Peer Review File

Article information: http://dx.doi.org/10.21037/jlpm-21-18

Reviewer A

Comment 1:

Study design. It would be great to have statistical power calculation with appropriate formula in the section "Methods", despite the sample size is large.

Reply 1:

We thank the reviewer for their suggestion. Our analysis used all available cases from the screening database for the 2011-2015 time period and, as such, the large cohort size ensures that our PGA model is very robust.[1] Furthermore, although the number of screen positive cases and true positive cases were relatively small, our analytical approach did not center on the significance of hypothesis tests. For these reasons, we believe that power calculations in the context of our analysis would be of little value to the reported outcomes.

1. Murphy, M. S., Hawken, S., Cheng, W., Wilson, L. A., Lamoureux, M., Henderson, M., Pervin, J., Chowdhury, A., Gravett, C., Lackritz, E., Potter, B. K., Walker, M., Little, J., Rahman, A., Chakraborty, P., & Wilson, K. (2019). External validation of postnatal gestational age estimation using newborn metabolic profiles in Matlab, Bangladesh. *eLife*, *8*, e42627. https://doi.org/10.7554/eLife.42627

Comment 2:

AUC for each model could be estimated. Yet, predictive models could be compared each other with C-statistic mode, such as INR and IDI.

Reply 2:

We appreciate the reviewer's suggestions however, as the use of PGA is a binary decision, we are unable to provide a continuum of input in order to generate a receiver operating characteristic curve.

Reviewer B

Comment 1:

Were data parametric or non-parametric and based on which statistical test?

If data were parametric – although obvious then t-test used should be stated as unpaired.
All data (if parametric) should be presented as mean (SD).

Reply 1:

We thank the reviewer for their suggestion to clarify the statistical tests used. The data were parametric based on visual inspection of the data. As such, unpaired t-tests were used to perform statistical comparisons between groups. We have amended the *Methods, Statistical Analysis* section (line 163) to reflect this:

"Statistical significance between means was determined using unpaired t-tests."

We have also updated the reported data to include the SD, where relevant. The SD for certain data can also be found in Table 1: Demographic information for all newborns included in the study population.

Comment 2:

Pg 9 lines 180-181: It is obvious and meaningless that CAH positive babies would have higher 17OHP than CAH negative babies as the screening process is dependent on 17OHP to differentiate the two groups.

Reply 2:

We agree that this provides redundant information, considering that high 17-OHP levels are required in order for newborns to receive a positive screening result. We have removed the corresponding data from Figure 1 and have changed the relevant sentence (line 182-183) as follows:

"CAH screen positive newborns had a significantly lower mean BW and mean GA ~~and higher mean 17-OHP concentration~~ than screen negative newborns (Figure 1)."

Comment 3:

Perhaps heading "Conclusions" should be replaced by "Discussion".

Reply 3:

Unfortunately, we are unable to make this change as JLPM requires that manuscripts are structured with a "Conclusions" section.

Comment 4:

Figure 1 – Has lost its legend and is difficult to visualise.

Reply 4:

We have increased the size of Figure 1 in order to improve its readability. The corresponding legend for Figure 1 has been included on page 2 of the document, as part of the formatting requirements for JLPM.

Comment 5:

Figure 2 – The legend should be just that and should not discuss results.

Reply 5:

As suggested, we have updated the legend of Figure 2 so that results are not presented.

Comment 6:

Table 1 need tidying up as columns and rows do not match

When tidied up data presentation should be appropriate if parametric or non-parametric rather than both.

Reply 6:

We have updated Table 1 in order to present parametric data; we have removed the medians and quartiles.

Comment 7:

Table 5 is difficult to understand.

Reply 7:

We thank the author for their observation that Table 5 could be made clearer. We have removed the first row of the table and incorporated the word "Algorithm" into the second and third columns to clarify that each represents a distinct screening algorithm. We have also updated the legend to reflect this change in terminology, and to direct readers to Figure 2, which has been updated with a visual representation of the two screening algorithms: GA & BW, and GA & Sequential.

Reviewer C

Comment 1:

Screening for CAH is time-critical and in the US, the recommendation is that CAH, as a time critical disorder, be reported by day 5 of life. It is therefore important to describe turn-around-time for screening. What is the turn-around time for the various stages of testing?

Timeliness of referral is essential – referrals are done at NSO the day after (for screen positives).

1) 17-OHP measurement using GSP
2) Predicting GA
3) Performing second tier test

Reply 1:

We agree that the importance of a rapid turn-around time for the availability of screening results cannot be understated. 17-OHP measurements using an immunoassay are available the morning after sample receipt, and GA predictions would be available immediately after first-tier results are ready. The results of second-tier testing are then available within 24 hours, thus guaranteeing that positive screening results can be quickly acted upon. High PPV initial results are confirmed and referred the same day. To consolidate this information for readers, we have made the following addition to lines 101-102:

"First and second tier screening results are available within 48 hours of sample receipt."

Comment 2:

The background section states that "CAH is often performed using a two-tiered approach" (lines 42-43). Although NBS labs are increasingly moving towards using a second-tier test, most NBS laboratories do not perform a second-tier CAH assay and only perform 17-hydroxyprogesterone screening.

Reply 2:

We appreciate the reviewer's consideration of variations in screening practices. We have edited lines 42-43 as follows:

"Newborn screening for congenital adrenal hyperplasia (CAH) is increasingly performed using a two-tiered approach."

Comment 3:

What percentage of NBS programs use GA-based screening thresholds (lines 44-45)? Many labs are still using BW based 17-OHP screening thresholds.

Reply 3:

Unfortunately, there is no global newborn screening organisation that collects and summarizes screening logic. As a result, we are unable to comment on the algorithms used by newborn screening laboratories.

Comment 4:

Line 89-90 – The primary target of CAH screening is to identify classic salt-wasting cases.

Simple virilizing forms are sometimes missed by NBS programs especially if cut-offs are set so as to improve the PPV of testing (Heather et al., The Journal of Clinical Endocrinology & Metabolism, Volume 100, Issue 3, 1 March 2015, Pages 1002–1008).

Reply 4:

Classic congenital adrenal hyperplasia due to 21-hydroxylase deficiency is the primary screening target in Ontario. The screening algorithm is optimized to identify both the salt-wasting form and the simple-virilizing form of classic 21-OHD CAH.

Comment 5:

Line 115 states that the model is capable of predicting GA with an accuracy of +/- 2 weeks. Supplementary table 1 gives the cut-offs based on GA. Is the accuracy of GA good enough? i.e. if the prediction is off by 2 weeks wouldn't the incorrect cut-off be used, leading to missed cases?

Reply 5:

We appreciate the reviewer's recognition of the importance of accurately predicting GA. Based on our analyses of over 700,000 newborns and 34 true cases, we have not observed a misclassification.

Comment 6:

What thresholds are used when specimens are collected at <24 hours of age? What percentage of specimens are collected at <24 hours of age. Frequently, specimens are collected at <24 hours prior to transfusion or TPN.

Reply 6:

When specimens are collected < 24 hours of age, the same thresholds are used. We do not adjust thresholds based on age of collection < 24 hours of age .

Comment 7:

Are lines 97-104 describing the NBS program in Ontario or NBS programs in general? If it is the Ontario program then t should be stated at the beginning of the paragraph.

Reply 7:

This section describes the NBS program in Ontario; we have amended line 98 to read as follows:

"At Newborn Screening Ontario, CAH screening follows a two-tier approach."

Comment 8:

Line 234 states BW-based screening was performed on 675,927 newborns with a reported GA. How many of these screened positive? How many would screen positive if PGA-based screening was performed? Data is being shown on PGA-screening minimizing BW-based screening. But, wouldn't the opposite also be true? PGA-based screening also has false positive results and wouldn't these numbers be reduced if BW-based screening was performed subsequent to PGA-based screening? As stated in line 214 false positives are introduced by PGA-based screening alone. Does this suggest that both BW-based and GA/PGA-based thresholds should be used for the entire population?

Reply 8:

We thank the reviewer for raising a valid point about applying PGA-based screening to the whole study population. Compared to BW-based screening, GA-based screening is the preferred approach because it has a higher positive predictive value and results in fewer false positive cases. For the purpose of this manuscript, we demonstrate that PGA-based screening has a higher positive predictive value than BW-based screening, but this is still lower than GA-based screening. Line 201 states: "Using PGA in place of reported GA for all newborns results in a lower PPV, thus reported GA is still the preferable parameter to use when available." Because BW-based screening introduces more false positive results than GA- and PGA-based screening, the suggested approach of performing BW-based screening after PGA-based screening is unlikely to reduce the proportion of newborns who go on to second-tier testing, and may increase it. Furthermore, we aimed to improve the positive predictive value of CAH screening for newborns without a reported GA and, at this time, a review of the overarching screening approach for all newborns screened by NSO would be beyond the manuscript's scope.

Comment 9:

In table 5, should there be 32 salt wasting cases (not 31)? In table 5, should the screen positive and screen negative in the middle column add to 702020? 2588+ 699442=702030.

Reply 9:

We are grateful to the reviewer for noticing this inconsistency in Table 5 and appreciate their attention to detail. We have reviewed the corresponding data and, unfortunately, had made a typographical error previously. We have amended the table to include the correct number of salt wasting cases and screen negative cases.