodological strength of our paper, otherwise we are externally validating in a large study prediction models developed on small data which will be unreliable and overfitted. Including the NICHD study in all rounds of the IECV process provides the most robust and well-powered assessment of model generalisability.

5. Page 6 Line 54-59, the authors state, “Furthermore, the predictive performance of previously published models have not been validated externally, and so none can currently be recommended for use in routine clinical practice.” Given the flooding of the literature with new models without first seeking to validate existing models, I urge the authors to include a short validation of the previous available models in this paper to demonstrate its performance. That is, of course, unless the previously developed model has methodological flaws or requires predictors which are not readily available at the time of first antenatal visit or in this data resource.

External validation of existing birthweight prediction models is outside the scope of the current publication.

Minor Comments (please note the page numbers referred to below correspond to those on the bottom of the pages - these don't appear to be consistent with the page numbers on the top of the pages):

6. Abstract (Methods) Page 3 Line 43 – Should this be 95% prediction intervals or confidence intervals?

We have corrected to 95% prediction interval.

7. Abstract (Results) Page 3 Line 46 – Some description of the analysis sample would be good here to allow the reader to understand the population the model could potentially be applied in. For example, geographical location of the 4 included cohorts, average age of mothers, range of gestational ages at delivery in analysis sample, average birthweight of babies in the sample…etc

We have added details of geographical location of the 4 included cohorts. We believe the other descriptive measures are less crucial to include in the abstract results given word count limitations. The key details we want to convey are the populations represented through the geographical locations, along with the core model performance metrics. Providing too many additional descriptive statistics could distract from the main results in the constrained abstract space.

“Data of four IPPIC cohorts (237,228 pregnancies) USA (NICHD 2018), UK (Allen 2017), Norway (Stork-Groruddalen 2010), and Australia (Rumbold 2006).” (Abstract – Results, Pg 3, Line 20-22)

8. Abstract (Results) Page 3 Line 57 - specify the (0.90 to 1.07) are ranges across the three cohorts or instead just list the three slopes as there are only three and two are already given. Same applies for the CITL and O:E ratio. Alternatively, just present the pooled estimates and mention the level of heterogeneity across studies.

We have revised the abstract to now read as:

“On internal-external cross validation (3 cycles in Allen, Stork-Groruddalen and Rumbold cohorts), the model showed good calibration and predictive performance when validated in the three cohorts with calibration slope (0.90 in Allen, 1.04 in Stork- Groruddalen, and 1.07 in Rumbold), calibration-in-the-large (-22.3g in Allen, -33.42 in Rumbold, and 86.4g in Stork-Groruddalen), and observed vs expected ratio (0.99 in Rumbold, 1.00 in Allen and 1.03 in Stork-Groruddalen); the respective…” (Abstract – Results, Pg 3, Line 6-14)

9. Methods Page 8 Line 39 – Mention the geographical location of the four studies here in the methods section

We report the geographical location in the results section (Page 12, Line 5)

10. Methods Page 9 Line 6 – “To calculate the minimum sample size required, we assumed a lower bound of 0.5 for the anticipated adjusted R^2”. Why was such a low R^2 value used to calculate the

minimum sample size? Was this based on performance of previous models from the literature? If not, consider using the R^2 of the previous models to predict birthweight.

In our methods we state that we used the R^2 from previous literature to calculate the minimum sample size. We have specified that this was from previous birthweight prediction model.

“To calculate the minimum sample size required, … based on previous published birthweight prediction model.” (Page 9, Line 4-7)

11. Methods Page 10 - While the R^2 and calibration statistics are crucial, I would also like to see some measure of the individual level error in the prediction such as the RMSE used for linear prediction models. This will allow you to really ascertain the individual level accuracy of the predictions from the model.

We have added RMSE, as suggested.

“…, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and…” (Pg 11, Line 1)

12. Results Page 13 Line 5-12 – “The ratio of mean observed to mean predicted birthweight in each cohort was near perfect (range from 1.01 to 1.04), with confidence intervals showing that the model does not statistically significantly predict birthweight more or less than the birthweight observed (Table 3).” Consider rephrasing this sentence.

We have rephrased the sentence to read as:

“The ratio of mean observed to mean predicted birthweight in each cohort was near 1.00 (range from 1.01 to 1.04), with confidence intervals that overlap one. (Table 3).” (Page 13, Line 5-7)

13. Results Page 14 Line 27-33 – “Visual inspection of the calibration plots also showed near perfect calibration on average for all cycles of internal-external cross validation, although some miscalibration can be seen for individual observations, especially at higher predicted birthweights (Figure 1).” There appears to be quite significant levels of miscalibration from the plots.

We disagree with the reviewer. It is clear that there is miscalibration at the extremes of predicted birthweight, however on average there is near perfect calibration.

14. Discussion Page 18 Line 33-38 – “Our IPD meta-analysis combines data from multiple studies to develop a mathematical model, providing a more robust estimate of the association between predictors and birthweight.” This statement needs rephrasing. Firstly, one of the studies was very large compared to the other three and this likely to have influenced the results and performance (especially as the results from the internal-external validation without this study are omitted). Secondly, the model may provide you with a more robust estimate of birthweight using the combined information from the included predictors, but it does not provide a more robust estimate of the association between predictors and the outcome.

We have discussed in detail in the discussion (Page 16, Line 15-21) the limitation of including the largest study in all cycles of our IECV process. Please also see our response to comment 15 above on showing results with this study omitted. We have also rephrased the sentence to read as below considering the first point of the reviewer:

“Our IPD meta-analysis combines data from multiple studies to develop a mathematical model, providing a more robust estimate of the association between included predictors and birthweight.” (Page 19, Line 16-18)

15. Reference 53 appears to have a formatting issue

This has now been corrected.

16. Box 1 – Please add a note that these have been superimposed over current growth charts

We have updated the title to state that the predictions are superimposed over current growth charts.

References

1. Riley R D, Ensor J, Snell K I E, Harrell F E, Martin G P, Reitsma J B et al. Calculating the sample size required for developing a clinical prediction model BMJ 2020; 368 :m441 doi:10.1136/bmj.m441

VERSION 2 – REVIEW

REVIEWER 1	Cole, Tim J; UCL Great Ormond Street Institute of Child Health. Competing Interest: None
REVIEW RETURNED	06-Nov-2023

GENERAL COMMENTS	The authors, in response to the BMJ editorial and reviewer comments, have expanded on their explanations of the methods and results, but they have made no substantial changes to the paper. 1. In particular they have not addressed the fundamental flaw of the paper that it presents data from just four cohorts, one of which (the US) contains 98.4% of the data, and they use the data to construct a sophisticated calibration model which they conclude "shows promising generalisability across different populations, settings, and healthcare systems". Their basis for this claim is the comparisons between the four IPPIC cohorts (237,228 pregnancies) from USA (NICHD 2018) (n = 233483), UK (Allen 2017) (n = 1045), Norway (Stork-Groruddalen 2010) (n = 823), and Australia (Rumbold 2006) (n = 1877). Expressed as percentages the numbers are respectively 98.4%, 0.4%, 0.3% and 0.8% of the total. So nearly all the data come from the USA, with only a tiny contribution from the other 3 countries. The model is essentially a US model. It is unsurprising that when the cohorts are compared by internal-external cross validation the results for the four models are very similar (Table 2), since the US cohort is retained in all models. The authors claim this is a strength of the method, but try as I might I cannot see how having data from essentially one cohort is better than having data from multiple cohorts. The implication of the Abstract Conclusion is that the model generalises across populations, settings and healthcare systems. This is presumably based partly on the IECV, and partly on the ethnic adjustments included in the model. However it is not valid to assume that adjusting for ethnicity within the four countries is equivalent to adjusting for the same ethnicities in the countries of origin. Some evidence would need to be provided to support such

an inference.

Setting aside the ethnicity issue, let's explore the model's generalisability. First let us assume that the model is a valid representation of births in the four countries (even though it is dominated by the US). Together they have around 4.7 million births a year, 3.7 million of them from the USA. Globally there are around 140 million births a year, so that the model is based on countries providing just 3.4% of global births. In addition the four countries contain only around 5.6% of the world population. They are all categorised as High Income Countries by OECD, so they are highly selected compared to the world population.

The authors added this to the Discussion:
“... multiple external validation of the model with data specifically from low-income setting are needed to fully appreciate transportability of the model. Such external validation will confirm the model’s robustness, and suitability for use across different scenarios, strengthening its practical applicability in clinical practice.”

I am impressed by the authors' confidence that the model's robustness and transportability will be confirmed when data from low-income settings become available for cross-validation. But this cannot be known until the data are available, and it absolutely cannot be assumed. The Abstract Conclusion that the model "shows promising generalisability across different populations, settings, and healthcare systems" is poorly evidenced and unjustified.

2. As previously pointed out, the Abstract defining the setting as 94 studies in 53 countries with 4.5 million pregnancies is misleading. Any reasonable reader would expect a substantial proportion of those numbers to feature in the final model, which they do not.

Also in the Abstract, the total number of pregnancies is given as 237,228, but the numbers per country are not given. It is important for transparency that the mismatch in numbers between the four countries should be made clear upfront.

3. The editorial query about skewness was not addressed. Figure 1 shows clear evidence of left skewness at higher birthweights - failing to take it into account in the modelling is an important limitation.

Is it conventional in calibration studies to put the observed data on the y-axis and the expected on the x-axis? Normally it's the other way round.

4. Reviewer 1 remarked that as a clinician they found the extensive statistics hard to understand, and asked that some of the results be moved to the supplement. The author response is disappointingly unhelpful, that "The TRIPOD reporting guideline emphasises the importance of clear and full reporting of statistical methods." It also does not address the question.

5. In the same vein, the authors have failed to respond to Reviewer 2's request to see the IECV model omitting the USA data.

	6. The authors have added to Box 1, at reviewer 2's request, a note saying that the predictions are superimposed on current growth charts. However it does not state which growth charts they are. It also labels the examples A and B in the plot, but 1 and 2 in the text. I would be nervous about example B, which predicts a birthweight way below the 1st centile at 30 weeks yet on the 5th centile at 43 weeks. This represents a difference of about two channel widths on the chart, or 1.3 z-scores. Why should the predicted centile vary so markedly by gestation? It ought to be largely independent of gestation, since none of the predictors except maternal weight vary during the pregnancy. The modelling at early gestations looks dodgy.

VERSION 2 – AUTHOR RESPONSE

Reviewer #1 Comments

The authors, in response to the BMJ editorial and reviewer comments, have expanded on their explanations of the methods and results, but they have made no substantial changes to the paper.

We thank the reviewer for acknowledging the additions that we made to the revised version. We are sorry that we did not fully meet his expectations. In the previous version, we performed additional analysis and revised the discussion section in a substantive manner. In this version, we have further responded to his comments.

17. In particular they have not addressed the fundamental flaw of the paper that it presents data from just four cohorts, one of which (the US) contains 98.4% of the data, and they use the data to construct a sophisticated calibration model which they conclude "shows promising generalisability across different populations, settings, and healthcare systems".

We are very confused as to why it is a ‘flaw’ to utilise data from four cohorts. Pooling IPD from multiple studies increases our sample sizes for prediction model development and allows us to examine generalisability of the model. Like any IPD meta-analysis, we are synthesising all the IPD from the four cohorts that provided their IPD in representative target populations and settings of interest for the research question.

We agree that most of the data is from the US study – but this is why it is always included in the model development dataset during the cross-validation process (the ‘internal external cross validation’ process, IECV) where we check and synthesise performance in new data (from the other three cohorts).

18. Their basis for this claim is the comparisons between the four IPPIC cohorts (237,228 pregnancies) from USA (NICHD 2018) (n = 233483), UK (Allen 2017) (n = 1045), Norway (Stork-Groruddalen 2010) (n = 823), and Australia (Rumbold 2006) (n = 1877). Expressed as percentages the numbers are respectively 98.4%, 0.4%, 0.3% and 0.8% of the total. So nearly all the data come from the USA, with only a tiny contribution from the other 3 countries. The model is essentially a US model.

We do not agree that there is an issue with the model development being based mainly on US data. Any prediction model is developed in a particular dataset, usually from just one country or just one site. For example, QRISK is based only on UK data. Of course, the important question is then, does it generalise to other countries and settings? This is exactly what we are looking to assess in the IECV approach.

We develop a model each time in the US data plus two other cohorts, then test this model in the remaining cohort which would not have been involved in that round of model development. This process is repeated excluding a different non-US cohort each time. The model developed in each of these cycles approximates our final model (being US-dominated), and is externally validated in a non-US setting. In each case, the US-dominated model appears to perform consistently well in the UK, Norwegian, and Australian populations.

Our final, proposed, model is developed across all four settings to make maximum use of the available data, which is indeed dominated by the US sample. Just because the US data dominates this model development, does not mean that the model will validate well in other countries – this is exactly what we are examining with our IECV approach.

The IECV approach we have used was proposed by Royston (https://pubmed.ncbi.nlm.nih.gov/15027080/) and has been shown to be a very important method for assessing generalisability in prediction model research involving clustered datasets with IPD meta-analysis of (e.g. https://www.sciencedirect.com/science/article/pii/S0895435621001074). We are therefore strongly defending our use of this approach here.

19. It is unsurprising that when the cohorts are compared by internal-external cross validation the results for the four models are very similar (Table 2), since the US cohort is retained in all models. The authors claim this is a strength of the method, but try as I might I cannot see how having data from essentially one cohort is better than having data from multiple cohorts.

We think that there has been a misunderstanding here, which has hopefully been resolved by our answer above. – **the US data is never used for validation in the IECV approach, therefore,** the results we present are separate validation results in the UK, Norwegian, and Australian populations alone. As such, there is no reason why the results should be similar each time because different cohorts were used in each validation (in each cycle of the IECV approach).

As discussed above, the model being developed each time is very similar (due to the US-dominance in the development data), but, vitally, the validation data is independent and therefore different. That is, the US data was retained for model development always – exactly for the reasons stated by the reviewer (that it dominates the sample size). So, only the non-US cohorts were used for model validation.

Hence, for this validation using IECV, there were 3 cycles of IECV. In each cycle, we develop a model on the combined data from the US and two other cohorts, then test this model in the remaining cohort. Data for validation are independent to that used for model development each time. So, we fundamentally disagree with the reviewer comment here that our results are 'unsurprising'.

20. The implication of the Abstract Conclusion is that the model generalises across populations, settings and healthcare systems. This is presumably based partly on the IECV, and partly on the ethnic adjustments included in the model. However it is not valid to assume that adjusting for ethnicity within the four countries is equivalent to adjusting for the same ethnicities in the countries of origin. Some evidence would need to be provided to support such an inference. Setting aside the ethnicity issue, let's explore the model's generalisability. First let us assume that the model is a valid representation of births in the four countries (even though it is dominated by the US). Together they have around 4.7 million births a year, 3.7 million of them from the USA. Globally there are around 140 million births a year, so that the model is based on countries providing just 3.4% of global births. In addition the four countries contain only around 5.6% of the world population. They are all categorised as High Income Countries by OECD, so they are highly selected compared to the world population.

We appreciate the reviewers point here and agree that this needs amending. We have only examined generalisability across the four cohorts in our IPD meta-analysis, and all four cohorts came from high-income countries, which must be made clear.

It should be noted that most prediction model studies do not validate at all, let alone across multiple cohorts, thus our comments around generalisability were not without basis, however our statements around broad generalisability were too strong. Therefore, we have changed the statement, as follows:

"…shows promising generalisability across different populations, settings, and healthcare systems."

To:

"…shows promising performance across the four different populations included in this IPD meta-analysis. Further research to examine generalisability of performance in other countries, settings and subgroups is required". (Abstract Conclusion, Page 4)

21. The authors added this to the Discussion:
"... multiple external validation of the model with data specifically from low-income setting are needed to fully appreciate transportability of the model. Such external validation will confirm the model's robustness, and suitability for use across different scenarios, strengthening its practical applicability in clinical practice."

I am impressed by the authors' confidence that the model's robustness and transportability will be confirmed when data from low-income settings become available for cross-validation. But this cannot be known until the data are available, and it absolutely cannot be assumed. The Abstract Conclusion that the model "shows promising generalisability across different populations, settings, and healthcare systems" is poorly evidenced and unjustified.

As mentioned above, we agree that we need to be clearer with our claims of generalisability, to confirm that we refer only to the settings we have tested in our validation analyses. For clarity, the quoted statement was intended to encourage further external validation to confirm if the model performs well in different settings, and was not meant as a confirmation that it would perform well. Further research here is definitely needed. We have revised the abstract (please see response 12 above), and the discussion to be more appropriate in our communication of what we have already found and what further research is needed in the future.

"While our prediction model showed promising performance following 3 cycles of IECV in women from UK, Norway, and Australia, multiple external validations with data specifically from low-income setting are needed to fully evaluate transportability of the model to those settings. Such external validation will help verify the model's robustness, and suitability for use in other countries and subgroups, strengthening its practical applicability in clinical practice." (Discussion, Page 13, Para 1)

22. As previously pointed out, the Abstract defining the setting as 94 studies in 53 countries with 4.5 million pregnancies is misleading. Any reasonable reader would expect a substantial proportion of those numbers to feature in the final model, which they do not.

Also in the Abstract, the total number of pregnancies is given as 237,228, but the numbers per country are not given. It is important for transparency that the mismatch in numbers between the four countries should be made clear upfront.

Again, we agree with the reviewer that the statement regarding the wider IPPIC network is unnecessary, and we have removed this from the abstract. We now focus only on what IPD was actually used in this study.

"IPD from four cohorts from the International Prediction of Pregnancy Complications (IPPIC) Network" (Abstract Setting, Page 3)

We have also added numbers of pregnancies from each cohort to the results for full transparency.

"…from USA (NICHD 2018; 233,483 pregnancies), UK (Allen 2017; 1,045 pregnancies), Norway (Stork-Groruddalen 2010; 823 pregnancies), and Australia (Rumbold 2006; 1,877 pregnancies) were included…" (Abstract Results, Page 3)

23. The editorial query about skewness was not addressed. Figure 1 shows clear evidence of left skewness at higher birthweights - failing to take it into account in the modelling is an important limitation.

The previous query about skewness was addressed in our response - we assumed errors variances were the same, conditional on adjustment for gestational age and other predictor values. We added an appendix (Appendix 5) to the supplementary materials to show error distributions, to demonstrate the consistency of RMSE and various percentiles of Absolute Error over different observed gestational ages at delivery.

We further say in the paper:

"Sensitivity analysis of predictive performance of the IPPIC Birthweight model by gestational age at delivery did not show …. difference in individual-level error distributions across different numbers of week of gestation (Appendix 5)" (Page 15, Para 1)

24. Is it conventional in calibration studies to put the observed data on the y-axis and the expected on the x-axis? Normally it's the other way round.

Yes, this is the standard format for a calibration plot in a prediction modelling setting – see for example https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1466-7

25. Reviewer 1 remarked that as a clinician they found the extensive statistics hard to understand, and asked that some of the results be moved to the supplement. The author response is disappointingly unhelpful, that "The TRIPOD reporting guideline emphasises the importance of clear and full reporting of statistical methods." It also does not address the question.

This paper will not just be reviewed or read by clinicians. For example, there will be systematic reviewers and methodologists appraising our article in the future for prediction model reviews. These readers will need clear and transparent reporting to fully understand the methods we have used, and to interpret our complete results. Such information is vital to allow others to appraise the quality of our study and to extract results for meta-analysis. Hence, we do not want to dilute the article by either (a) not adhering to the relevant reporting TRIPOD guidelines or (b) moving key aspects to the supplementary material. We maintain that our current level of reporting is both appropriate and necessary for this type of research.

26. In the same vein, the authors have failed to respond to Reviewer 2's request to see the IECV model omitting the USA data.

We feel that we did address this in our original response, and believe that the comment stemmed from a misunderstanding of our validation approach. As the current reviewer rightly states, our model development data was largely dominated by the US cohort, thus an IECV cycle excluding this study from the model development would be uninformative and arguably irrelevant. Such a cycle would involve developing a model in only the UK, Norwegian and Australian populations: a model that would not be representative of our final proposed model, with a validation performance from this round being equally unrepresentative.

In the paper there is also an explicit section on IECV in the methods, with reference to relevant methodology papers. In the results we say

'The above analysis was done by forcing the largest of the four cohorts (NICHD 2018),[35] to remain throughout all cycles of the internal-external cross-validation, to ensure the model development sample size was always large enough to develop a reliable model and that the validation performance calculated was representative of an external validation of the final model, which was highly influenced by this cohort. We therefore developed a model in three cohorts and applied this within the fourth cohort but did not include a cycle where a model was developed without the NICHD 2018 study'.

Also, see our comments above about misunderstandings about the US study in the IECV approach in our previous responses. We are not sure what else we can do here.

27. The authors have added to Box 1, at reviewer 2's request, a note saying that the predictions are superimposed on current growth charts. However it does not state which growth charts they are. It also labels the examples A and B in the plot, but 1 and 2 in the text.

We have now corrected the error in labelling of the examples in Box 1, and added that the example chart was superimposed on GROW chart. (Box 1)

28. I would be nervous about example B, which predicts a birthweight way below the 1st centile at 30 weeks yet on the 5th centile at 43 weeks. This represents a difference of about two channel widths on the chart, or 1.3 z-scores. Why should the predicted centile vary so markedly by gestation? It ought to be largely independent of gestation, since none of the predictors except maternal weight vary during the pregnancy. The modelling at early gestations looks dodgy.

From a clinical perspective, at both these gestations, the baby will be flagged up to be a very small for gestational age baby that requires intense monitoring and early delivery as appropriate. Furthermore, 43 weeks is more theoretical than reflective of real life, since we do not go beyond 42 weeks even for a normal sized baby. We however welcome further validation at different gestational ages.

VERSION 3 – REVIEW

REVIEWER 2	Davies, Neil; University of Bristol. Comeptng Interest: None
REVIEW RETURNED	27-Feb-2024

GENERAL COMMENTS	This is a comprehensive paper that carefully develops a prediction model for birthweight using a large sample. This is an important issue that is not currently well covered by existing predictive models. The paper uses the relevant standards and checklists for developing predictive models. The paper is well conducted and I only have very minor comments and suggestions. Minor comments. 1. Could you provide the code used in your analysis in a repo e.g. GitHub? 2. Could you include details of the permissions you had to access the data and the data access committees that granted permission? 3. P9l50 – Could you include a sentence or two on how these cohorts were recruited? Are they designed or thought to be representative of the underlying population? How were they sampled? Doesn't need to be much, but just a comment on this in the methods could help reassure your readers about how these individuals were sampled. 4. P27l47 - We kept our prediction of birthweight on the continuous scale, so our model is not limited by arbitrary cut-offs used to define small or

	large for gestational age. This approach allows clinicians to calculate predicted birth centiles using any fetal growth standard of their choice, such as GROW, INTERGROWTH 21st, WHO.(28-30) Could you provide growth charts or a prediction model (e.g. R shiny app) that a clinician or parent could use? This would *massively* increase the usefulness of your paper to your readers, and shouldn't be too hard/time consuming to implement. Please feel free to say this is out of scope/there's no capacity for this. 5. P29l5 – Would it be possible to provide some statistics comparing the predictive power of your model relative to any of the others? E.g. what is the explanatory power and calibration of these models? This could help illustrate the benefits of your modelling versus what has been done before.

VERSION 3 – AUTHOR RESPONSE

Reviewer #2 Comments

This is a comprehensive paper that carefully develops a prediction model for birthweight using a large sample. This is an important issue that is not currently well covered by existing predictive models. The paper uses the relevant standards and checklists for developing predictive models. The paper is well conducted and I only have very minor comments and suggestions.

Thank you for your positive comments.

Minor comments.

29. Could you provide the code used in your analysis in a repo e.g. GitHub?

The files used for analysis are on the Keele University system where statistical analysis was originally performed, which we're currently unable to fully access as the statisticians moved to University of Birmingham. However, Prof Richard Riley and Dr Lucy Archer who performed the statistical analysis are working with Keele University to obtain these codes and will put it on GitHub in the next month.

30. Could you include details of the permissions you had to access the data and the data access committees that granted permission?

We have added the below to our methods section.

"Access to the IPPIC dataset was provided after application to the IPPIC data access committee." (Methods, Page 8, Para 4)

31. P9l50 – Could you include a sentence or two on how these cohorts were recruited? Are they designed or thought to be representative of the underlying population? How were they sampled? Doesn't need to be much, but just a comment on this in the methods could help reassure your readers about how these individuals were sampled.

We have added the below to our methods section.

“Women in the studies were recruited with sampling methods and inclusion criteria designed to capture a broad cross-section of the target population.” (Methods, Page 8, Para 4)

32. P27l47 - We kept our prediction of birthweight on the continuous scale, so our model is not limited by arbitrary cut-offs used to define small or large for gestational age. This approach allows clinicians to calculate predicted birth centiles using any fetal growth standard of their choice, such as GROW, INTERGROWTH 21st, WHO.(28-30).

    Could you provide growth charts or a prediction model (e.g. R shiny app) that a clinician or parent could use? This would *massively* increase the usefulness of your paper to your readers, and shouldn’t be too hard/time consuming to implement. Please feel free to say this is out of scope/there’s no capacity for this.

    Development of a shiny app is outside the scope of the current study; however, development of a user-friendly web-based application to facilitate use of this prediction model in clinical practice is currently underway. We are in the process of designing this tool with input from women and healthcare professionals to ensure it is acceptable, accessible and meets the needs of both women and providers."

33. P29l5 – Would it be possible to provide some statistics comparing the predictive power of your model relative to any of the others? E.g. what is the explanatory power and calibration of these models? This could help illustrate the benefits of your modelling versus what has been done before.

    The aim of our study was to develop and validate a new prediction model for birthweight, rather than to compare predictive performance of our model with existing birthweight prediction models. We appreciate that added context on this head-to-head comparison can highlight the benefits of our model over existing prediction models, but this work forms part of an existing project and is outside the scope of this research.