PEER REVIEW HISTORY

BMJ Paediatrics Open 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.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Cross-validated prediction model for severe adverse neonatal outcomes in a term, non-anomalous, singleton cohort
AUTHORS	Flatley, Christopher; Gibbons, Kristen; Hurst, Cameron; Flenady, Vicki; Kumar, Sailesh

VERSION 1 – REVIEW

REVIEWER	Reviewer name: Sicco Scherjon Institution and Country: Dept of Obsterics Groningen The Netherlands Competing interests: None
REVIEW RETURNED	06-Jan-2019

GENERAL COMMENTS	I am wondering what this model is adding for clinicians in their abilities to predict outcome after antenatally suspicion of FGR. After many models have bee published what any modeling is now needing are those variables already available at birth (or at the moment of clinical decision making) with really make a difference in the models predictive capacity and which might have an impact on clinical management. The final model has only a PPV of 27.3 and a low sensitivity (27.3%: not very much different form FGR estimation by fundal palpation. The model as suggested in this manuscript is missing this improvement. It would be of interest to have a look at a model with uses only antenatal data (e.g. EFW/PI/GA) and how much the AUC is improving after also including intrapartum (e.g. fetal distress, mode of birth (elective vs indicated), indication for birth interventions) and postpartum variables (e.g. Apgar and art pH) data. The authors more or less agree with this statement in their discussion (pg 13 line 6-13). I think clinicians are aware of this limitation. It would be interesting to get form the authors a suggestion how too go further. The suggestion of including biomarkers in these models (such as PlGF, PAPPa, HCG) has been done (in the context of obstetrical pathology), but this has nog given the hoped solution. I am wondering if the authors agree that their last sentence (pg 13 line 16-19) is not in complete contradiction to the preceding lines (pg 13 line 6-16). -what is the definition for FGR (comparing which ultrasound curve (Hadlock4?) with which fetal/neonatal growth curve). E.g in the initial model (pg 9 line 51 I would leave out MOD and IOL (as there is indication for the IOL given: lecture vs indicative). For the model selection the readers are most probably referred to the AIC, resulting in the definitive model. The model gives then aOR whereby e.g. indigenous ethnicity has an increase OR on a SANO, while EFW has a reduced OR: it is relatively unclear how these variables have their influence on the risk in the end: ethnicity is probably a 0/1 effect while the reduction effect of EFW on the SANO (OR of 0.88) cannot be estimated in grams (but 100 grams?).

bmjpo: first published as 10.1136/bmjpo-2018-000424 on 15 March 2019. Downloaded from http://bmjpaedsopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.

	-pg 7 line 50: with the Mixed effects LR model much more is doen then only looking for correlation of observation in women with moe then one child included in the study. On pg 8 the model building - probably using LR- variables are excluded (in a backward procedure; variable by variabel): how is then the AIC used: which decrease in AIC is indicative of model improvement. (What is the meaning of "parsimony") -pg 8 line 56: with a SANO prevalence of nearly 12% I would consider this study being performed in a high risk population which also means that the model would also only be applicable in a high risk setting.This also probably the explanation for the " negligible improvements" in the AUROC (pg 10; line 39) and understandable as the PPV is prevalence dependent a minor improvement in PPV was found. -pg 10: which patients data are used for the final model and for the cross validation model. I am accepting that in both models the same variables are use to predict SANO. -pg 11 line 17: on the basis of table 4 and 5 show a high calibration as the absolute numbers (table 4) and the test characteristics (table 5) are nearly identical. However the models accuracy of prediction of very early after delivery occurring SANOs is only moderate, whereby all the variables -also those becoming only available -immediate after delivery are needed to have this moderate accuracy. I am a little doubting what the clinical relevance is of a such a model. - the discussion of the threshold (pg 11 line 50) is of some interest, but then there should be dat in the result section on which threshold in the model -defining a higher risk population- would be of clinical relevance. Minor comments: -admissions to the NICU is not always a perfect outcome measurement as admittance criteria may be different beween hospitals and are sometime indicative of logistic variables. -the database consist of data collected in a period of more then 7 years. Has there been any selection of included cases (n=5.439). How many did not have a EFW estimated , a PI MCA not measured etc. -which Hadlock formula was used for EFW (Hadlock 4?) -it is important to have not only MOD as a variable but also indicated (fetal disteress e.g) vs elective. -pg 7 line 28: do the authors mean "or" and would "and/or" be better. Any of the six makes the neonate being labelled as SANO. Why then call it a "composite SANO" (pg 8 line 59). I think the authors should leave out "composite" for clarity. -pg 10 line 57: I am not familiair with "Confusion matrixes" (should be eluded to in the method/statistical section. -there is no textual capitation to the figures 3 and 4 (explaining what is demonstrated) minor textual suggestions: pg 3 line 48: severe arterial acidosis pg 9 line 54: 0.70 should be 0.69?

REVIEWER	Reviewer name: Amelia Hui Institution and Country: Hong Kong Competing interests: Nil

bmjpo: first published as 10.1136/bmjpo-2018-000424 on 15 March 2019. Downloaded from http://bmjpaedsopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.

REVIEW RETURNED	08-Jan-2019

GENERAL COMMENTS	Major strength of this paper study is that it's based on prospective 7 year study on 5439 women at one institution. As regards to outcome measures 639 women were subjected to SANO (Severe acidosis arterial, Admission to neonatal ICU, Apgar score of <3 at 5 minutes, or perinatal death). Authors have given extensive explanation on statistical analysis and results. It would be worthwhile for a statistician to look at appropriateness of these analytic measures and results. The article has provided an insight of the feasibility to predict SANO at term using maternal, intrapartum and ultrasound variables. However, the performance of the proposed model is still fair. Although the composite outcomes are associated with IOL and mode of delivery, the author should cautiously interpret whether IOL and mode of delivery should be included in the generation of prediction model as these two factors should be considered as clinical decision to be made from result of any prediction model. It would be of higher clinical implication if a model can help to select suitable candidates for IOL or predict the likelihood of operative delivery. In table 1, there is probably a typo mistake in the number of SVD (2089?) in those with SANO. Would the author include a description of the indication for ultrasound between 36 and 38 weeks? What is the total number of deliveries within the studied period? Tables and references are plentiful which are useful but could be simplified to make the article more reader-friendly.

REVIEWER	Reviewer name: Sarah J Nevitt Institution and Country: University of Liverpool. United Kingdom Competing interests: I have no competing interests
REVIEW RETURNED	14-Jan-2019

GENERAL COMMENTS	I have performed a statistical review of the manuscript "Cross-validated prediction model for severe adverse neonatal outcomes at term." The authors present a generalised linear prediction model for severe adverse neonatal outcomes (SANO). I agree with the authors that the model demonstrates moderate accuracy, appears to be highly calibrated from the cross validation analysis and that the authors have presented and interpreted their model appropriately in light of the relative strengths and limitations of the work. I have a few minor comments for the attention of the authors, mainly around wording: 1) Page 6, line 46: Minor wording point - were any particular criteria used for the definition 'non-anomalous'? From my understanding, this term would translate to 'standard' or 'not unusual' but I presume that a 'standard' birth experience may not exist! 2) Page 8, line 37: Minor wording point – I was a little confused by the sentence "The true number of outcomes (SANO) were compared to the predicted number…"

bmjpo: first published as 10.1136/bmjpo-2018-000424 on 15 March 2019. Downloaded from http://bmjpaedsopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.

	As SANO is actually a composite outcome made-up of several possible adverse 'sub'-outcomes, I interpreted the 'number of outcomes' here to mean the number of different 'sub' adverse outcomes experiences, rather than the number of people with a SANO outcome as I understand from the results. Perhaps the authors could reword slightly for clarity? 3) The p-values presented in the text on the first paragraph of page 9, do not all correspond to the p-values presented in Table 1. I assume this is due to different methodology used to examine the difference in the frequencies (chi-squared tests or similar?) and to calculate odds ratios. Although I appreciate that the different p-values lead to the same conclusions regarding significance, the numerical differences may be confusing when Table 1 is linked later in the paragraph. Please add further explanation of this or it may be easier to just present the odds ratios and associated p-values in the text and refer the reader to Table 1 for the percentage differences. 4) It may also be helpful to add a footnote of which variables are standardised to Table 1 to aid interpretation. 5) Page 10, line 26: "Using a fixed false positive cut-off of 10%..." Could the authors add further rationale for this cut-off to the methods section please?

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

Comment 1

I am wondering what this model is adding for clinicians in their abilities to predict outcome after antenatally suspicion of FGR. After many models have been published what any modelling is now needing are those variables already available at birth (or at the moment of clinical decision making) with really make a difference in the models predictive capacity and which might have an impact on clinical management. The final model has only a PPV of 27.3 and a low sensitivity (27.3%: not very much different form FGR estimation by fundal palpation. The model as suggested in this manuscript is missing this improvement. It would be of interest to have a look at a model with uses only antenatal data (e.g. EFW/PI/GA) and how much the AUC is improving after also including intrapartum (e.g. fetal distress, mode of birth (elective vs indicated), indication for birth interventions) and postpartum variables (e.g. Apgar and art pH) data. The authors more or less agree with this statement in their discussion (pg 13 line 6-13). I think clinicians are aware of this limitation. It would be interesting to get form the authors a suggestion how too go further. The suggestion of including biomarkers in these models (such as PlGF, PAPPa, HCG) has been done (in the context of obstetrical pathology), but this has nog given the hoped solution. I am wondering if the authors agree that their last sentence (pg 13 line 16-19) is not in complete contradiction to the preceding lines (pg 13 line 6-16).

Response: This study was performed on all term, non-anomalous singleton pregnancies that had an ultrasound performed between 36-38 weeks gestation, but we did not base our cohort on antenatal suspicion of FGR. We wanted to move away from developing predictive models using high-risk cohorts as this requires the dichotomisation of pregnancies using a definitive cut-off. This is statistically unsound and clinically dangerous as the probability of outcome is linear and therefore clinical decisions should be based off this probability as well as the clinician's assessment of the pregnancy throughout the gestational period.

bmjpo: first published as 10.1136/bmjpo-2018-000424 on 15 March 2019. Downloaded from http://bmjpaedsopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.

Saying this in clinical decision making it is better to have standard procedures in which case it would be preferable to create categories of risk based from probability cut-offs derived from sensitivity, specificity etc.

In regards to the inclusion of biomarkers for the prediction of adverse outcomes, our group have shown improvements in pre-labour screening for intrapartum fetal compromise in low-risk pregnancies at term using the CPR and PLGF. Gaccioli et al also found associations between placental biomarkers and adverse outcomes. I believe more research needs to be done in this area as the associations are present but what they mean when included in more complex predictive models is yet to be thoroughly explored.

Comment 2

what is the definition for FGR (comparing which ultrasound curve (Hadlock4?) with which fetal/neonatal growth curve). E.g in the initial model (pg 9 line 51 I would leave out MOD and IOL (as there is indication for the IOL given: lecture vs indicative). For the model selection the readers are most probably referred to the AIC, resulting in the definitive model. The model gives then aOR whereby e.g. indigenous ethnicity has an increase OR on a SANO, while EFW has a reduced OR: it is relatively unclear how these variables have their influence on the risk in the end: ethnicity is probably a 0/1 effect while the reduction effect of EFW on the SANO (OR of 0.88) cannot be estimated in grams (but 100 grams?).

Response: We did not use FGR in this study. MOD and IOL need to be in the model. If the model were accurate enough to be used clinically, these variables would enable the clinician to assess the probability of the SANO if they induced the pregnancy or not, or let the women attempt an SVD or go straight to caesarean if the probability of outcome was unacceptably high.

All the variables within the model contribute to the overall probability of the outcome occurring. The more information that a model has the more accurate that the probability will be. So a woman who is from indigenous ethnicity has a higher odds of the outcome occurring – increasing the probability of the outcome occurring compared to a Caucasian woman (the reference category). The EFW is recorded as standardised Z-scores, so for every increase of 1 Z-score in EFW there is a reduced odds and probability of the outcome occurring. Each pregnancy has a different combination of the predictor variables and the model combines all these factors for individual probabilities of the outcome occurring.

Comment 3:

pg 7 line 50: with the Mixed effects LR model much more is doen then only looking for correlation of observation in women with moe then one child included in the study. On pg 8 the model building - probably using LR- variables are excluded (in a backward procedure; variable by variabel): how is then the AIC used: which decrease in AIC is indicative of model improvement. (What is the meaning of "parsimony")

Response:

The correlation of the observation in women with more than one birth within the study period is the reason why I chose the mixed effect modelling approach over a straight logistic regression. As there are mothers who birthed more than once it would violate the assumption of logistic regression that the observations are independent of each other. (I am assuming the second use of LR in this comment is referring to the likelihood ratio test?) The likelihood ratio test is not appropriate for a test of model improvement for a mixed effects model. The AIC is widely used assessment of model improvement.

bmjpo: first published as 10.1136/bmjpo-2018-000424 on 15 March 2019. Downloaded from http://bmjpaedsopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.

A parsimonious model is one that best trades off explanatory power (the model that best fits the data) against using the simplest (and typically most robust) model possible (i.e avoid overfitting to the sample). AIC is a widely used measure of parsimony that penalizes for BOTH lack of fit and model complexity. At one end of the spectrum an uninformative model will have a high AIC (model with few good predictors) and at the other end of the spectrum an overfit model (one with too many predictors) will have a high AIC. A model with a good AIC (i.e. in between these two extremes) will have a relatively small number of variables, all of which 'pull their weight" in explaining the outcome (i.e. a sufficiently good fit).

Comment 4:

-pg 8 line 56: with a SANO prevalence of nearly 12% I would consider this study being performed in a high risk population which also means that the model would also only be applicable in a high risk setting.This also probably the explanation for the " negligible improvements" in the AUROC (pg 10; line 39) and understandable as the PPV is prevalence dependent a minor improvement in PPV was found.

Response:

The population includes women who had an ultrasound between 36-38 weeks gestation. While 11.7% of pregnancies had the SANO, 88.3% had an uncomplicated birth. While the prevalence is slightly higher than the Australian wide prevalence I would not class it as high risk. Furthermore, we assessed the model on cohorts that could be classified as high risk (ie EFW<10th centile, CPR<10th centile etc) and found little difference.

Comment 5:

-pg 11 line 17: on the basis of table 4 and 5 show a high calibration as the absolute numbers (table 4) and the test characteristics (table 5) are nearly identical. However the models accuracy of prediction of very early after delivery occurring SANOs is only moderate, whereby all the variables -also those becoming only available -immediate after delivery are needed to have this moderate accuracy. I am a little doubting what the clinical relevance is of a such a model.

Response:

The reviewer is making two points here.

Firstly, table 4 and 5 are showing the calibration of the cross-validation model and the original predictive model. It is checking the predictive model's appropriateness to the data by checking for such things as overfitting of the data. This is a separate concept to the model building and accuracy.

Secondly, the model adds to the current knowledge of research regarding predicting adverse perinatal outcomes using demographic and ultrasound variables. Previously models predicting adverse outcomes such as fetal distress and admission to NICU have been derived from high-risk cohorts (EFW<10th centile). This model is a more comprehensive model and shows that there is a need to incorporate all fetuses into risk stratification models. Furthermore, the use of EFW and CPR in models should be done using them as continuous variables for both accuracy of probability and to avoid loss of information that results when dichotomising variables. Accurate prediction of at-risk fetuses in the clinical setting cannot be done in consideration of only one or two assessments of variables. It needs to be done in combination of several risk factors. For example, we know that an EFW<10th centile will place a fetus at higher risk of adverse outcomes but we also know that most of the fetuses that have adverse outcomes are in the appropriate for gestational age range (EFW≥10th centile). The complexity of pregnancy and complications that arise during pregnancy mean that a comprehensive predictive model needs to assess numerous risk factors to accurately predict fetuses at high risk of adverse outcomes.

bmjpo: first published as 10.1136/bmjpo-2018-000424 on 15 March 2019. Downloaded from http://bmjpaedsopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.

While I accept that our model is only moderately predicting these outcomes it is a step forward and has the opportunity to guide further research using such things as placental biomarkers, genetic markers or epigenetics to enhance the accuracy.

Comment 6:

the discussion of the threshold (pg 11 line 50) is of some interest, but then there should be data in the result section on which threshold in the model -defining a higher risk population- would be of clinical relevance.

Response:

We advocate not using thresholds as this is not appropriate and illustrated by the linear trend in figure 3. Dichotomising continuous variables loses a lot of information and unjustly categorises those on either side of the threshold as being as different to one another as those on the lowest and highest extremes of measures. However, the reviewer does have a point. As mentioned above, in clinical decision making it is better to have standard procedures in which case it would be preferable to create categories of risk based from probability cut-offs derived from sensitivity, specificity etc.

Minor comment:

-admissions to the NICU is not always a perfect outcome measurement as admittance criteria may be different beween hospitals and are sometime indicative of logistic variables.

Response:

I agree with this and it applies to some of the self-reporting variables as well. They will always be problematic in both clinical research as well as clinical decision making. However, the imperfections are what we witness in the hospitals and between hospitals and the lack of definiteness is at random (ie if we sampled 1000 patients one year the proportion of NICU admission would be similar if we sampled 1000 patients in the next year). The way to minimise the effect you are describing is have a large sample size.

Minor comment:

-the database consist of data collected in a period of more then 7 years. Has there been any selection of included cases (n=5.439). How many did not have a EFW estimated, a PI MCA not measured etc.

Response:

At our institution routine third trimester scans are not the norm. All women that had ultrasound data would have been referred for either a clinical indication or a previous pregnancy complication. Only cases that had complete dataset of all the variables required for the model were included. Our standard protocol is for EFW and all fetal Dopplers to be routinely measured if the interval between scans was >2 weeks.

Minor comment:

-which Hadlock formula was used for EFW (Hadlock 4?)

Response:

Yes it was BPD, HC, AC, FL)Minor comment:

-it is important to have not only MOD as a variable but also indicated (fetal disteress e.g) vs elective.

bmjpo: first published as 10.1136/bmjpo-2018-000424 on 15 March 2019. Downloaded from http://bmjpaedsopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.

Response:

This is a predictive model and the indication of method of birth would not be known at time of assessment. The method of birth is more to allow the clinician to evaluate different scenarios.

Minor comment:

-pg 7 line 28: do the authors mean "or" and would "and/or" be better. Any of the six makes the neonate being labelled as SANO. Why then call it a "composite SANO" (pg 8 line 59). I think the authors should leave out "composite" for clarity.

Response:

I agree. “or” changed to “and/or”

Minor comment:

-pg 10 line 57: I am not familiair with "Confusion matrixes" (should be eluded to in the method/statistical section.)

Response:

I agree, this is confusing as it is a statistical term. A confusion matrix is just a 2X2 table of predicted verses true outcomes. In the text I have included “cross tabulation of actual and predicted outcomes (aka confusion matrix)…”

Minor comment:

-there is no textual capitation to the figures 3 and 4 (explaining what is demonstrated)

Response:

Thank you. I have added these as well as one for figure 1 in the manuscript.

Reviewer 2

Comment 1:

Major strength of this paper study is that it’s based on prospective 7 year study on 5439 women at one institution. As regards to outcome measures 639 women were subjected to SANO (Severe acidosis arterial, Admission to neonatal ICU, Apgar score of <3 at 5 minutes, or perinatal death). Authors have given extensive explanation on statistical analysis and results. It would be worthwhile for a statistician to look at appropriateness of these analytic measures and results.

Response:

Reviewer 3 is a statistician

Comment 2:

The article has provided an insight of the feasibility to predict SANO at term using maternal, intrapartum and ultrasound variables. However, the performance of the proposed model is still fair. Although the composite outcomes are associated with IOL and mode of delivery, the author should cautiously interpret whether IOL and mode of delivery should be included in the generation of prediction model as these two factors should be considered as clinical decision to be made from result of any prediction model.

bmjpo: first published as 10.1136/bmjpo-2018-000424 on 15 March 2019. Downloaded from http://bmjpaedsopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.

It would be of higher clinical implication if a model can help to select suitable candidates for IOL or predict the likelihood of operative delivery.

Response:

The decision to include IOL and Mod was based on the premise that it gives the clinician an opportunity to put different scenarios into the model and determine what the probability of the outcome occurring would be. For example the clinician could compare two scenarios - if the clinician decided not to induce but the pregnancy evolved to an emergency CS what is the probability of the outcome compared to the woman undergoing an induction and having a SVD. Using a web based or phone app this could be done in seconds and could involve all combinations of IOL and/or MOD for each pregnancy.

Comment 3:

In table 1, there is probably a typo mistake in the number of SVD (2089?) in those with SANO.

Would the author include a description of the indication for ultrasound between 36 and 38 weeks? What is the total number of deliveries within the studied period?

Response:

Yes, thank you for picking that up. It has been rectified (208/639).

Comment 4:

Tables and references are plentiful which are useful but could be simplified to make the article more reader-friendly.

Response:

Table 3 could possibly be simplified. However previous attempts as model building using ultrasound variables have used high-risk cohorts (EFW<10th centile). Part of our argument is that this is not only statistically wrong, the improvement it provides is not much which is illustrated when our model is applied to the higher risk cohorts.

Table 5 and 6 are necessary to illustrate the closeness in outcomes between the predictive model and the cross validated model. Our study is over powered for the usual statistical tests (Delong's and the Hanley and McNeil) and therefore these tables need to illustrate that there is no need for formal statistical tests.

Reviewer 3

Comment 1:

Page 6, line 46: Minor wording point - were any particular criteria used for the definition 'non-anomalous'? From my understanding, this term would translate to 'standard' or 'not unusual' but I presume that a 'standard' birth experience may not exist!

Response:

This term is commonly used in obstetric publications to refer to fetuses that are without major congenital abnormalities.

bmjpo: first published as 10.1136/bmjpo-2018-000424 on 15 March 2019. Downloaded from http://bmjpaedsopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.

Comment 2:

Page 8, line 37: Minor wording point – I was a little confused by the sentence “The true number of outcomes (SANO) were compared to the predicted number…”

As SANO is actually a composite outcome made-up of several possible adverse ‘sub’-outcomes, I interpreted the ‘number of outcomes’ here to mean the number of different ‘sub’ adverse outcomes experiences, rather than the number of people with a SANO outcome as I understand from the results. Perhaps the authors could reword slightly for clarity?

Response:

I agree. This has been changed to: “The number of SANO outcomes were compared to the number of SANO predicted by the model through the use of…”

Comment 3:

The p-values presented in the text on the first paragraph of page 9, do not all correspond to the p-values presented in Table 1. I assume this is due to different methodology used to examine the difference in the frequencies (chi-squared tests or similar?) and to calculate odds ratios.

Although I appreciate that the different p-values lead to the same conclusions regarding significance, the numerical differences may be confusing when Table 1 is linked later in the paragraph. Please add further explanation of this or it may be easier to just present the odds ratios and associated p-values in the text and refer the reader to Table 1 for the percentage differences.

Response:

I agree it is a little confusing. However, we already have 6 tables and I felt the more important analysis to present for model building is the univariable analysis. I have removed the link to table 1 at the end of that paragraph so that the descriptive analysis is only presented as text.

Comment:

It may also be helpful to add a footnote of which variables are standardised to Table 1 to aid interpretation.

Response:

The only variables that are standardised are those which are presented as z-scores.

Comment:

Page 10, line 26: “Using a fixed false positive cut-off of 10%...”

Could the authors add further rationale for this cut-off to the methods section please?

Response:

The 10% false positive rate is a commonly reported cut-off used within this field of study. It is only included for comparability to other previously published models. I believe the use of the optimal cut-off is more appropriate (as is used and reported in tables 3,4 and 5).

bmjpo: first published as 10.1136/bmjpo-2018-000424 on 15 March 2019. Downloaded from http://bmjpaedsopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.