



## RESEARCH ARTICLE

**REVISED** External validation of machine learning models including newborn metabolomic markers for postnatal gestational age estimation in East and South-East Asian infants [version 2; peer review: 3 approved, 1 approved with reservations]

Steven Hawken<sup>1,2</sup>, Malia S. Q. Murphy<sup>1</sup>, Robin Ducharme<sup>1</sup>, A. Brianne Bota<sup>1</sup>, Lindsay A. Wilson<sup>1</sup>, Wei Cheng<sup>1</sup>, Ma-Am Joy Tumulak<sup>3</sup>, Maria Melanie Liberty Alcausin<sup>3</sup>, Ma Elouisa Reyes<sup>3</sup>, Wenjuan Qiu<sup>4</sup>, Beth K. Potter<sup>2</sup>, Julian Little<sup>2</sup>, Mark Walker<sup>1,5</sup>, Lin Zhang<sup>6,7</sup>, Carmencita Padilla<sup>8,9</sup>, Pranesh Chakraborty<sup>10,11</sup>, Kumanan Wilson<sup>1,12,13</sup>

<sup>1</sup>Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada

<sup>2</sup>School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada

<sup>3</sup>Newborn Screening Reference Centre, University of the Philippines Manila, Manila, Philippines

<sup>4</sup>Pediatric Endocrinology and Genetic Metabolism, XinHua Hospital, Shanghai, Shanghai, China

<sup>5</sup>Better Outcomes Registry & Network, Ottawa, Canada

<sup>6</sup>Department of Gynecology and Obstetrics, XinHua Hospital, Shanghai, Shanghai, China

<sup>7</sup>MOE-Shanghai Key Lab of Children's Environmental Health, Xinhua Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

<sup>8</sup>Department of Pediatrics, University of the Philippines Manila, Manila, Philippines

<sup>9</sup>Institute of Human Genetics, National Institutes of Health, University of Philippines Manila, Manila, Philippines

<sup>10</sup>Newborn Screening Ontario, Children's Hospital of Eastern Ontario, Ottawa, ON, Canada

<sup>11</sup>Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

<sup>12</sup>Department of Medicine, University of Ottawa, Ottawa, ON, Canada

<sup>13</sup>Bruyère Research Institute, Ottawa, ON, Canada

**v2** First published: 29 Oct 2020, 4:164  
<https://doi.org/10.12688/gatesopenres.13131.1>  
 Latest published: 21 Jun 2021, 4:164  
<https://doi.org/10.12688/gatesopenres.13131.2>

## Abstract

**Background:** Postnatal gestational age (GA) algorithms derived from newborn metabolic profiles have emerged as a novel method of acquiring population-level preterm birth estimates in low resource settings. To date, model development and validation have been carried out in North American settings. Validation outside of these settings is warranted.

**Methods:** This was a retrospective database study using data from newborn screening programs in Canada, the Philippines and

## Open Peer Review

Approval Status ? ✓ ✓ ✓

	1	2	3	4
<b>version 2</b> (revision) 21 Jun 2021		✓ view	✓ view	
<b>version 1</b> 29 Oct 2020	? view	? view	? view	✓ view


China. ELASTICNET machine learning models were developed to estimate GA in a cohort of infants from Canada using sex, birth weight and metabolomic markers from newborn heel prick blood samples. Final models were internally validated in an independent sample of Canadian infants, and externally validated in infant cohorts from the Philippines and China.

**Results:** Cohorts included 39,666 infants from Canada, 82,909 from the Philippines and 4,448 from China. For the full model including sex, birth weight and metabolomic markers, GA estimates were within  $\pm 5$  days of ultrasound values in the Canadian internal validation (mean absolute error (MAE) 0.71, 95% CI: 0.71, 0.72), and within  $\pm 6$  days of ultrasound GA in both the Filipino (0.90 (0.90, 0.91)) and Chinese cohorts (0.89 (0.86, 0.92)). Despite the decreased accuracy in external settings, our models incorporating metabolomic markers performed better than the baseline model, which relied on sex and birth weight alone. In preterm and growth-restricted infants, the accuracy of metabolomic models was markedly higher than the baseline model.


**Conclusions:** Accuracy of metabolic GA algorithms was attenuated when applied in external settings. Models including metabolomic markers demonstrated higher accuracy than models using sex and birth weight alone. As innovators look to take this work to scale, further investigation of modeling and data normalization techniques will be needed to improve robustness and generalizability of metabolomic GA estimates in low resource settings, where this could have the most clinical utility


## Keywords


biological modelling, gestational age, preterm birth, newborn screening

1. **Sunil Sazawal** , Centre for Public Health Kinetics, New Delhi, India

Johns Hopkins School of Public health, Baltimore, USA

2. **José Villar** , University of Oxford, Oxford, UK

**Eric Ohuma** , London School of Hygiene & Tropical Medicine, London, UK

3. **Julie Courraud** , Statens Serum Institut, Copenhagen, Denmark

4. **Ramesh Agarwal**, All India Institute of Medical Sciences, New Delhi, India

**Suman Chaurasia**, All India Institute of Medical Sciences (AIIMS), Rishikesh, India

Any reports and responses or comments on the article can be found at the end of the article.

**Corresponding author:** Steven Hawken ([shawken@ohri.ca](mailto:shawken@ohri.ca))

**Author roles:** **Hawken S:** Conceptualization, Formal Analysis, Funding Acquisition, Methodology, Supervision, Validation, Writing – Original Draft Preparation; **Murphy MSQ:** Data Curation, Funding Acquisition, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Ducharme R:** Formal Analysis, Project Administration, Visualization, Writing – Review & Editing; **Bota AB:** Project Administration, Writing – Review & Editing; **Wilson LA:** Project Administration, Supervision, Writing – Original Draft Preparation; **Cheng W:** Formal Analysis, Methodology, Writing – Review & Editing; **Tumulak MAJ:** Investigation, Project Administration, Writing – Review & Editing; **Alcausin MML:** Investigation, Project Administration, Writing – Review & Editing; **Reyes ME:** Investigation, Project Administration, Writing – Review & Editing; **Qiu W:** Investigation, Project Administration, Writing – Review & Editing; **Potter BK:** Methodology, Writing – Review & Editing; **Little J:** Methodology, Writing – Review & Editing; **Walker M:** Methodology, Writing – Review & Editing; **Zhang L:** Investigation, Project Administration, Supervision, Writing – Review & Editing; **Padilla C:** Investigation, Project Administration, Supervision, Writing – Review & Editing; **Chakraborty P:** Data Curation, Investigation, Resources, Supervision, Validation, Writing – Review & Editing; **Wilson K:** Conceptualization, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** This work was supported by the Bill and Melinda Gates Foundation [OPP1141535].

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2021 Hawken S *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Hawken S, Murphy MSQ, Ducharme R *et al.* **External validation of machine learning models including newborn metabolomic markers for postnatal gestational age estimation in East and South-East Asian infants [version 2; peer review: 3 approved, 1 approved with reservations]** Gates Open Research 2021, 4:164 <https://doi.org/10.12688/gatesopenres.13131.2>

**First published:** 29 Oct 2020, 4:164 <https://doi.org/10.12688/gatesopenres.13131.1>

**REVISED Amendments from Version 1**

We have revised the manuscript as suggested by the reviewers. In the updated version we provide more details on methodology, model development and validation and statistical analysis.

The figures have been updated and a new figure has been added showing observed vs. predicted gestational age in all three cohorts.

**Any further responses from the reviewers can be found at the end of the article**

## Introduction

Global- and population-level surveillance of preterm birth is challenging. Inconsistent use of international standards to define preterm birth and gestational age (GA) categories, the range of methods and timing used for GA assessment, and inadequate jurisdictional or national health data systems all hamper reliable population estimates of preterm birth<sup>1</sup>. As complications related to preterm birth continue to be the most common cause of mortality for children under five<sup>2</sup>, robust data on the burden of preterm birth are needed to maximize the effectiveness of resource allocation and global health interventions.

Newborn screening is a public health initiative that screens infants for rare, serious, but treatable diseases. Most of the target diseases are screened through the analysis of blood spots taken by heel-prick sampling. Samples are typically collected within the first few days after birth. Newborn samples are analyzed for a range of diseases, such as inborn errors of metabolism, hemoglobinopathies, and endocrine disorders, using tandem mass spectrometry, colorimetric and immunoassays, and high-performance liquid chromatography<sup>3</sup>. Postnatal GA algorithms derived from newborn characteristics and metabolic profiles have emerged as a novel method of estimating GA after birth. Using anonymized data from state and provincial newborn screening programs, three groups in North America have developed algorithms capable of accurately estimating infant GA to within  $\pm 1$  to 2 weeks<sup>4-6</sup>. Recent work to refine metabolic GA models<sup>7</sup>, as well as internally and externally validate their performance in diverse ethnic groups and in low-income settings, has demonstrated the potential of these algorithms beyond proof-of-concept applications<sup>8,9</sup>.

Published approaches to model development and validation to date have been carried out in cohorts of infants in North American settings<sup>4-6</sup>. Although internal validation of these models has been conducted among infants from diverse ethnic backgrounds<sup>4,8</sup>, external validation of model performance outside of the North American context is essential to evaluate the generalizability of models to low-income settings where they would have the most clinical utility. Birth weight, a significant covariate in all published models, is strongly correlated with GA and varies significantly by ethnicity<sup>10</sup>. Metabolic variations in newborn screening profiles that result from variation in genetic and *in utero* exposures may also affect the performance of established algorithms across ethnic or geographic subpopulations<sup>11</sup>. Importantly, as innovators seek to take this

work to scale, validation of metabolic models using data stemming from different laboratories is warranted. In this study, we sought to validate a Canadian metabolic GA estimation algorithm in data derived from newborn screening databases based in the Philippines and China.

## Methods

### Study design

This was a retrospective cohort study involving the secondary analysis of newborn screening data from three established newborn screening programs: Newborn Screening Ontario (Ottawa, Canada); Newborn Screening Reference Centre (Manila, the Philippines); and the Shanghai Neonatal Screening Center (Shanghai, China). Approval for the study was obtained from the Ottawa Health Sciences Network Research Ethics Board (20160056-01H), and research ethics committees at both the University of the Philippines Manila (2016-269-01), and the Xinhua Hospital, Shanghai (XHEC-C-2016). The need for express informed consent from participants was waived by the ethics committees for this retrospective study.

### Study population and data sources

Newborn Screening Ontario (NSO): a provincial newborn screening program that coordinates the screening of all infants born in Ontario, Canada. The program screens approximately 145,000 infants (>99% population coverage) annually for 29 rare conditions, including metabolic and endocrine diseases, sickle cell disease, and cystic fibrosis<sup>12</sup>. Newborn screening data collected between January 2012 and December 2014 were used in model building and internal validation. Further details on sample analysis methodology can be found here<sup>13</sup>.

Newborn Screening Reference Center: coordinates screening across six operations sites in the Philippines. The program screens approximately 1.5 million infants (68%) annually, offering two screening panels: a basic panel of six disorders at no cost, or an expanded panel of 28 disorders paid for by families. Data from this study were obtained from one of the newborn screening centers, the National Institutes of Health at the University of the Philippines Manila. Data were included for infants born between January 2015 and October 2016 who were screened using the expanded panel of 28 disorders. Disorders screened included metabolic disorders, and hemoglobinopathies.

Shanghai Neonatal Screening Center, National Research Center for Neonatal Screening and Genetic Metabolic Diseases: coordinates the screening of infants born in Shanghai, China. The program screens approximately 110,000 infants (>98%). Four screening tests - for phenylketonuria, congenital adrenal hyperplasia, hypothyroidism and Glucose-6-phosphate dehydrogenase deficiency - are provided at no cost. Expanded newborn screening is available but must be paid for by families. Data collected from the Shanghai Jiaotong University School of Medicine Xinhua Hospital were used for this study. Infants born between February 2014 and December 2016 with expanded newborn screening results were included.

### Reference GA assessment

In newborn cohorts from Ontario and China, GA was measured using gold-standard first trimester gestational dating ultrasound

in approximately 98% of cases, and was reported in weeks and days of gestation (for example 37 weeks and 6 days would be reported as 37.86 weeks). In the Philippines cohort, mothers who delivered in private hospitals generally received gestational dating ultrasounds while other infants' GAs were generally measured using Ballard Scoring, however individual-level data identifying which GA measurement method was used was not available. GA was reported in completed weeks (for example 37 weeks and 6 days would be recorded as 37 weeks). Therefore, for the Philippines cohort only, model-based GA estimates were rounded down in the same way for comparison to reference GA in the presentation of validation results.

Specific data elements used in this study from each respective newborn screening program are provided in Table 1. All of the analytes used in our analyses were routinely collected, including quantitative fetal and adult Hb levels. The Newborn Screening Ontario (Canada) disease panel included the greatest number of analytes. All analytes included in the newborn screening panels of the Newborn Screening Reference Centre (the Philippines) and the Shanghai Newborn Screening Program (China) were also available from Newborn Screening Ontario.

## Statistical methods

### Data preparation

Infants whose blood spot samples were collected more than 48 hours after birth were excluded from model development in the Ontario cohort. The reasons for this were twofold. First, the recommended screening sample collection window in Ontario is 24 to 48 hours after birth, and samples collected

beyond that window have increasingly heterogeneous analyte results influenced by multiple exogenous factors that cannot be statistically adjusted for. And second, our GA estimation models are intended to be applied in low to middle income countries (LMICs) where samples are expected to be collected almost exclusively within the first few hours after birth.

Infants who screened positive for one or more conditions (<1%) were excluded from model development and validation in all three cohorts, which had the effect of removing a large proportion of extreme outliers and atypical metabolic profiles.

In the China and Philippines cohorts, a larger proportion of samples were collected after 48 hours so we relaxed the exclusion criteria to >72 hours, to avoid exclusion of an unacceptably large proportion of samples from the external validation cohorts. Even after relaxing the criteria, samples in preterm infants were more likely to have been collected later than 72 hours in both China and the Philippines, which resulted in an artificially lower prevalence of preterm infants in the external validation cohorts. Samples from all three cohorts were excluded if they had missing sex, birth weight, gestational age, or missing analyte values. The proportion of subjects with missing analyte or covariate data was very low so missing data imputation was not undertaken.

In preparation for use in modeling, newborn screening analytes from Ontario, the Philippines and China were winsorized using an adapted "Tukey Fence" approach<sup>14</sup>. For each analyte, this involves assigning values more than three interquartile ranges

**Table 1. Newborn screening data used in model development.**

Newborn Screening Ontario, Canada	Newborn Screening Reference Centre, the Philippines	Shanghai Newborn Screening Program, China
Sex, birth weight <b>Hemoglobins</b> <b>F1, F, A</b> <b>Endocrine markers</b> <b>TSH</b> , 17OHP <b>Amino Acids</b> alanine, arginine, citrulline, glycine, leucine, methionine, ornithine, phenylalanine, tyrosine, valine, <b>Acyl-carnitines, Acylcarnitine ratios</b> C0, C2, C3, C4, C5, C6, C8, C10, C12, C14, C16, C18, C10:1, C12:1, C14OH, C14:1, <b>C14:2</b> , C16OH, <b>C16:1OH</b> , <b>C18OH</b> , C18:1, C18:2, <b>C18:1OH</b> , C3DC, C4DC, C4OH, C5DC, C5OH, <b>C5:1</b> , C6DC, C8:1, <b>Enzyme markers</b> <b>GALT, BIO</b> <b>Cystic fibrosis Markers</b> <b>IRT</b>	Sex, birth weight <b>Hemoglobins</b> <b>F1, F, A</b> <b>Endocrine markers</b> <b>TSH</b> , 17OHP <b>Amino Acids</b> alanine, arginine, citrulline, glycine, leucine, methionine, ornithine, phenylalanine, tyrosine, valine, <b>Acyl-carnitines, Acylcarnitine ratios</b> C0, C2, C3, C4, C5, C6, C8, C10, C12, C14, C16, C18, C10:1, C12:1, C14OH, C14:1, C16OH, <b>C16:1OH</b> , C18:1, C18:2, C3DC, C4DC, C4OH, C5DC, C5OH, C6DC, C8:1 <b>Enzyme markers</b> <b>BIO</b> <b>Cystic fibrosis Markers</b> <b>IRT</b>	Sex, birth weight  <b>Endocrine markers</b> <b>TSH</b> , 17OHP <b>Amino Acids</b> alanine, arginine, citrulline, glycine, leucine, methionine, ornithine, phenylalanine, tyrosine, valine <b>Acyl-carnitines, Acylcarnitine ratios</b> C0, C2, C3, C4, C5, C6, C8, C10, C12, C14, C16, C18, C10:1, C12:1, C14OH, C14:1, <b>C14:2</b> , C16OH, <b>C18OH</b> , C18:1, C18:2, C3DC, C4DC, C4OH, C5DC, C5OH, <b>C5:1</b> , C6DC, C8:1

TSH, thyroid stimulating hormone; 17OHP, 17 hydroxyprogesterone; GALT, galactose-1-phosphate uridyl transferase; IRT, Immuno-reactive trypsinogen, BIO, biotinidase. Analytes in **bold italics** are not available in one or more of the external validation cohorts. Further details on the Ontario newborn screening metabolites can be found here<sup>15</sup>.

above the third quartile (the upper Tukey fence) or below the first quartile (the lower Tukey fence), to the Tukey fence value. This approach preserves much of the “extremeness” of outliers but prevents extreme values from disproportionately impacting model fitting and GA estimation. The majority of measured analytes exhibit strongly right-skewed distributions, which was addressed through natural log transformation, which also stabilizes the variance, reducing the impact of heteroskedasticity. Finally, analyte levels and birth weight were normalized by subtracting the mean and dividing by the square root of the standard deviation for each variable (pareto scaling)<sup>16,17</sup> which centers all predictors to have a mean of zero, and scales them to reduce the impact of variations in dispersion of individual analytes across cohorts.

### Statistical modelling

#### Model Development in the Ontario, Canada model derivation cohort

The Ontario cohort was randomly divided into three sub-cohorts: 1) a model development sub-cohort (50%); 2) an internal validation sub-cohort (25%); and 3) a test sub-cohort (25%). Stratified random sampling was used to ensure that these three sub-cohorts retained the same distribution of GA and sex as the overall cohort.

A total of 47 newborn screening analytes, as well as sex, birth weight were used in the original Ontario model development. Model development included a ratio of fetal to adult hemoglobin calculated as  $(HbF+HbF1)/(HbF+HbF1+HbA)$ , to measure the proportion of normal fetal hemoglobin, relative to the proportion of total normal fetal + adult hemoglobin types. GA at birth (in weeks) determined by first trimester gestational dating ultrasound was the dependent variable. A subset of screening analytes were not available in the external cohorts (Table 1). Three main models were developed and internally validated in the Ontario cohort (Table 2). We developed restricted models including only the covariates available in each of the two external cohorts (Figure 1). All models were trained and internally validated within the Ontario datasets but deployed in the external cohorts.

For Models 1 and 3, birth weight was the strongest predictor and was modeled using a restricted cubic spline with five knots to allow for non-linearity of the association of birth weight with

gestational age. For Models 2 and 3, we included all pre-specified covariate main effects in the models, however we additionally identified the most predictive analytes using the metric of generalized partial Spearman correlation that detects non-linear and non-monotonic associations with GA, mutually adjusted for all other analytes and clinical covariates. Based on this partial Spearman correlation analysis there were seven analyte covariates that had distinctly stronger partial correlations with GA compared to all others. These seven analytes were modeled using restricted cubic splines with 5 knots. These were (in order of strength of partial spearman correlation): fetal-to-adult hemoglobin ratio, 17-OHP, C4DC, TYR, ALA, C5, and C5DC. For birthweight, and the seven strongest analyte predictors, knot placement was at the 5th, 27.5th, 50th, 72.5th, and 95th percentiles based on the Ontario population distribution of these analytes<sup>18</sup>. Hemoglobin measurements were unavailable in the China cohort, so the China restricted models included restricted cubic splines for the other top six analytes.

For Model 1, all covariates and pairwise interactions were included in the model without variable selection or regularization. For Models 2 and 3 we employed Elastic Net regularization, which employs two forms of penalization (called L1 and L2 regularization) to simultaneously estimate regression coefficients while also shrinking them towards zero to penalize the increase in model complexity from each additional term included in the model<sup>19</sup>. The Elastic Net regression methodology allows models to be fit with a large number of predictors, and allows models to be fit even in situations where models where the number of predictors exceeds the number of observations ( $p \gg n$ ). This regularization strategy provides strong protection against overfitting, and against the instability inherent in fitting models with a large number of predictors relative to the number of observations available for model fitting<sup>19,20</sup>.

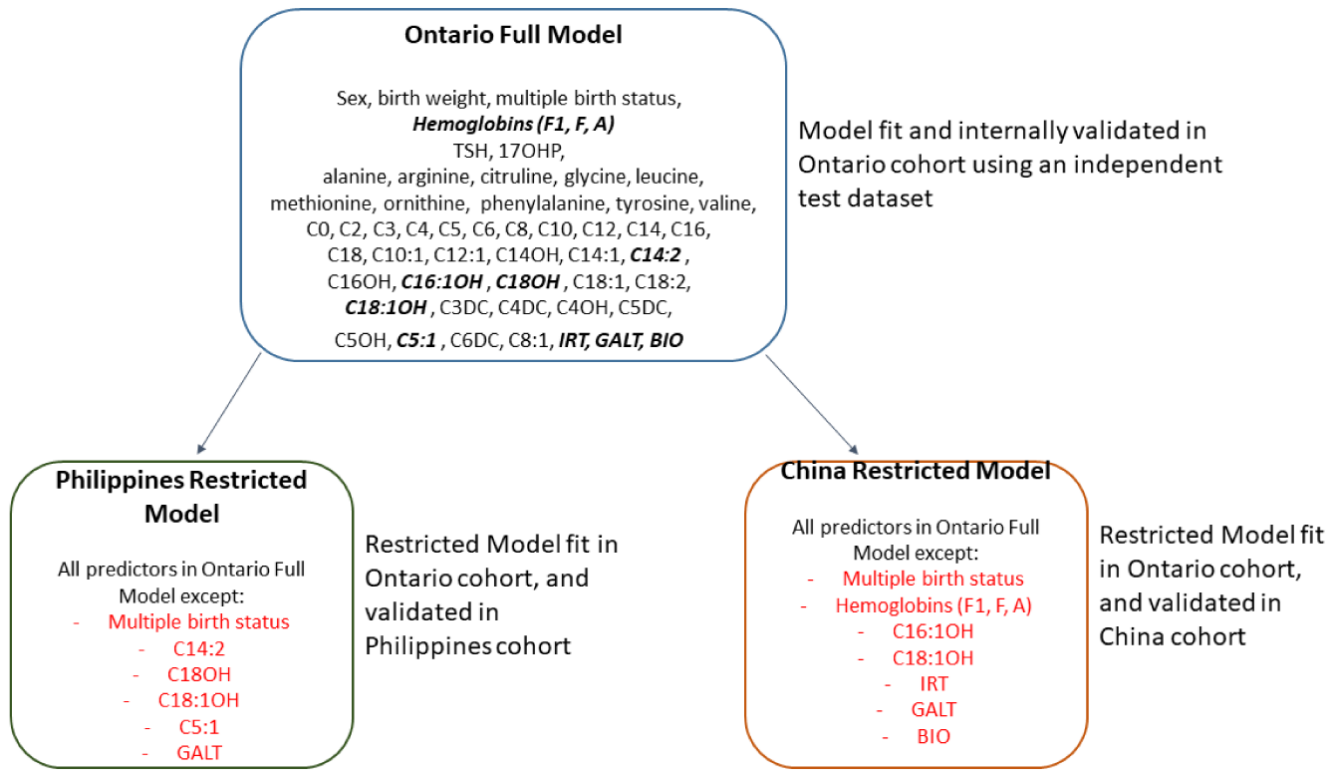
#### Validation of GA estimation models

We applied the identical approach in validating models internally (in Ontario test cohort) and externally (China and Philippines cohorts). Final regression model equations derived from the Ontario model development sub-cohort were used to calculate an estimated gestational age in the independent Ontario test cohort and in the China and Philippines cohorts. Model accuracy metrics were based on residual errors: the difference between model-estimated GA and reference GA. Although

**Table 2. Summary of models developed and validated.**

Model	Description
Model 1	Multivariable regression model including sex, birth weight and their interaction
Model 2	ELASTIC NET multivariable regression model including sex, analytes and pairwise interactions among predictors*
Model 3	ELASTIC NET multivariable regression model including sex, birth weight, analytes and pairwise interactions among* predictors
	*Restricted versions of Models 2 and 3 were derived based on the analyte measurements available in both the China and Philippines cohorts.





**Figure 1. Full model vs restricted models.**

mean square error (MSE) is typically the loss function used in maximum-likelihood model fitting for continuous outcomes, it is not necessarily the best metric for assessing average agreement in model validation, as it is based on sum of squared differences, and hence is sensitive to large and small residuals. Therefore, the primary metric we have presented is the mean absolute error (MAE). MAE is the average of absolute values of residuals (values of the model estimate minus the reference GA) across all observations. MAE reflects the average deviation of the model estimate compared to the reference estimate, expressed in the same units as GA (weeks).

$$MAE = \frac{1}{n} \sum_{i=1}^n |GA_{ref_i} - GA_{est_i}|$$

For completeness, as well as for comparability to other published validations, we also report the square root of the MSE (RMSE). Also known as the standard error of estimation, RMSE is also expressed in the same units as GA (weeks).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (GA_{ref_i} - GA_{est_i})^2}$$

Lower values of both MAE and RMSE reflects more accurate model estimated GA. For example, a reported MAE of 1.0 week reflects that the average discrepancy between model estimated GA and reference GA was 7 days. We also calculated the percentage of infants with GAs correctly estimated

within  $\pm 7$  and  $\pm 14$  days of reference GA. We assessed model performance overall and in important subgroups: preterm birth (<37 weeks GA), and small-for-gestational age: below the 10th (SGA10) and 3rd (SGA3) percentile for birth weight within categories of gestational week at delivery and infant sex based on INTERGROWTH-21 gestational weight for GA percentiles<sup>21</sup>. Parametric standard error estimates were not readily calculable for our performance metrics, therefore we calculated 95% bootstrap percentile confidence intervals based on the 2.5th and 97.5th percentiles over 1000 bootstrap replicates for each validation cohort<sup>22</sup>. Replication code is available as Extended data<sup>23</sup>. To assess the overall calibration of our models we produced residual plots to visually assess accuracy across the spectrum of observed and estimated GA. To assess calibration in the large, we estimated the overall preterm birth rate using estimated GA derived from each model and compared this to the true preterm birth rate based on observed GA in each validation cohort.

## Results

### Cohort characteristics

Cohort characteristics are presented in Table 3. In all, the final infant cohorts for model validation included 39,666 infants from Ontario, Canada, 82,909 infants from the Manila, Philippines cohort and 4,448 infants from the Shanghai, China cohort. Mean (SD) of clinically reported GAs for the Ontarian, Filipino and Chinese cohorts were 39.3 (1.6), 38.5 (1.4) and 38.9 (1.4) weeks, respectively. Preterm infants (GA <37 weeks) comprised

2,226/39,666 (5.6%) of the Ontario cohort, 3,832/82,909 (4.6%) of the Philippines cohort, and 215/4,448 (4.8%) of the China cohort.

### Model performance in Ontario, Canada

The most predictive individual analytes in Models 2 and 3 were fetal-to-adult hemoglobin ratio, 17-OHP, C4DC, TYR, ALA, C5, and C5DC. Modeling this subset of analytes using restricted cubic splines further increased their predictive power. Sensitivity analyses were conducted modeling the most predictive analytes using linear terms versus restricted cubic splines, as well a sensitivity analysis comparing models developed with only main effects versus allowing pairwise interactions, and in both cases model performance in terms of MAE/RMSE and adjusted R-square was markedly improved. Therefore, final models included splines and pairwise interactions.

Estimation of GA using Model 1 (including only sex and birth weight) yielded an MAE (95% CI) of 0.96 (0.96, 0.97) weeks in the Ontario cohort, indicating that the model provided GA estimates that were accurate to within  $\pm 7$  days of reference GA. Model 2, (including sex and metabolomic markers), was accurate within an average of  $\pm 6$  days (MAE 0.79 (0.79, 0.80)

weeks). Model 3, which included sex, birth weight and metabolomic markers was the most accurate, estimating GA within about  $\pm 5$  days of ultrasound-assigned GA (MAE 0.71 (0.71, 0.72) weeks), and estimated GA within  $\pm 1$  week in 74.6% of infants overall. Model 3 was the best performing model in preterm infants (GA<37 weeks), with an MAE (95% CI) of 1.03 (0.99, 1.06) compared to MAE of 1.78 (1.73, 1.82) for Model 1 and 1.25 (1.21, 1.29) for Model 2. In contrast, Model 2, which did not include birth weight, performed the best in growth restricted infants, with MAE of 0.90 (0.85 to 0.94) in SGA10 infants and 1.03 (0.92, 1.13) in SGA3 infants, and was slightly better than Model 3, which did include birth weight. However, Model 1, including only sex and birth weight, was extremely inaccurate in both SGA10 and SGA3 infants with MAE of 2.71 (2.66, 2.76) and 3.84 (3.75, 3.95) respectively (Table 4).

Restricted models including the subset of analytes available in the Philippines and China cohorts performed comparably to the unrestricted Ontario models overall. When applied to the Ontario internal validation cohort, accuracy of both the China- and Philippines-restricted models was slightly lower overall and lower in important subgroups, most notably in preterm and growth restricted infants for cohort (Model 2 and Model 3 China restricted and Philippines restricted) (Table 4).

**Table 3. Cohort Characteristics.**

	Canada n=39,666 (Ontario test cohort)	Philippines n=82,909	China n=4,448
<b>Sex, n (%)</b>			
Male	19,536 (49.3%)	42,867 (51.7%)	2,351 (52.9 %)
Female	20,130 (50.7%)	40,042 (48.3%)	2,097 (47.1 %)
<b>Birth weight (g), mean<math>\pm</math>SD</b>			
Overall	3,379 $\pm$ 530	3,001 $\pm$ 452	3,337 $\pm$ 437
Term infants only	3,431 $\pm$ 476	3,044 $\pm$ 414	3,369 $\pm$ 407
Preterm infants only	2,504 $\pm$ 623	2,250 $\pm$ 539	2,710 $\pm$ 536
Low birth weight (<2500g), n (%)	1812 (4.6%)	8423 (10.2%)	128 (2.9%)
SGA (<10 th Centile), n (%)	1,561 (3.9%)	11,295 (13.6%)	123 (2.8%)
SGA (<3 rd Centile), n (%)	363 (0.9%)	3,407 (4.1%)	19 (0.4 %)
LGA (>90 th Centile), n (%)	8734 (22.0%)	4780 (5.8%)	741 (16.7%)
Completed gestational age (wks), mean $\pm$ SD	39.3 $\pm$ 1.6	38.5 $\pm$ 1.4	38.9 $\pm$ 1.4
Term ( $\geq$ 37 wks), n (%)	37,440 (94.4%)	79,077 (95.4%)	4,233 (95.2%)
Late Preterm (32–36 wks), n (%)	2,049 (5.2%)	3,566 (4.3%)	197 (4.4 %)
Very Preterm (28–31 wks), n (%)	126 (0.3%)	233 (0.3%)	11 (0.3 %)
Extremely Preterm (<28 wks), n (%)	51 (0.1%)	33 (0.0%)	7 (0.2 %)

SGA, small for gestational age (lowest 10 and 3 centiles within gestational age and sex strata, calculated in the Ontario cohort using Intergrowth-21 centiles and applied uniformly in the Ontario, China and Philippines cohorts)

**Table 4. Summary of model performance to estimate gestational age.**

Models	Ontario Cohort				Philippines Cohort				China Cohort			
	Overall, N=39,666	<37 weeks, N=2,226	SGA10, N=1,561	SGA3, N=363	Overall, N=82,909	<37 weeks, N=3,828	SGA10, N=11,294	SGA3, N=3,407	Overall, N=4,448	<37 weeks, N=215	SGA10, N=123	SGA3, N=19
<b>Model 1: Sex and Birth weight</b>												
MAE (CI)	0.96 (0.96, 0.97)	1.78 (1.73, 1.82)	2.71 (2.66, 2.76)	3.84 (3.75, 3.95)	0.96 (0.95, 0.97)	1.87 (1.83, 1.92)	1.47 (1.45, 1.49)	2.65 (2.62, 2.69)	0.90 (0.87, 0.92)	2.02 (1.76, 2.33)	2.72 (2.60, 2.85)	3.90 (3.66, 4.15)
RMSE (CI)	1.23 (1.22, 1.24)	2.13 (2.07, 2.18)	2.86 (2.80, 2.91)	3.96 (3.86, 4.06)	1.30 (1.29, 1.31)	2.31 (2.24, 2.37)	1.83 (1.80, 1.86)	2.84 (2.80, 2.89)	1.23 (1.17, 1.29)	2.85 (2.37, 3.39)	2.82 (2.68, 2.96)	3.93 (3.69, 4.19)
% +/− 1 wk (CI)	60.0 (59.5, 60.5)	30.2 (28.5, 32.0)	0.4 (0.1, 0.7)	0.6 (0.0, 1.4)	78.2 (78.0, 78.5)	43.4 (42.0, 45.0)	61.0 (60.1, 61.8)	5.0 (4.3, 5.8)	65.5 (64.1, 66.9)	33.4 (27.2, 39.7)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
% +/− 2 wks (CI)	90.9 (90.6, 91.2)	60.5 (58.4, 62.6)	21.7 (19.5, 23.8)	1.1 (0.3, 2.3)	95.5 (95.3, 95.6)	68.9 (67.4, 70.4)	85.4 (84.7, 86.0)	53.6 (51.8, 55.3)	91.7 (90.8, 92.5)	58.6 (52.1, 65.2)	13.7 (7.7, 20.2)	0.0 (0.0, 0.0)
<b>Model 2: Sex and Analyses</b>												
<b>Ontario</b>												
MAE (CI)	0.79 (0.79, 0.80)	1.25 (1.21, 1.29)	0.90 (0.85, 0.94)	1.03 (0.92, 1.13)	-	-	-	-	-	-	-	-
RMSE (CI)	1.02 (1.01, 1.03)	1.57 (1.51, 1.62)	1.19 (1.13, 1.26)	1.42 (1.27, 1.59)	-	-	-	-	-	-	-	-
% +/− 1 wk (CI)	69.4 (69.0, 69.9)	46.9 (45.0, 48.9)	66.6 (64.0, 69.0)	61.9 (56.9, 67.1)	-	-	-	-	-	-	-	-
% +/− 2 wks (CI)	95.1 (94.9, 95.4)	80.3 (78.8, 82.0)	91.7 (90.3, 93.1)	88.1 (84.6, 91.6)	-	-	-	-	-	-	-	-
<b>Philippines-restricted</b>												
MAE (CI)	0.81 (0.80, 0.81)	1.28 (1.24, 1.32)	0.92 (0.89, 0.97)	1.04 (0.93, 1.14)	1.02 (1.02, 1.03)	1.96 (1.91, 2.01)	1.08 (1.06, 1.09)	1.18 (1.14, 1.2)	-	-	-	-
RMSE (CI)	1.04 (1.03, 1.05)	1.59 (1.54, 1.64)	1.22 (1.16, 1.28)	1.42 (1.26, 1.57)	1.37 (1.36, 1.37)	2.43 (2.37, 2.50)	1.44 (1.42, 1.47)	1.57 (1.53, 1.62)	-	-	-	-
% +/− 1 wk (CI)	68.5 (68.1, 69.0)	45.4 (43.4, 47.4)	63.8 (61.4, 66.1)	61.6 (56.2, 66.7)	75.3 (75.1, 75.6)	41.93 (40.3, 43.5)	73.1 (72.3, 73.9)	70.0 (68.7, 71.5)	-	-	-	-
% +/− 2 wks (CI)	94.8 (94.6, 95.0)	80.0 (78.3, 81.7)	91.2 (89.7, 92.6)	88.1 (84.6, 91.5)	94.3 (94.1, 94.5)	69.3 (67.8, 70.7)	92.3 (91.8, 92.8)	89.8 (88.8, 90.7)	-	-	-	-
<b>China-restricted</b>												
MAE (CI)	0.87 (0.87, 0.88)	1.48 (1.44, 1.52)	0.98 (0.94, 1.03)	1.13 (1.03, 1.25)	-	-	-	-	1.07 (1.04, 1.10)	2.49 (2.21, 2.80)	1.00 (0.84, 1.15)	1.03 (0.63, 1.45)
RMSE (CI)	1.12 (1.11, 1.13)	1.80 (1.75, 1.85)	1.31 (1.25, 1.38)	1.58 (1.40, 1.76)	-	-	-	-	1.44 (1.38, 1.52)	3.32 (2.80, 3.86)	1.34 (1.15, 1.52)	1.36 (0.84, 1.92)
% +/− 1 wk (CI)	64.9 (64.5, 65.1)	39.8 (37.9, 42.0)	60.9 (58.3, 63.4)	57.1 (51.6, 62.3)	-	-	-	-	56.2 (54.7, 57.6)	16.8 (11.9, 21.8)	62.0 (53.4, 70.9)	58.9 (56.4, 82.4)
% +/− 2 wks (CI)	93.0 (92.8, 93.3)	71.8 (70.1, 73.8)	89.0 (87.3, 90.6)	83.2 (79.1, 87.3)	-	-	-	-	86.8 (85.9, 87.8)	46.9 (40.3, 53.3)	86.1 (79.8, 91.7)	89.3 (74.5, 100.0)
<b>Model 3: Sex, Birth weight and Analyses</b>												
<b>Ontario</b>												
MAE (CI)	0.71 (0.71, 0.72)	1.03 (0.99, 1.06)	1.13 (1.09, 1.17)	1.48 (1.37, 1.60)	-	-	-	-	-	-	-	-
RMSE (CI)	0.92 (0.91, 0.93)	1.32 (1.27, 1.37)	1.42 (1.36, 1.47)	1.81 (1.67, 1.96)	-	-	-	-	-	-	-	-
% +/− 1 wk (CI)	74.6 (74.2, 75.1)	58.0 (55.9, 60.2)	50.5 (48.2, 53.0)	35.8 (30.6, 40.6)	-	-	-	-	-	-	-	-
% +/− 2 wks (CI)	97.0 (96.8, 97.2)	88.8 (87.5, 90.3)	85.0 (83.2, 86.6)	73.8 (69.1, 78.1)	-	-	-	-	-	-	-	-
<b>Philippines-restricted</b>												
MAE (CI)	0.72 (0.71, 0.72)	1.04 (1.00, 1.07)	1.15 (1.10, 1.19)	1.49 (1.38, 1.60)	0.90 (0.90, 0.91)	1.49 (1.45, 1.53)	0.97 (0.96, 0.99)	1.27 (1.23, 1.3)	-	-	-	-
RMSE (CI)	0.92 (0.91, 0.93)	1.33 (1.28, 1.38)	1.43 (1.37, 1.49)	1.82 (1.68, 1.97)	1.21 (1.20, 1.22)	1.92 (1.86, 1.98)	1.31 (1.29, 1.33)	1.63 (1.60, 1.67)	-	-	-	-
% +/− 1 wk (CI)	74.5 (74.1, 74.9)	57.0 (54.9, 59.0)	50.1 (47.6, 52.8)	35.2 (30.3, 39.8)	80.4 (80.1, 80.7)	57.2 (55.8, 58.6)	77.4 (76.5, 78.2)	65.1 (63.5, 66.6)	-	-	-	-
% +/− 2 wks (CI)	69.9 (96.8, 97.1)	88.2 (86.8, 89.6)	85.0 (83.2, 86.8)	73.5 (68.5, 77.8)	97.0 (96.8, 97.1)	84.1 (82.8, 85.3)	94.7 (94.2, 95.1)	88.3 (87.2, 89.3)	-	-	-	-
<b>China-restricted</b>												
MAE (CI)	0.76 (0.75, 0.76)	1.12 (1.08, 1.15)	1.26 (1.21, 1.31)	1.65 (1.54, 1.78)	-	-	-	-	0.89 (0.86, 0.91)	1.74 (1.49, 2.05)	1.48 (1.32, 1.64)	2.04 (1.68, 2.45)
RMSE (CI)	0.97 (0.96, 0.98)	1.43 (1.37, 1.48)	1.54 (1.48, 1.59)	1.98 (1.84, 2.13)	-	-	-	-	1.20 (1.14, 1.27)	2.69 (2.16, 3.28)	1.70 (1.54, 1.88)	2.19 (1.78, 2.65)
% +/− 1 wk (CI)	71.8 (71.3, 72.2)	53.1 (50.1, 55.2)	44.1 (41.7, 46.6)	29.4 (24.7, 34.2)	-	-	-	-	64.7 (63.3, 66.0)	43.4 (36.9, 50.2)	30.9 (22.7, 39.6)	4.9 (0.0, 16.7)
% +/− 2 wks (CI)	96.3 (96.1, 96.5)	85.1 (83.5, 86.7)	83.2 (81.2, 85.1)	69.9 (64.8, 74.9)	-	-	-	-	92.7 (92.0, 93.4)	72.1 (66.2, 78.0)	77.3 (69.3, 84.7)	47.2 (25.0, 70.7)

Data are presented as the mean and 2.5<sup>th</sup> and 97.5<sup>th</sup> bootstrap percentiles for MAE, RMSE and the percentage of model estimates within 1 and 2 weeks of ultrasound GA for 1000 bootstrap samples generated from each cohort



### External validation of model performance in the Philippines cohort

When applied to infant samples from the Philippines cohort, Model 1 yielded a MAE (95% CI) of 0.96 (0.95, 0.97). Accuracy was slightly decreased for Model 2, with MAE of 1.02 (1.02, 1.03). Model 3 which included sex, birth weight and screening analytes available in the Philippines database performed the best, with an MAE of 0.90 (0.90, 0.91). Model 3 was also the best performing model in preterm infants, with MAE of 1.49 (1.45, 1.53) compared to 1.87 (1.83, 1.92) for Model 1 and 1.96 (1.91, 2.01) for Model 2. Model 3 also yielded the most accurate GA estimates in growth restricted infants, with MAE of 0.97/1.27 for SGA10/SGA3 infants compared to 1.47/2.65 for Model 1 and 1.08/1.18 for Model 2 for SGA10/SGA3 infants (Table 4).

Based on GA estimates from Model 3, the estimated preterm birth prevalence was 4.2% (95% CI: 4.1%, 4.4%), compared to the true prevalence of 4.3% using reference GA in the Philippines cohort. Both Model 1 and Model 2 overestimated the preterm birth rate, at 5.1% (4.9%, 5.4%) and 5.0% (4.9%, 5.2%), respectively.

### External validation of model performance in the China cohort

In the China cohort, Model 1 estimated GA to within 6 days overall, with an MAE of 0.90 (0.87, 0.92). Model 3 demonstrated similar accuracy to Model 1 with MAE of 0.89 (0.86, 0.91), and Model 2 performed the worst with MAE of 1.07 (1.04, 1.10). Model 3 performed the best in preterm infants, with MAE of 1.74 (1.49, 2.05) versus 2.49 (2.21, 2.80) for Model 2 and 2.02 (1.76, 2.33) for Model 1. In growth restricted infants, Model 2 was the most accurate, with MAE of 1.00/1.03 in SGA10/SGA3 infants compared to 1.48/2.04 for Model 3 and 2.72/3.90 for Model 1.

Based on GA estimates from Model 3, the estimated preterm birth prevalence was 4.2% (95% CI: 3.7%, 4.8%), and Model 1, which demonstrated similar overall accuracy, estimated a rate of 4.9% (4.3%, 5.6%). Model 2, the least accurate of the three in the China cohort, underestimated the preterm birth rate to be 3.6% (2.9%, 4.3%), compared to the actual preterm birth rate of 4.8% based on reference GA in the China cohort.

### Model performance across spectrum of GA

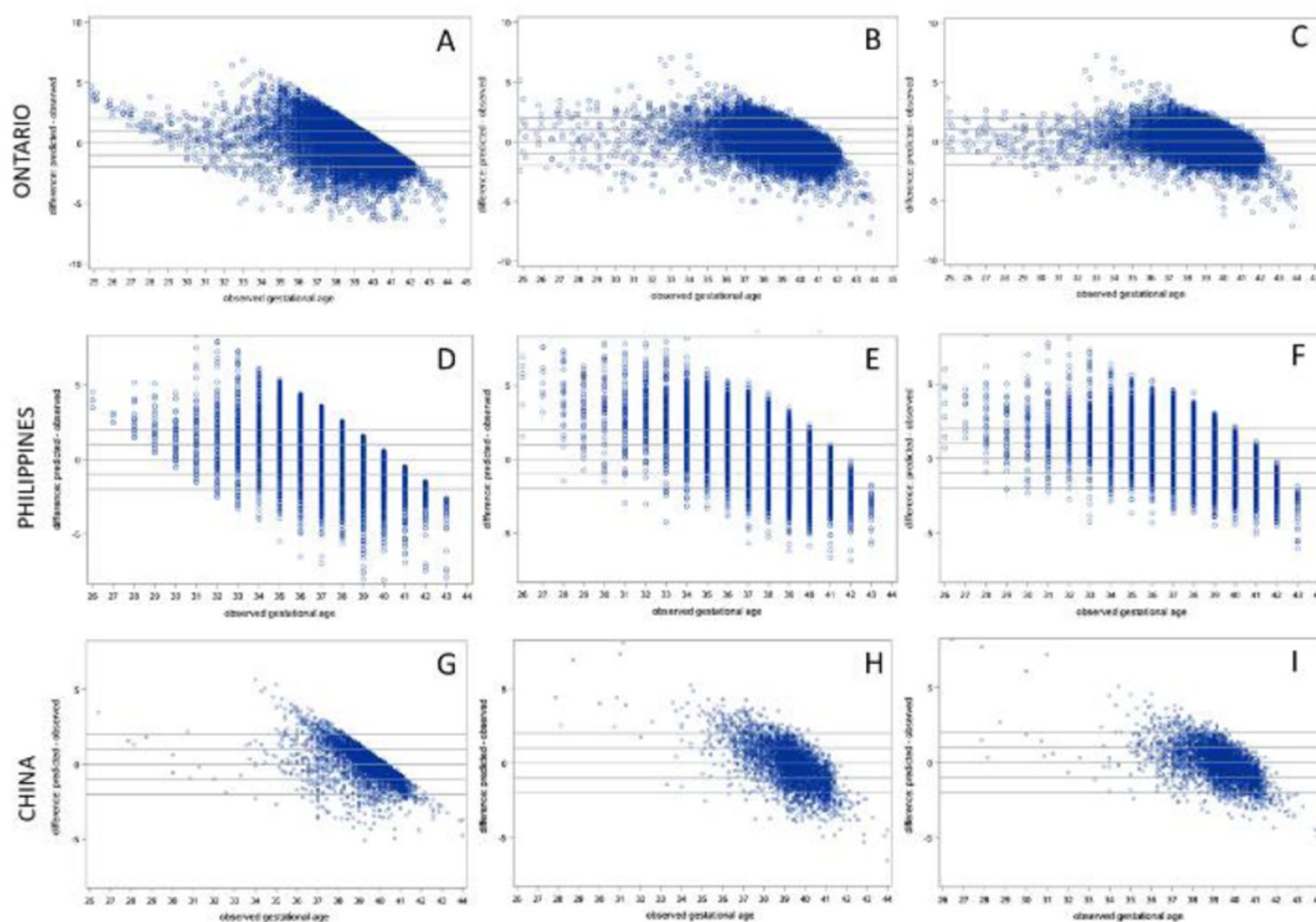
Scatter plots of observed GA versus estimated GA for all three models in the Ontario, Philippines and China cohorts are presented in Figure 2, which shows that in general, lower (preterm) GAs tend to be overestimated by all three models when applied to all cohorts. In all models applied to both external validation cohorts, GA estimates were most accurate in term infants and accuracy tended to be lower in preterm infants. Across the spectrum of ultrasound-assigned GA, Model 3 provided the most accurate estimates overall (Figure 3).

### Discussion

In this study, we demonstrated that the performance of gestational dating algorithms developed in a cohort of infants from Ontario, Canada including newborn screening metabolomic

markers from dried blood-spot samples was attenuated when the models were applied to data derived from external laboratories and populations. When these Canadian-based models were tailored to the analytes available from newborn screening programs in Shanghai, China and Manila, Philippines, the models were less accurate in estimating absolute GA in infant cohorts from these locations than when the same models were applied to an Ontario infant cohort. Models including analytes generally demonstrated improved accuracy over those relying on sex and birth weight alone, but the added benefit of models including blood-spot metabolomic markers (Model 2 and Model 3) was not substantial when looking at overall accuracy. However, our models that included metabolomic markers did demonstrate markedly improved accuracy over sex and birth weight in important subgroups (preterm and growth restricted infants), with Model 3 which included sex, birth weight and metabolomic markers demonstrating the best performance in almost all settings. The exception to this observation was in growth restricted infants (SGA10 and SGA3), where Model 2 often performed the best. This is not surprising, as birth weight is clearly a misleading predictor of GA in growth restricted infants, and although Model 3 still outperformed Model 1, its accuracy was impacted by the inclusion of birth weight in addition to metabolomic markers. Therefore, the decision of whether to prefer Model 2 or Model 3 may hinge on whether the prevalence of growth restriction is known to be high in the setting where the GA estimation algorithm is to be deployed. When we compared preterm birth rates (<37 weeks GA) calculated based on model estimates, to those calculated based on reference GA in each cohort, the model-based estimates from the best performing model (Model 3) agreed reasonably well with the reference preterm birth rates (4.2% vs 4.8% for China and 4.2% vs 4.6% for the Philippines). Unfortunately, as with any dichotomization of a continuous measure (GA), there are significant edge effects that can contribute to perceived misclassification (e.g. GA of 36.9 weeks is classified as preterm while a GA of 37.1 weeks is classified as term, despite a difference in GA of only about 1 day).

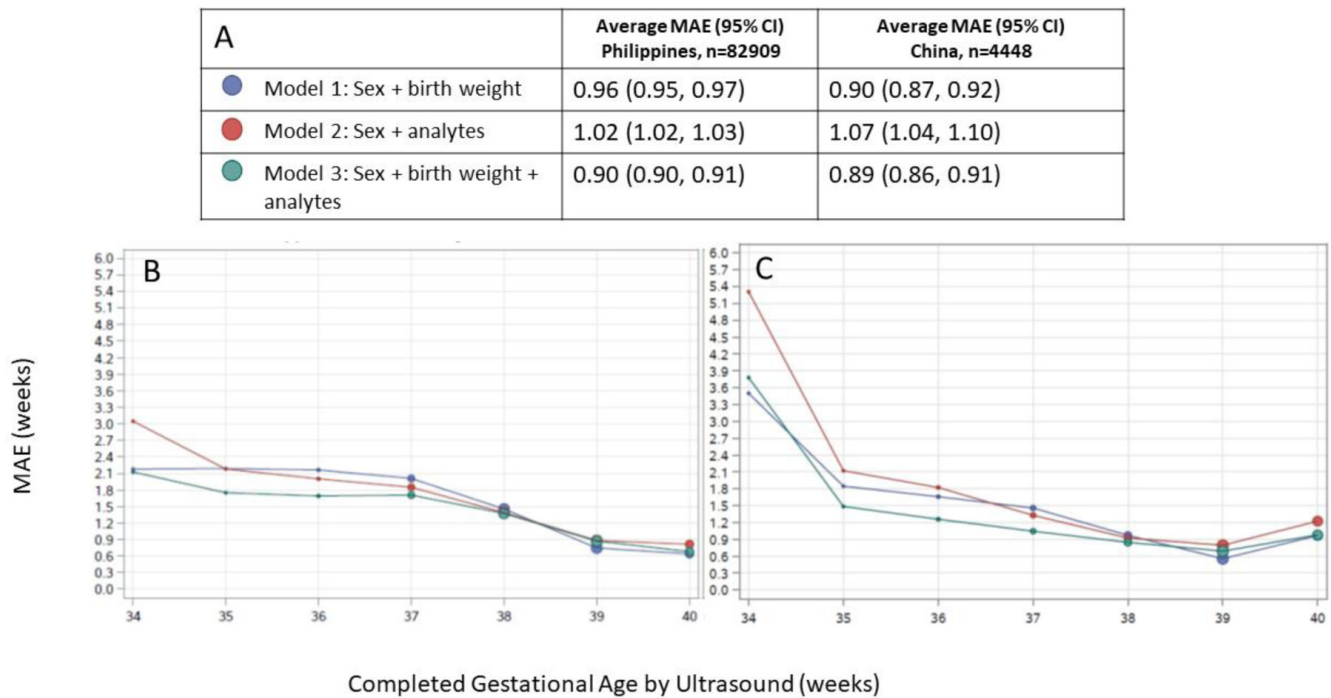
There are several reasons why the metabolic gestational dating algorithm we developed from a North American newborn cohort may not have performed as well using data derived from other infant populations. First, as observed in the differences in Model 1's performance across the three cohorts, the predictive utility of anthropomorphic measurements for estimating GA may vary across populations. Second, metabolic profiles may be influenced by the differences in genetic and environmental exposures experienced by each cohort. Previous validation of our models among infants born in Ontario to landed-immigrant mothers from eight different countries across Asia and North and Sub-Saharan Africa suggested that inherent biological differences may not be a significant contributor to newborn metabolic data and the performance of our algorithms<sup>8</sup>. However in an external validation of previously developed GA estimation models in a prospective cohort from South Asia<sup>9</sup>, the drop in performance was more pronounced, despite the centralized analysis of samples in the Ontario Newborn Screening lab. Third, variations in the clinical measures of GA used across the cohorts



**Figure 2. Residual plots of predicted – observed by ultrasound-assigned gestational age.** Models applied to test cohort from Ontario: (A) Model 1 (sex + birthweight), (B) Model 2 (sex + analytes) (C) Model 3 (sex + birthweight + analytes). Models applied to cohort from the Philippines: (D) Model 1, (E) Model 2, and (F) Model 3. Models applied to cohort from China: (G) Model 1, (H) Model 2, and (I) Model 3.

may have impeded the accuracy of our algorithms. Our GA models were developed with first trimester ultrasound-assigned GA as the dependent variable. Whereas first trimester ultrasounds were the gold standard in the Ontario and China cohorts, GAs for the Philippines cohort were determined by a mixture of gestational dating ultrasound and Ballard scores, and were only available to the nearest completed week of GA. Ballard tends to overestimate gestational age with a wide margin of error, particularly in preterm infants<sup>24</sup>. Lastly, and perhaps most importantly, variations in the collection procedures and analytical methods used by each of the newborn screening programs are likely to have impacted the measurable relationship between the analytes and newborn GA. At the newborn screening program in Shanghai, China, samples were collected, on average, about one day later than samples used for model development, particularly among preterm infants with the majority being collected between 48–72 hours. Variations in temperature, climate, sample handling, and storage among the three newborn screening laboratories may have also contributed to heterogeneity of findings. The screening laboratories in Ontario,

Shanghai, China and Manila also varied with respect to sample collection and handling, as well as lab equipment, assays and reagents to quantify the measured analytes. Reduced performance of models in the China and Philippines settings was likely due to a combination of these sources of heterogeneity including genetic, environmental and non-biological (eg laboratory-based) variation, which cannot easily be teased apart. We attempted to address these sources of heterogeneity and bias through our data preparation steps, which involved local standardization of analyte values and birth weight. Extreme outliers, skewed distributions, heteroscedasticity, and systematic biases within and between laboratories are all factors that may obscure biological signals. Normalization and other data pre-processing steps are therefore crucial to the analysis of metabolomic data, and we continue to investigate the impact of alternative data normalization techniques in improving the generalizability of our GA estimation models, while still taking care to preserve the biological signals of interest. This is an active area of active research as it relates to the use of 'omics data in prognostic models more generally<sup>25,26</sup>.



**Figure 3. Agreement between algorithmic gestational age estimations compared to ultrasound-assigned gestational age. (A)** Legend, and overall MAE (95% CI) for each model applied to data from the Philippines and China. Dot size in plots is proportional to sample size in each gestational age category. Performance of each model by ultrasound-assigned gestational age when applied to data from **(B)** the Philippines **(C)** China. MAE, mean absolute error (average absolute deviation of observed vs. predicted gestational age in weeks).

Our study has several strengths and limitations. Notable strengths include the size of our Ontario, China and Philippines cohorts, the commonality of a preponderance of the analytes across populations, the ability to tailor models to the specific analytes available for each cohort, and the methodological rigor we imposed in our modeling and validation. Limitations include the inability to examine the impact of environmental factors (socio-economic conditions, dietary and environmental exposures during pregnancy), variations in approaches to newborn screening that may not have been accounted for in our analyses, and generally smaller sample sizes for more severely preterm children. The preterm birth rate in our external validation cohorts (Philippines: 4.6%, China: 4.8%) was lower than published estimates (Philippines: 15%, China: 7.1%)<sup>27</sup>. This was in part due to selection bias because the expanded screening panel was not universally covered, requiring the family to pay in many cases, which may have led to infants in our validation cohorts being more likely to be from affluent families and/or from urban versus rural areas. Another source of bias was that samples in infants born preterm in China, and to a lesser extent the Philippines, were much more likely to be collected later than in term infants, and often after 48-72 hours. Samples collected beyond 48 hours were more heterogeneous than samples collected within 48 hours but excluding these would have excluded an unacceptably large proportion of infants overall, especially preterm infants. Our compromise approach of relaxing the criteria to exclude

samples that were collected later than 72 hours after birth, included more infants at the cost of increased heterogeneity, but even so, still excluded a disproportionate number of preterm infants in the China and Philippines. A combination of these factors likely contributed to the lower than expected preterm birth rates observed in China and in the Philippines validation cohorts, as well as leading to decreased apparent model performance due to more heterogeneous samples (collected 48-72 hours after birth) being included.

While there are numerous options currently available to health care providers to determine postnatal GA, none are as accurate as first trimester dating ultrasound<sup>28</sup>. Where access to antenatal dating technologies are limited, and the reliability of postnatal assessments is variable, there is a recognized need for new and innovative approaches to ascertaining population-level burdens of preterm birth in low resource settings<sup>28,29</sup>. Metabolic GA estimation models in particular have proven particularly promising<sup>29</sup>, and we continue to refine and evaluate these models in a variety of populations<sup>6,7,15</sup> and laboratories in an effort to ready this innovation for broader application. The findings of this study suggest that the accuracy of metabolic gestational dating algorithms may be improved where newborn samples can be analyzed in the same laboratories from which the algorithms were originally derived and underscore our previous findings of their potential particularly among low birth weight or SGA

infants<sup>7</sup>. Validation of our ELASTIC NET machine learning models is also being completed in prospective cohorts of infants from low-income settings in Bangladesh and Zambia<sup>15</sup>, with validation of previously developed models already completed in Bangladesh<sup>9</sup>. The effects of laboratory-specific variables are being mitigated through the standardization of collection and analytical procedures applied to newborn samples; preliminary results are promising. As efforts to optimize gestational dating algorithms based on newborn metabolic data continue, and innovators seek to take this work to scale, future work should identify opportunities to develop algorithms locally where newborn screening laboratories exist, and to build capacity in low resource settings for these purposes.

## Data availability

### Underlying data

The data from Ontario, Canada used to develop models, and the data for the external validation cohorts in which model performance was evaluated were obtained through bilateral data sharing agreements with the Ontario Newborn Screening Program and BORN Ontario, and newborn screening laboratories at Xinhua Hospital in Shanghai, China and University of the Philippines, Manila, Philippines. These data sharing agreements prohibited the sharing of patient-level data beyond our research team.

### Ontario data

Those wishing to request access to Ontario screening data can contact [newbornscreening@cheo.on.ca](mailto:newbornscreening@cheo.on.ca), and the request will be assessed as per NSO's data request and secondary use policies. For more information, please visit the NSO website: <https://www.newbornscreening.on.ca/en/screening-facts/screening-faq> ('What happens when a researcher wants to access stored samples for research'); <https://www.newbornscreening.on.ca/en/privacy-and-confidentiality>.

## Philippines data

Researchers can request access to the de-identified data (sex, birthweight, gestational age and screening analyte levels) from the Philippines for future replication of the study by sending a request letter to the Director of Newborn Screening Reference Center stating the study objectives in addition to:

- A copy of the study protocol approved by a technical and ethics review board that includes methods and statistical analysis plans;
- Full name, designation, affiliation of the person with whom the data will be shared; and,
- Time period that the data will be accessed.

Data requests must be addressed to: Dr. Noel R. Juban, Director of the Newborn Screening Reference Center National Institutes of Health, Unit 304 New Gold Bond Building, 1579 F. T. Benitez St, Ermita, Manila, Philippines, [info@newbornscreening.ph](mailto:info@newbornscreening.ph).

## China Data

Researchers can request access to the de-identified data (sex, birthweight, gestational age, age at sample collection, and screening analyte levels) from China by sending a written request to the corresponding author, Dr. Steven Hawken ([shawken@ohri.ca](mailto:shawken@ohri.ca)), which must include a copy of the study protocol and approval from the researcher's local ethics board.

## Extended data

SAS and R code for data preparation and cleaning, model fitting and external model validation are available at: <https://github.com/stevenhawken/Gates-Repository-China-Phil>.

Archived code at time of publication: <http://doi.org/10.5281/zenodo.4085320><sup>23</sup>.

License: [GNU General Public License v3](#).

## References

- March of Dimes, PMNCH, Save the Children, WHO: **Born Too Soon: The Global Action Report on Preterm Birth**. Geneva, Switzerland; 2012.  
[Reference Source](#)
- Liu L, Oza S, Hogan D, et al.: **Global, regional, and national causes of under-5 mortality in 2000-15: an updated systematic analysis with implications for the Sustainable Development Goals**. *Lancet*. 2016; **388**(10063): 3027-3035.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Pitt JJ: **Newborn screening**. *Clin Biochem Rev*. 2010; **31**(2): 57-68.  
[PubMed Abstract](#) | [Free Full Text](#)
- Jelliffe-Pawlowski LL, Norton ME, Baer RJ, et al.: **Gestational dating by metabolic profile at birth: a California cohort study**. *Am J Obstet Gynecol*. 2016; **214**(4): 511.e1-511.e13.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ryckman KK, Berberich SL, Dagle JM: **Predicting gestational age using neonatal metabolic markers**. *Am J Obstet Gynecol*. 2016; **214**(4): 515.e1-515.e13.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wilson K, Hawken S, Potter BK, et al.: **Accurate prediction of gestational age using newborn screening analyte data**. *Am J Obstet Gynecol*. 2016; **214**(4): 513.e1-513.e9.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Wilson K, Hawken S, Murphy MSQ, et al.: **Postnatal Prediction of Gestational Age Using Newborn Fetal Hemoglobin Levels**. *EBioMedicine*. 2017; **15**: 203-209.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hawken S, Ducharme R, Murphy MSQ, et al.: **Performance of a postnatal metabolic gestational age algorithm: a retrospective validation study among ethnic subgroups in Canada**. *BMJ Open*. 2017; **7**(9): e015615.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Murphy MSQ, Hawken S, Cheng W, et al.: **External validation of postnatal gestational age estimation using newborn metabolic profiles in Matlab, Bangladesh**. *eLife*. 2019; **8**: e42627.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ray J, Sgro M, Mamdani M, et al.: **Birth weight curves tailored to maternal world region**. *J Obstet Gynaecol Canada*. 2012; **34**(2): 159-171.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Ryckman KK, Berberich SL, Shchelochkov OA, et al.: **Clinical and environmental influences on metabolic biomarkers collected for newborn screening**. *Clin Biochem*. 2013; **46**(1-2): 133-138.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Newborn Screening Ontario: **Background on Newborn Screening Ontario (NSO) What is newborn screening? What is NSO?** 2017; 8330.  
[Reference Source](#)



13. Murphy MSQ, Hawken S, Cheng W, *et al.*: **Metabolic profiles derived from residual blood spot samples: A longitudinal analysis [version 1; peer review: 2 approved]**. *Gates open Res.* 2018 [cited 2019 Oct 1]; **2**: 28. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Tukey JW: **Exploratory data analysis**. 1st ed. Reading: Addison-Wesley Publishing Company; 1977. [Reference Source](#)
15. Murphy MSQ, Hawken S, Atkinson KM, *et al.*: **Postnatal gestational age estimation using newborn screening blood spots: a proposed validation protocol**. *BMJ Glob Heal.* 2017; **2**(2): e000365. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
16. van den Berg RA, Hoefsloot HCJ, Westerhuis JA, *et al.*: **Centering, scaling, and transformations: Improving the biological information content of metabolomics data**. *BMC Genomics.* 2006; **7**: 142. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
17. Eriksson L, Johansson E, Kettapeh-Wold S, *et al.*: **Introduction to multi- and megavariable data analysis using projection methods (PCA & PLS)**. *Umetrics*; 1999; 213–25.
18. Harrell FE: **Regression Modeling Strategies with Applications to Linear Models, Logistic and Ordinal Regression and Survival Analysis**. 2nd ed. New York: Springer. 2015. [Publisher Full Text](#)
19. Zou H, Hastie T: **Regularization and variable selection via the elastic net**. *J R Stat Soc Ser B Stat Methodol.* 2005. [Reference Source](#)
20. Hastie T, Tibshirani R, Friedman J: **The Elements of Statistical Learning: Data Mining, Inference and Prediction**. 2nd Ed. New York: Springer; 2008. [Reference Source](#)
21. Villar J, Ismail LC, Victora CG, *et al.*: **International standards for newborn weight, length, and head circumference by gestational age and sex: The Newborn Cross-Sectional Study of the INTERGROWTH-21<sup>st</sup> Project**. *Lancet.* 2014; **384**(9946): 857–68. [PubMed Abstract](#) | [Publisher Full Text](#)
22. Efron B, Hastie T: **Computer Age Statistical Inference**. *Computer Age Statistical Inference.* 2016. [Publisher Full Text](#)
23. Hawken S: **stevenhawken/Gates-Repository-China-Phil: prerelease 1, adding first few SAS macros (Version v0.1.alpha)**. *Zenodo.* 2020. <http://www.doi.org/10.5281/zenodo.4085320>
24. Lee AC, Mullany LC, Ladhani K, *et al.*: **Validity of Newborn Clinical Assessment to Determine Gestational Age in Bangladesh**. *Pediatrics.* 2016; **138**(1): e20153303. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
25. Bolstad BM, Irizarry RA, Astrand M, *et al.*: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics.* 2003; **19**(2): 185–93. [PubMed Abstract](#) | [Publisher Full Text](#)
26. Wu Y, Li L: **Sample normalization methods in quantitative metabolomics**. *J Chromatogr A.* 2016; **1430**: 80–95. [PubMed Abstract](#) | [Publisher Full Text](#)
27. Blencowe H, Cousens S, Oestergaard MZ, *et al.*: **National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: A systematic analysis and implications**. *Lancet.* 2012; **379**(9832): 2162–72. [PubMed Abstract](#) | [Publisher Full Text](#)
28. Lee AC, Panchal P, Folger L, *et al.*: **Diagnostic Accuracy of Neonatal Assessment for Gestational Age Determination: A Systematic Review**. *Pediatrics.* 2017; **140**(6): e20171423. [PubMed Abstract](#) | [Publisher Full Text](#)
29. Mundel T: **Innovation: How a 50-Year-Old Drop of Blood Helps Solve an Urgent Global Health Challenge | Impatient Optimists**. In: *Impatient Optimists*. 2017 [cited 10 Apr 2018]. [Reference Source](#)



## Open Peer Review

Current Peer Review Status: ? ✓ ✓ ✓

---

### Version 2

Reviewer Report 01 July 2021

<https://doi.org/10.21956/gatesopenres.14541.r30793>

© 2021 Courraud J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Julie Courraud**

Section for Clinical Mass Spectrometry, Statens Serum Institut, Copenhagen, Denmark

The authors have addressed my comments appropriately.

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Metabolomics, LC-MS/MS, clinical study design, clinical assays and quality

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 30 June 2021

<https://doi.org/10.21956/gatesopenres.14541.r30796>

© 2021 Villar J et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**José Villar**

Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK

**Eric Ohuma**

London School of Hygiene & Tropical Medicine, London, UK

Thank you to the authors for taking time to address the previous comments and subsequent changes to the manuscript. No further comments

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Medical Statistician with experience in the field of maternal and child health

**We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

---

**Version 1**

Reviewer Report 27 May 2021

<https://doi.org/10.21956/gatesopenres.14318.r30339>

© 2021 Agarwal R et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Ramesh Agarwal**

Newborn Health Knowledge Centre, WHO Collaborating Centre For Training and Research in Neonatal Care, Department of Pediatrics, All India Institute of Medical Sciences, New Delhi, Delhi, India

**Suman Chaurasia**

Department of Neonatology, All India Institute of Medical Sciences (AIIMS), Rishikesh, Uttarakhand, India

*Authors: Suman Chaurasia, Ramesh Agarwal*

- We are pleased to go through this interesting article by Hawken *et al.*, dealing with estimation of gestational age among neonates born in low-resource settings that illustrates novel clinical and metabolomic parameter-based models. We congratulate the authors for taking up this study on account of several outstanding features:
  - Conceptualizing the study idea of estimating gestational age postnatally, using a mix of conventional and novel metabolomic based objective parameters.
  - Enrolling neonates in large numbers from population based cohort to develop the model as well as validating internally.
  - Stratified random distribution of neonates among the derivation sub-cohorts to match the overall gestational age (GA) of the development cohort.
  - Using efficient study design of retrospective databases to source the samples - often stringently available in neonatal prospective cohort studies.
  - Utilizing machine-learning approaches to refine the algorithm.
  - Undertaking the enormous task of external validation rigorously - involving settings that may find the algorithm most useful, recruiting huge cohorts, harmonizing tools and processes across the sites, etc.
  - Finally, ensuring that the data is accessible to all interested in taking up future studies.

However, we have a few comments to make, especially from the clinical rather than public health viewpoints. Major comments are: the model's performance in term infants seems more reliable than preterms or SGAs. However the latter are the subgroups where gestation estimation maybe

much more useful in clinical settings. One of the reasons could have been because the derivation cohort itself had very few preterm infants (less than 5%) and so future studies should be planned for preterms and SGA categories. Also, as elaborated further, there appears to be much scope for conducting large multi-centric prospective study to remarkably improve and validate the GA algorithm, given that the article presents a promising alternative to grapple with this long-standing issue.

- We are of the opinion that GA estimation by the algorithm beyond  $\pm 1$  week may not be clinically as useful; so, the model performance in Table 4 in the main text should primarily depict this parameter and avoid parameters like “%  $\pm 14$  d”. The latter may actually be shifted to the web appendix if needed at all. This may also make the table less unsettling to the eyes.
- The authors admit that the model performance got attenuated from derivation to external validation such that the best performing model i. e. #3 lost its accuracy remarkably from average MAE of 0.71 (0.71, 0.72) in Ontario cohort to MAE of 0.90 (0.90, 0.91) and 0.89 (0.86, 0.91) in Manila and Shanghai cohorts respectively. The authors have rightly attributed this finding to factors like biological differences, sample collection or lab/equipment variations and as yet unknown data (pre-)processing techniques.
- However, as depicted in Figure 1 and accepted by the authors, *during external validation* among the different models, the full model (#3) only marginally improved the agreement over that already achieved by birthweight model or model 1 (average MAE 0.90 (Philippine)/0.89 (China) vs 0.96 (Canada)). This is *in contrast to internal validation*; addition of metabolic analytes (i.e. model #3; average MAE 0.71) or for that matter, restricting to analytes only model (model #2; 0.79) had significantly improved over the basic birthweight model (0.96). Two points emerge from this discussion:
  - Firstly, birthweight remains the most fundamental factor to predict GA with the lion’s share of explanatory attribute; this implies possible variability in measuring birthweight, quite plausible in resource-constrained settings and appears to be the crucial factor and should be minimized.
  - Secondly, the addition of analytes may not substantially improve beyond the birthweight’s robust contribution in the algorithm unless we consider critical remedies. One most likely pointer towards the solution may be attributed to the postnatal age cut off taken to collect the samples for metabolomic studies. The Ontario cohort’s sample collection cut off was 48 h compared to the other two cohorts of 72 h. Though the authors do mention that in the latter two cohorts, “most samples would have been excluded” with 48 h cut off, it may be pertinent to give the break up to describe the “most”.
  - Further it will be worthwhile to see how the model performs by removing the samples between 48 h and 72 h, and the same be included in the appendix.
  - If possible, analyzes involving samples more closer to birth e.g. within 24 h should also be alluded to give a broader understanding to the audience. Such exercises may aid towards improving the current performance of the model at nearly 75% of samples being predicted with GA  $\pm 1$  w. We do believe certainly that the latter target should be much higher- maybe close to 90%.
  - Thus, as raised earlier, we would like to emphasize that a prospective cohort study design for validation with the samples collected well within 48 h may lead to better algorithm development. In fact, we would suggest this esteemed group of authors

led by Hawken *et al.* to consider developing algorithm especially for the preterm infants recruiting subjects prospectively.

- Reference GA assessment for the Philippines cohort has been mentioned to have “generally received gestational dating ultrasounds” for infants born in private hospitals while “other infants GA were generally assessed using Ballard scoring”. This discrepancy probably explains why the Philippines restricted models (#3 or #2) did not perform well while validating with Manila cohort than with Ontario cohort. For example, for model # 3, the average MAE is higher for Manila cohort at 0.90 (0.90, 0.91) against Ontario cohort at 0.72 (0.71, 0.72) (Table 4). In contrast, the average MAE for China restricted model against Ontario cohort (0.76; 0.75, 0.76) is slightly closer to Shanghai cohort (0.89; 0.86, 0.91).
  - Therefore, firstly, we would suggest to provide the break up of two methods of reference GA ascertainment for the Philippines cohort.
  - Secondly, we would like the authors to consider reviewing the analyses of the Manila cohort excluding the infants with GA assessed by Ballard scoring. We assume the remaining cohort may still be large to validate the algorithm robustly given that it originally comprises of over 80,000 infants.
  - In addition, the need for the suggested review may also be relevant because of another example, of SGA cohorts. As highlighted by the authors, model 2 should better perform in the SGA cohorts. However, for Manila SGA10 cohort, model 2 of Philippines restricted model has higher MAE (**1.08**; 1.06, 1.09) compared to that of model 3 (**0.97**; 0.96, 0.99). This is contrary to the Canadian scenario: (**0.90**; 0.85, 0.94) vs. (**1.13**; 1.09, 1.17) or even the Chinese scenario: (**1.00**; 0.84, 1.15) vs. (**1.48**; 1.32, 1.64).
- It will also be worthwhile to have a view at the agreement plots (as in Figure 3) after removing the SGAs - SGA10, SGA3 and both in that order, especially in the Philippine cohort where SGAs constitute around 13% infants. This may also improve the average MAEs across the models 1 – 3. Additionally, separate agreement plots for SGAs should be explored as well, and considered to be included in the appendix if they make sense. This may particularly helpful for several LMICs regions having high prevalence of IUGR like South-east Asian Region.
- We have noted a typo error: the proportion of infants predicted by Philippines restricted model for the Ontario cohort within +/- 2 weeks is mentioned as 69.9; this should perhaps be 96.9, going by the 95% CIs.

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

I cannot comment. A qualified statistician is required.

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Neonatology, Neonatal Sepsis, Infectious disease epidemiology

**We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 26 April 2021

<https://doi.org/10.21956/gatesopenres.14318.r30526>

© 2021 Courraud J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Julie Courraud**

Section for Clinical Mass Spectrometry, Statens Serum Institut, Copenhagen, Denmark

The authors report the results of a validation study of previously developed algorithms to predict gestational age in post-natal settings. This is a complex task considering the many differences between the countries and the authors have gathered substantial datasets for this study. One model is based on sex and birthweight only. The two other models include metabolic profiles measured on dried blood spot samples during newborn screening for congenital disorders. The models were applied on data acquired in different laboratories in China and the Philippines in comparison with data from a Canadian laboratory where it was developed. The ultimate goal is legitimate and, although the results are promising, given the limitations observed for preterm infants, several aspects must be addressed and explored before the method can be used in the clinic.

I could not assess the relevance or correct application of the deep learning method used, as well as bootstrap percentile confidence intervals, as these are not within my area of expertise.

### **Comments:**

#### **Introduction:**

- In paragraph 2, you state that "Samples are typically collected within the first few days after birth, but under special circumstances (e.g., preterm birth, neonatal transfer) may be collected later". In your study, you selected samples collected within 48 hours only, and explain that the reason is that LMIC usually collect the samples in this timeframe. However,



you also state that “most samples would have been excluded if the >48-hour exclusion were applied to these validation cohorts” (Methods, paragraph 2), so it seems that many samples were collected between 48 and 72 hours as mentioned in the discussion.

- Please discuss whether your algorithm could then be used to target the right population, i.e. preterm birth, when the samples might not be collected within 48 or even 72 hours.
- In untargeted metabolomics of newborn dried blood spots, it has been shown that the baby's age at sampling is a critical variable when one considers metabolic profiles, and only a few days difference has significant impact. Have you investigated the extent of impact of this variable on your targeted metabolic profiling? How do you intend to address this in your future research or when applying your algorithm? Have you considered integrating age at sampling as a variable in the algorithm (or as a stratification variable for the partitioning into subsets)? See also my comment regarding the discussion.
- Please discuss the limitation of applying the algorithm **outside** the age at sampling range on which it was developed (or mention/rephrase this limitation more specifically in the discussion: “one day later after birth *than the samples used for model development*” as no sample >48 hours was included during model development).
- In untargeted metabolomics of newborn dried blood spots, another crucial covariate impacting metabolic profiling is month of birth (see Courraud *et al.* 2021), or so at least in Denmark. Being born in summer or winter is remarkably visible. Such effect might be or not be visible in various countries. Have you investigated this potential covariate in your targeted profiling and/or considered integrating it in the algorithm?

### Methods:

- Paragraphs 3-5. Please specify which analytical methods are used in each center included in the study. Is it mass spectrometry everywhere? Do they use a marketed kit or laboratory-developed tests? Consider giving more methodological details as supplementary material as different platforms may not give the same analytical performance.
- Paragraph 4. Please clarify why some infants get the expanded screening panel of 28 diseases and discuss the risk of selection bias when choosing these infants for the validation.
- Paragraph 5. Please discuss the risk of selection bias when choosing to include only infants for whom tandem mass spectrometry data were available. Does this mean that all metabolites have been measured with this method? (Is this the method used to screen for phenylketonuria, congenital adrenal hyperplasia, hypothyroidism and Glucose-6-phosphate dehydrogenase deficiency?).
- If the applicability of the algorithm is dependent on the family's income (being able to pay for extra screening), will it achieve its goal to reflect preterm birth globally, given that preterm birth is more frequent in families struggling economically? Please discuss.
- Paragraph 6 on GA assessment: for the Philippines, please indicate the proportion of infants

for whom GA was assessed by ultrasound or using Ballard Scoring. A note on the precision of the Ballard scoring with a relevant reference would help the reader.

- On the same topic, you later state that “model performance was assessed by comparing the estimated GA from the model to the ultrasound-derived GA”. So it is unclear whether or not the infants for whom GA was assessed using Ballard Scoring are included at all. Please clarify.

**Table 1:**

- Please indicate what “C0”, “C2”, etc. refer to precisely. It might be obvious for someone in the field, but not to many readers for whom C18 might just be a free fatty acid and not the acyl-carnitine. You could for instance provide a list in supplementary data with full names and PubChemIDs. It helps bridging with the untargeted metabolomics community who is also working on the topic.
- Please be more specific as to which metabolites are included in model 2 and 3. It’s not clear, especially considering the “restricted models” in Table 4.
- Models including newborn screening analytes: How did you cope with the metabolites missing in the validation cohorts? In the result section, you mention “Philippines-restricted” models, etc., please introduce them in the method section. Are the equations the same, just removing the missing metabolites or did you “re-develop” the models? Or?
- Are your models “resistant” to missing values? (In the real world, there will be missing values.)
- Would it be possible to report which metabolites have the biggest influence in each model?
- Have you considered a “model 4” restricted to the few metabolites measured for the “basic” screening panels offered in China and the Philippines? It would be accessible to more people as far as I understand, and might still perform better than just birthweight and sex.
- Statistical modeling: while MAE is clearly explained, it is unclear how RMSE is calculated and what it brings. An extra sentence would be welcome, for instance with an example as given for MAE. RMSE values in Table 4 are not discussed in the manuscript, so if this metric does not bring important elements to understand the work, consider giving the values in supplementary data. Else, please discuss this metric.

**Results and discussion:**

- Can you comment on the high percentage of SGA in the Filipino cohort? Could it be that the thresholds used (ref 14) are not applicable to this population? Could it also be why models generally perform better in the Filipino cohort for the SGA infants as compared to Canadian and Chinese cohorts? (More power).
- In relation to my comment above (introduction), you mention that a majority of Chinese samples have been collected between 48-72 hours. Without going into extensive details, could you present your hypotheses as to why this variable matters? (Are there special metabolic changes during this window for instance?).

## **Minor comments:**

### **Introduction and methods:**

- Paragraph 3. "in cohorts of infants from in North American settings". Please remove the "from".
- Data cleaning and normalization: Please develop the LMIC abbreviation.
- Statistical modeling: "a reported MAE of 1.0 weeks". Please write "week" for values below 2.0 weeks throughout the manuscript (several places).

### **Results:**

- Table 3:
  - Females in Canadian cohort, please correct the percentage to 50.7%.
  - Birthweight values. I do not think that the decimal makes sense considering the precision of the measurements. I doubt that the used equipment goes below 1 g of precision. I would present birthweight with no decimal.
  - Please use only one decimal for percentage of SGA counts.
- Please add thousands separators throughout the manuscript in a homogeneous way.
- Internal validation: The second sentence is 61 words long. Please split it in 2-3 sentences for clarity.
- External validation in the Philippines cohort: please indicate the CI for model 1 and 2 regarding the estimated preterm birth rate.
- Same comment for estimated preterm birth rate in the Chinese cohort using model 2.
- Figure 1.
  - (A) It would be more informative to describe models as follows: Model 1: sex + birth weight; Model 2: sex + analytes; Model 3: sex + birth weight + analytes. "analyte model" and "full model" are not very clear.
  - (C) redundant x axis legend.

### **Discussion:**

- You write: "First, as observed in the differences in performance across the birth weight-only models developed in the three cohorts, the predictive utility of anthropomorphic measurements for estimating GA may vary across populations".
  - "birth weight-only models developed " Do you mean the unique model 1 (sex + birth weight) applied in the 3 cohorts? This sentence is confusing as it implies that there are several models that were developed, when I had understood that you **developed** one model 1 based on the Canadian infants and **applied** "the final equations" to the

other cohorts no involved in the development. Please clarify.

- Also, why do you think that the “predictive **utility** of anthropomorphic measurements for estimating GA may vary across populations”? It could be that anthropomorphic measurements are indeed too different between Canada and Asian populations, so the models developed with Canadian data are not performing in Chinese infants. But why question the utility of the measurement itself? (To make a comparison with, for instance, month of birth, one could argue that seasonal variation is relevant in some climates but not in others. I’m not sure why birthweight would be more or less relevant and I’m just curious as to whether you have a more specific hypothesis.)
- Sentence starting with “Previous validation of our models among”: Please split this sentence as it is too long and difficult to know what you are referring to when you end with “differences were more pronounced” (between what? These different subgroups? More pronounced compared to?). When you write “inherent biological differences”, do you mean both genetic and environmental? Please clarify.

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Partly

**If applicable, is the statistical analysis and its interpretation appropriate?**

I cannot comment. A qualified statistician is required.

**Are all the source data underlying the results available to ensure full reproducibility?**

No

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Metabolomics, LC-MS/MS, clinical study design, clinical assays and quality

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 27 May 2021

**Kumanan Wilson**, Ottawa Hospital Research Institute, Ottawa, Canada

**Introduction:**

1. In paragraph 2, you state that “Samples are typically collected within the first few days after birth, but under special circumstances (e.g., preterm birth, neonatal transfer) may be collected later”. In your study, you selected samples collected within 48 hours only, and explain that the reason is that LMIC usually collect the samples in this timeframe. However, you also state that “most samples would have been excluded if the >48-hour exclusion were applied to these validation cohorts” (Methods, paragraph 2), so it seems that many samples were collected between 48 and 72 hours as mentioned in the discussion.

**Response:** *Thank you for your careful review of the manuscript. We have removed the second half of the sentence as we now see that it could confuse reviewers.*

*We have now provided more details on the exclusion criteria applied to the Ontario cohort for model development. The two criteria leading to the most exclusions were 1) requiring gold-standard GA measurement via 1st-trimester dating ultrasound, and 2) screening bloodspot collection within 48 hours of birth.*

*The first exclusion criteria may have excluded infants born in rural or underserved areas of the province where access to comprehensive prenatal care was lower. In many cases however, this is more likely to be a data quality issue, where dating ultrasound was used but not recorded as such. The second criteria led to the disproportionate exclusion of preterm infants who more often had delayed sample collection, despite this not being recommended practice. Although this exclusion biased the rate of preterm gestation observed in our Ontario study cohort downward, but it was unlikely to have had any important impact on GA model development, as we still had a large sample size across the full spectrum of gestational ages at birth to allow robust model development and performance evaluation. Further, the inclusion of samples collected later than 48 hours would introduce a large amount of heterogeneity in analyte levels which had to be balanced against the impact of exclusions. Although our intention was to exclude samples collected later than 48 hours for the external cohorts as well, it was not possible to take this approach because only calendar day of sample collection was available. Since hourly data was not available, a 48-hour cut off would have excluded most samples. Therefore, we relaxed the exclusion criteria to >72 hours. We have reorganized the methods and provided additional details which will clarify some of these points.*

1. Please discuss whether your algorithm could then be used to target the right population, i.e. preterm birth, when the samples might not be collected within 48 or even 72 hours.

**Response:** *We have reviewed data collected outside of the recommended time frame. In our first external validation study in Bangladesh (Murphy et al, eLife 2019), mean sample collection time was ~14 hours as mothers were often discharged before the 24-hour time frame. If samples are collected too early, hemoglobin values still reflect maternal values and decreases the accuracy of the algorithm.*

*Ultimately for the algorithm to be viable, it needs to be effective across a range of sample collection timings to accommodate for early or late collection times.*

*For accuracy of newborn screening, it is recommended that samples are collected between*



***24-48 hours as heterogeneity begins to appear after 48 hours. Limiting the collection time for algorithm development reduces the variability in the data and improves the accuracy of the algorithm.***

2. In untargeted metabolomics of newborn dried blood spots, it has been shown that the baby's age at sampling is a critical variable when one considers metabolic profiles, and only a few days difference has significant impact. Have you investigated the extent of impact of this variable on your targeted metabolic profiling? How do you intend to address this in your future research or when applying your algorithm? Have you considered integrating age at sampling as a variable in the algorithm (or as a stratification variable for the partitioning into subsets)? See also my comment regarding the discussion.

***Response: This is an important consideration. We did incorporate time of sample collection in earlier exploratory models, and though a significant term retained in the model, the effect of time at collection appeared to mostly be the addition of noise/heterogeneity rather than having a monotonic relationship with gestational age that improved model estimates.***

3. Please discuss the limitation of applying the algorithm **outside** the age at sampling range on which it was developed (or mention/rephrase this limitation more specifically in the discussion: "one day later after birth *than the samples used for model development*" as no sample >48 hours was included during model development).

***Response: We have made the suggested edit in the discussion as follows:***

"Another source of bias was that samples in infants born preterm in China, and to a lesser extent the Philippines, were much more likely to be collected later than in term infants, and often after 48-72 hours. Samples collected beyond 48 hours were more heterogeneous than samples collected within 48 hours but excluding these would have excluded an unacceptably large proportion of infants overall, especially preterm infants. Our compromise approach of relaxing the criteria to exclude samples that were collected later than 72 hours after birth, included more infants at the cost of increased heterogeneity, but even so, still excluded a disproportionate number of preterm infants in the China and Philippines. A combination of these factors likely contributed to the lower than expected preterm birth rates observed in China and in the Philippines validation cohorts, as well as leading to decreased apparent model performance due to more heterogeneous samples (collected 48-72 hours after birth) being included."

4. In untargeted metabolomics of newborn dried blood spots, another crucial covariate impacting metabolic profiling is month of birth (see Courraud *et al.* 2021), or so at least in Denmark. Being born in summer or winter is remarkably visible. Such effect might be or not be visible in various countries. Have you investigated this potential covariate in your targeted profiling and/or considered integrating it in the algorithm?

***Response: Thank you for this comment. We are aware of studies demonstrating seasonable***

*variability in newborn screening outcomes (i.e. Ryckman et al., 2013), and also based on discussions with subject matter experts within our Ontario newborn screening, however the effects noted in Ontario data were small. Despite this, we have actually accounted for seasonality in our data normalization process, in which analyte values were standardized within monthly strata over time, to address assay/reagent/standard changes over time, which had the additional benefit of removing any seasonal bias in analyte measurements within the Ontario, China and Philippines screening data.*

**Reference:**

*Ryckman KK, Berberich SL, Shchelochkov OA, Cook DE, Murray JC. Clinical and environmental influences on metabolic biomarkers collected for newborn screening. Clin Biochem. 2013 Jan;46(1-2):133-8. doi: 10.1016/j.clinbiochem.2012.09.013. Epub 2012 Sep 23. PMID: 23010448; PMCID: PMC3534803.*

**Methods:**

1. Paragraphs 3-5. Please specify which analytical methods are used in each center included in the study. Is it mass spectrometry everywhere? Do they use a marketed kit or laboratory-developed tests? Consider giving more methodological details as supplementary material as different platforms may not give the same analytical performance.

**Response:** *In Ontario, hemoglobin profiles were determined by high performance liquid chromatography; neonatal 17-OHP, and TSH were measured using a PerkinElmer AutoDELFIA® Immunoassays; amino acid and acylcarnitine analysis was performed by tandem mass spectrometry; total TREC copy number was measured by quantitative polymerase chain reaction using a ThermoFisher Scientific Viia 7; biotinidase and galactose-1-phosphate uridylyltransferase levels were measured using the Astoria-Pacific SPOTCHECK® Pro system.*

*In the Philippines, a commercial kit is used for the measurement of amino acids, succinylacetone, free carnitine, acylcarnitines and PKU by tandem mass spectrometry (NeoBase 1, Perkin Elmer). 17OHP and TSH to detect CAH and CH, respectively are detected using a Autodelfia kit by fluoroimmunoassay (Perkin Elmer). Screening for G6PDH is done fluorometrically.*

*We were unfortunately not able to confirm the details of analytical methods used in China. Based on published works (Shi et al., 2012), we have surmised the following: phenylalanine was measured fluorometrically to detect phenylketonuria. In congenital hypothyroidism, TSH is quantified by radioimmunoassay (RIA), enzyme linked immunosorbent assay (ELISA) or dissociation-enhanced lanthanide fluoroimmunoassay (DELFIA). The remaining tests for the expanded panel use tandem mass spectrometry.*

**Reference:**

*Shi XT, Cai J, Wang YY, Tu WJ, Wang WP, Gong LM et al. Newborn screening for inborn errors of metabolism in mainland china: 30 years of experience. JIMD Rep. 2012;6:79-83. doi: 10.1007/8904\_2011\_119. Epub 2012 Jan 31. PMID: 23430943; PMCID: PMC3565663.*

2. Paragraph 4. Please clarify why some infants get the expanded screening panel of 28

diseases and discuss the risk of selection bias when choosing these infants for the validation.

**RESPONSE:** *Although the newborn screening initiatives are meant to be universal in these populations, some tests were paid for by the families. The expanded panel is now covered by National Health Insurance in the Philippines. As our study only included samples for which the full panel of analytes was available, this could have contributed to a selection bias towards more affluent families. We have added the following to the discussion:*

"The preterm birth rate that we estimated in the current cohort (Philippines:4.6%, China: 4.8%) was less than previously estimated (Philippines: 15%, China: 7.1%) (Blencowe et al., 2012). Although newborn screening initiatives are meant to be universal in these populations, some tests are paid for by the families. Considering we only tested samples which the full panel was available this could have contributed to selection bias in our sample population towards more affluent families."

Reference:

Blencowe H, Cousens S, Oestergaard M, Chou D, et al.: National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *The Lancet*. 2012; **379** (9832): 2162-2172.

3.Paragraph 5. Please discuss the risk of selection bias when choosing to include only infants for whom tandem mass spectrometry data were available. Does this mean that all metabolites have been measured with this method? (Is this the method used to screen for phenylketonuria, congenital adrenal hyperplasia, hypothyroidism and Glucose-6-phosphate dehydrogenase deficiency?).

**Response:** *The metabolites listed for each site with the expanded panel are listed in Table 1 and in Figure 1. All metabolites listed were measured at each external site. In Shanghai, tests for phenylketonuria, congenital adrenal hyperplasia, hypothyroidism and Glucose-6-phosphate dehydrogenase deficiency are not measured by tandem mass spectrometry and are publicly funded. The expanded panel of tandem mass spectrometry are strongly recommended but participation is voluntary in both external settings at the time of analysis. There is a risk for selection bias toward a more affluent and medically aware population and we have included this in the manuscript.*

Reference:

Shi XT, Cai J, Wang YY, Tu WJ, Wang WP, Gong LM et al. Newborn screening for inborn errors of metabolism in mainland china: 30 years of experience. *JIMD Rep*. 2012;6:79-83. doi: 10.1007/8904\_2011\_119. Epub 2012 Jan 31. PMID: 23430943; PMCID: PMC3565663.

4.If the applicability of the algorithm is dependent on the family's income (being able to pay for extra screening), will it achieve its goal to reflect preterm birth globally, given that preterm birth is more frequent in families struggling economically? Please discuss.

**Response:** *Universal access and coverage for newborn screening is improving around the world. Since the initiation of this project, newborn screening is now fully funded in the Philippines. The authors agree that if the algorithm is dependent on a family's ability to pay for screening, this could bias preterm birth rates and is not an ideal scenario for implementation of this approach. However, the approach we are evaluating isn't intended*

**for implementation in either the Philippines or China - this was an opportunistic use of external cohorts with large retrospective screening databases where we could test our models in different geographic settings with results from local screening labs. Our approach is ultimately targeted to LMICs in Africa and South Asia, based on priorities defined by the Gates foundation assuming the necessary panel of analytes would be collected for the purpose of GA estimation.**

5. Paragraph 6 on GA assessment: for the Philippines, please indicate the proportion of infants for whom GA was assessed by ultrasound or using Ballard Scoring. A note on the precision of the Ballard scoring with a relevant reference would help the reader.

**Response: Although we knew general practice patterns of gestational dating method used in the Philippines, we did not have individual-level data on what method was used in each individual pregnancy, hence we accepted this as an additional source of validation error which would likely have led to larger MAE/RMSE. This was presented as a limitation in the discussion and we have clarified the description in the methods. It now reads:**

“In the Philippines cohort, mothers who delivered in private hospitals generally received gestational dating ultrasounds while other infants’ GAs were generally measured using Ballard Scoring, however individual-level data identifying which GA measurement method was used was not available.”

**We also added a statement in the discussion about the precision of the Ballard scoring:**

“Ballard tends to overestimate gestational age with a wide margin of error, particularly in preterm infants (Lee et al., 2016).”

Reference:

Lee AC, Mullany LC, Ladhani K, Uddin J, Mitra D, Ahmed P et al. Validity of Newborn Clinical Assessment to Determine Gestational Age in Bangladesh. *Pediatrics*. 2016 Jul;138(1):e20153303. doi: 10.1542/peds.2015-3303. Epub 2016 Jun 16. PMID: 27313070; PMCID: PMC4925072.

6. On the same topic, you later state that “model performance was assessed by comparing the estimated GA from the model to the ultrasound-derived GA”. So it is unclear whether or not the infants for whom GA was assessed using Ballard Scoring are included at all. Please clarify.

**Response: Thank you for pointing this out. We added a qualifying statement to correct this in the methods. The sentence now reads as follows.**

“For each infant, model performance was assessed by comparing the estimated GA from the model to the ultrasound-derived GA (or ultrasound or Ballard in the Philippines) and calculating validation metrics that reflect the precision of model estimates compared to reference GA values.”

**Table 1:**

7. Please indicate what "C0", "C2", etc. refer to precisely. It might be obvious for someone in the field, but not to many readers for whom C18 might just be a free fatty acid and not the acyl-carnitine. You could for instance provide a list in supplementary data with full names and PubChemIDs. It helps bridging with the untargeted metabolomics community who is also working on the topic.

**Response:** *As this journal does not allow supplementary materials, the authors feel that providing the full names of all species would be more detail than is needed. We have added a reference to a previous paper which defines the list more exhaustively. We've also added subheadings which identify the 'C' species as acylcarnitines.*

8. Please be more specific as to which metabolites are included in model 2 and 3. It's not clear, especially considering the "restricted models" in Table 4

**Response:** *We have added in a new figure. (Figure 1) to make it more clear which analytes are used in the restricted models.*

9. Models including newborn screening analytes: How did you cope with the metabolites missing in the validation cohorts? In the result section, you mention "Philippines-restricted" models, etc., please introduce them in the method section. Are the equations the same, just removing the missing metabolites or did you "re-develop" the models? Or?

**Response:** *The internal validation of models 1-3 was conducted on an independent test dataset of Ontario infants. Since not all of the predictors available for the Ontario dataset were also available for the external datasets, we tailored the models 2 and 3 to include the maximum number of available predictors in each of the external datasets (which we called 'restricted models'). The list of analytes available at each site is presented in Table 1. The tailored models were fit in the Ontario dataset, and validated in the external datasets. We have added a figure to clarify the different models tested. The Methods section now states:*

*"A total of 47 newborn screening analytes, as well as sex, birth weight and multiple birth status, were used in the original Ontario model development. GA at birth (in weeks) determined by first trimester gestational dating ultrasound was the dependent variable. Multiple birth status and a subset of screening analytes were not available in the external cohorts (Table 1). Three main models were developed and evaluated in the Ontario cohort (Table 2). For models 2 and 3, we also developed restricted models including only the covariates available in each of the two external cohorts (Figure 1). Restricted models were trained on the Ontario datasets but deployed in the external cohorts."*

10. Are your models "resistant" to missing values? (In the real world, there will be missing values.)

**Response:** *In the current study the sample size was large enough to exclude any samples with missing values and thus no imputation of missing values was done. In our external*



*validation studies (Murphy et al., 2019, Bota AB et al., 2020 and Hawken et al., 2021) where sample size was limited, we used multiple imputation for missing analyte values. If these models were to be implemented in real-world settings, we would use the same methods to impute missing values.*

**References:**

**Murphy M, Hawken S, Cheng W, Wilson L, Lamoureux M, Henderson M et al. Postnatal gestational age estimation using newborn metabolic profiles: A validation study in Matlab, Bangladesh. *Elife*. 2019 Mar 19;8. pii: e42627. doi: 10.7554/eLife.42627.**

**Bota AB, Ward V, Hawken S, Wilson LA, Lamoureux M, Ducharme R, et al. Metabolic gestational age assessment in low resource settings: a validation protocol. *Gates Open Res*. 2021 Jan 28;4:150. doi: 10.12688/gatesopenres.13155.2. PMID: 33501414; PMCID: PMC7801859.**

**Hawken S, Ducharme R, Murphy MSQ, Olibris B, Bota AB, Wilson LA, et al. Development and external validation of machine learning algorithms for postnatal gestational age estimation using clinical data and metabolomic markers. 2021. *BMC Pregnancy and Childbirth* (under review).**

11. Would it be possible to report which metabolites have the biggest influence in each model?

**Response: Birth weight was the strongest predictor of gestational age overall. Seven analytes were the strongest predictors. Based on partial Spearman correlation analysis there were seven analyte covariates that had distinctly stronger partial correlations with GA compared to all others. These seven analytes in order of strength of partial spearman correlation are: fetal-to-adult hemoglobin ratio, 17-OHP, C4DC, TYR, ALA, C5, and C5DC. These details have been added to the methods.**

12. Have you considered a “model 4” restricted to the few metabolites measured for the “basic” screening panels offered in China and the Philippines? It would be accessible to more people as far as I understand, and might still perform better than just birthweight and sex.

**Response: We have previously derived and reported on restricted model results that included 17-hydroxyprogesterone(17OHP), thyroid stimulating hormone (TSH), and fetal/adults hemoglobin, as these are highly predictive analytes that can be cost-effectively analyzed using non MS/MS marketed tests. Although these models were predictive, they were far inferior in performance to models evaluating the full set of expanded screening analytes and didn't reach a promising level of precision in estimating GA, therefore we haven't pursued them further.**

13. Statistical modeling: while MAE is clearly explained, it is unclear how RMSE is calculated and what it brings. An extra sentence would be welcome, for instance with an example as given for MAE. RMSE values in Table 4 are not discussed in the manuscript, so if this metric does not bring important elements to understand the work, consider giving the values in supplementary data. Else, please discuss this metric.

**Response: Thank you. We have elaborated on the difference between RMSE and MAE in the methods section, and have now provided formulas (please see manuscript text for formulas):**

“Model accuracy metrics were based on residual errors: the difference between model-estimated GA and reference GA. Although mean square error (MSE) is typically the loss function used in maximum-likelihood model fitting for continuous outcomes, it is not necessarily the best metric for assessing average agreement in model validation, as it is based on sum of squared differences, and hence is sensitive to large and small residuals. Therefore, the primary metric we have presented is the mean absolute error (MAE). MAE is the average of absolute values of residuals (values of the model estimate minus the reference GA) across all observations. MAE reflects the average deviation of the model estimate compared to the reference estimate, expressed in the same units as GA (weeks).

For completeness, as well as for comparability to other published validations, we also report the square root of the MSE (RMSE). Also known as the standard error of estimation, RMSE is also expressed in the same units as GA (weeks).

Lower values of both MAE and RMSE reflects more accurate model estimated GA.”

#### **Results and discussion:**

1. Can you comment on the high percentage of SGA in the Filipino cohort? Could it be that the thresholds used (ref 14) are not applicable to this population? Could it also be why models generally perform better in the Filipino cohort for the SGA infants as compared to Canadian and Chinese cohorts? (More power).

**Response: SGA was previously estimated to be 20.9% in an urban cohort of Filipino infants in the Cebu Longitudinal Health and Nutrition Survey (Jones et al 2008) and 22.5% in a smaller cohort (Blake et al 2016). The INTERGROWTH-21 SGA standards population did not include a Filipino cohort in its population, thus it is possible that this is not an acceptable reference for this particular population. Filipino infants born in Ontario have previously been shown to be at higher risk of being born small as compared to other Asian infants (Batsch et al).**

#### **References:**

**Jones LL, Griffiths PL, Adair LS, Norris SA, Richter LM, Cameron N. A comparison of the socio-economic determinants of growth retardation in South African and Filipino infants. *Public Health Nutr.* 2008 Dec;11(12):1220-8. doi: 10.1017/S1368980008002498. Epub 2008 May 8. PMID: 18462561; PMCID: PMC2939971.**  
**Blake RA, Park S, Baltazar P, Ayaso EB, Monterde DB, Acosta LP, Olveda RM, Tallo V, Friedman JF. LBW and SGA Impact Longitudinal Growth and Nutritional Status of Filipino Infants. *PLoS One.* 2016 Jul 21;11(7):e0159461. doi: 10.1371/journal.pone.0159461. PMID: 27441564; PMCID: PMC4956033.**

2. In relation to my comment above (introduction), you mention that a majority of Chinese samples have been collected between 48-72 hours. Without going into extensive details,

could you present your hypotheses as to why this variable matters? (Are there special metabolic changes during this window for instance?).

**Response:** *For accuracy of newborn screening, it is recommended that samples are collected between 24-48 hours as samples collected beyond that window have increasingly heterogeneous analyte results influenced by multiple exogenous factors that cannot be statistically adjusted for. Limiting the collection time for algorithm development reduces the variability in the data and improves the accuracy of the algorithm.*

#### **Minor comments:**

##### **Introduction and methods:**

1. Paragraph 3. "in cohorts of infants from in North American settings". Please remove the "from".

**Response:** *Thank you, this has been corrected.*

2. Data cleaning and normalization: Please develop the LMIC abbreviation.

**Response:** *Thank you, we have made this edit.*

3. Statistical modeling: "a reported MAE of 1.0 weeks". Please write "week" for values below 2.0 weeks throughout the manuscript (several places).

**Response:** *We have edited the manuscript to remove the s after any mention of 1 week. However, the authors think that any decimal number should be presented in plural (ie. 0.96 weeks). We defer to the editors recommendations for guidance on this suggestion.*

##### **Results:**

4. Table 3: Females in Canadian cohort, please correct the percentage to 50.7%.

**Response:** *Thank you, we have corrected this.*

5. Birthweight values. I do not think that the decimal makes sense considering the precision of the measurements. I doubt that the used equipment goes below 1 g of precision. I would present birthweight with no decimal.

**Response:** *Thank you, we have corrected this.*

6. Please use only one decimal for percentage of SGA counts.

**Response:** *We have made this edit in Table 3, thank you.*

7. Please add thousands separators throughout the manuscript in a homogeneous way.

**Response: We have corrected this in the manuscript, thank you.**

8. Internal validation: The second sentence is 61 words long. Please split it in 2-3 sentences for clarity.

**Response: We have edited this sentence.**

9. External validation in the Philippines cohort: please indicate the CI for model 1 and 2 regarding the estimated preterm birth rate.

**Response: We have corrected this in the manuscript, thank you.**

10. Same comment for estimated preterm birth rate in the Chinese cohort using model 2.

**Response: We have corrected this in the manuscript, thank you.**

11. Figure 1

1.
  - (A) It would be more informative to describe models as follows: Model 1: sex + birth weight; Model 2: sex + analytes; Model 3: sex + birth weight + analytes. "analyte model" and "full model" are not very clear.
  - (C) redundant x axis legend.

**Response: We have edited the figure and figure legend as suggested.**

#### Discussion:

1. You write: "First, as observed in the differences in performance across the birth weight-only models developed in the three cohorts, the predictive utility of anthropomorphic measurements for estimating GA may vary across populations".
  - "birth weight-only models developed " Do you mean the unique model 1 (sex + birth weight) applied in the 3 cohorts? This sentence is confusing as it implies that there are several models that were developed, when I had understood that you **developed** one model 1 based on the Canadian infants and **applied** "the final equations" to the other cohorts no involved in the development. Please clarify.

**Response: We developed 3 models in the Ontario cohort which were applied to the external cohorts. Model 1 was a multivariable regression model including sex, birthweight and their interaction. Model 2 included ELASTIC NET regression model including sex, analytes and pairwise interactions among predictors, whereas model 3 used ELASTIC NET regression model including sex, birth weight, analytes and pairwise interactions among predictors. These were then applied to the external cohorts using analytes available in these settings. To clarify this, we have replaced all references to "birth weight only models" with referenced to Model 1 throughout the manuscript.**

2. Also, why do you think that the “predictive **utility** of anthropomorphic measurements for estimating GA may vary across populations”? It could be that anthropomorphic measurements are indeed too different between Canada and Asian populations, so the models developed with Canadian data are not performing in Chinese infants. But why question the utility of the measurement itself? (To make a comparison with, for instance, month of birth, one could argue that seasonal variation is relevant in some climates but not in others. I’m not sure why birthweight would be more or less relevant and I’m just curious as to whether you have a more specific hypothesis.)

**RESPONSE:** *This is an important consideration. As our intent was to externally validate models derived in Ontario infants, we applied Model 1 based on sex and birthweight model coefficients derived in Ontario infants. Although birth weight was locally standardized within each of the three cohorts, it is possible that deriving local models in China and the Philippines and then applying them locally would yield better performance. And we did see evidence of this in previous exploratory modeling we have done. However our intent here was to deploy pre-trained models to use in estimating GA that don’t require an existing database in each new country that is large enough to derive a robust country-specific model.*

3. Sentence starting with “Previous validation of our models among”: Please split this sentence as it is too long and difficult to know what you are referring to when you end with “differences were more pronounced” (between what? These different subgroups? More pronounced compared to?). When you write “inherent biological differences”, do you mean both genetic and environmental? Please clarify.

**Response:** *Thank you we have edited the sentence and it now reads:*

“Previous validation of our models among infants born in Ontario to landed-immigrant mothers from eight different countries across Asia and North and Sub-Saharan Africa suggested that inherent biological differences may not be a significant contributor to newborn metabolic data and the performance of our algorithms (Hawken et al., 2017). However in an external validation of previously developed GA estimation models in a prospective cohort from South Asia (Murphy et al., 2019), the drop in performance was more pronounced, despite the centralized analysis of samples in the Ontario Newborn Screening lab.”

#### References:

Hawken S, Ducharme R, Murphy MSQ, et al.: Performance of a postnatal metabolic gestational age algorithm: a retrospective validation study among ethnic subgroups in Canada. *BMJ Open*. 2017;7(9):e015615. 28871012 10.1136/bmjopen-2016-015615 5589017  
Murphy MSQ, Hawken S, Cheng W, et al.: External validation of postnatal gestational age estimation using newborn metabolic profiles in Matlab, Bangladesh. *eLife*. 2019;8:e42627. 30887951 10.7554/eLife.42627 6424558

**Competing Interests:** No competing interests were disclosed.

Reviewer Report 29 March 2021

<https://doi.org/10.21956/gatesopenres.14318.r30282>

© 2021 Villar J et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**José Villar**

Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK

**Eric Ohuma**

London School of Hygiene & Tropical Medicine, London, UK

In this manuscript by Hawken *et al.*, the authors have performed an external validation of newborn metabolomic markers for postnatal GA estimation in East and South-East Asian infants using an Elastic Net regression modelling approach.

#### Main comments:

- There is no doubt that an accurate estimate of GA is key. However, the authors propose a postnatal estimation of GA which does very little in advancing and encouraging determination of GA in early pregnancy. This is the recommendation of the WHO, The Brighton Collaboration GAIA definitions Prematurity and assessment of gestational age, The National Institute for Health and Care Excellence (NICE) Guideline for Routine Antenatal Care (2008), and International Society of Ultrasound in Obstetrics and Gynaecology (ISUOG).
- There is a clear gap identified in the accuracy of determining GA especially in LMIC and I strongly doubt this approach will help in better characterising of the burden of vulnerable newborns and the potential impact towards achieving better estimates for population rates of preterm birth, low birth weight, small-for-gestational age, and combinations of these to identify other vulnerable newborn phenotypes.
- Overall reporting of results warrants improvement, the authors have decided to report overall mean agreement in GA between gold standard GA with model predicted GA and yet there are clearly large differences in precision and this differs according to GA (Figure 1). The authors should state explicitly what is meant by agreement within 7 days. For example, if on average, model estimate agrees within 7 days, this technically means  $\pm 7$  days and therefore for a given fetus say model GA estimate is 32 weeks + 0 days, this would mean that the true GA ranges between 31 weeks + 0 days and 33 weeks + 0 days which is effectively 2 weeks. Following this, the best model estimate (model 3) on average will be accurate to within 10 days at best. The authors should show a plot of true GA vs. predicted GA as this will evidently show the variability of the prediction as a function of GA as opposed to the aggregated estimates they have presented by GA in Figure 1.
- Across all models, great discrepancies and perhaps unacceptable discrepancies are observed for GA before 39 weeks. I am not convinced this approach offers any benefit/added value/utility compared to other methods in common use such as best obstetric methods for ascertaining GA.
- The team used blood spot samples collected within 48hrs of delivery – there is considerable



extra effort involved, time, and cost for drawing blood spots and processing of analytes. I do not see how this would be a feasible alternative especially for LMIC where accurate estimation of GA is a key data gap. The merits of the proposed approach have to clearly outweigh the performance of other known methods for postnatal GA determination such as the Ballard Score.

**Statistical modelling:**

- Statistical modelling uses split-sampling for model derivation and model validation. Data splitting is an unstable method for validating models because if you were to split the data again, develop a new model on the training sample, and test it on the holdout sample, the results are likely to vary significantly. Recommended resampling approaches are cross-validation and bootstrapping and the authors should consider this.
- The authors should also report recommended metrics for evaluating model performance i.e., discrimination and calibration of the models.
- Interaction terms for models 1-3 – could the authors comment on the added value of the interaction parameters and how much improvement in model performance can be attributed to the inclusion of the interaction parameters?
- Reference for classifying SGA – the authors provide the reference INTERGROWTH-21st very preterm size at birth reference charts. The reference provided is only for infants born 24 to <33 weeks. What about for infants born after 33 weeks? Can the authors confirm that for infants born <sup>3</sup>33 weeks they used the IG standards provided here: Villar *et al.* (2014<sup>1</sup>).

**Results:**

- The authors should comment on the very low % preterm across the three cohorts. According to Blencowe *et al.* (2012<sup>2</sup>), in 2010, the preterm birth rate in Philippines was estimated to be 15% (vs 4.6% in current cohort) and was 7.1% in China (vs 4.8% in current cohort).
- In table 3, can the authors also include % low birth weight and % LGA?

**Minor comments:**

- It is unnecessary to have elastic net in the title – it distracts the main focus of the paper.

**References**

1. Villar J, Ismail L, Victora C, Ohuma E, et al.: International standards for newborn weight, length, and head circumference by gestational age and sex: the Newborn Cross-Sectional Study of the INTERGROWTH-21st Project. *The Lancet*. 2014; **384** (9946): 857-868 [Publisher Full Text](#)
2. Blencowe H, Cousens S, Oestergaard M, Chou D, et al.: National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *The Lancet*. 2012; **379** (9832): 2162-2172 [Publisher Full Text](#)

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Partly

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Medical Statistician with experience in the field of maternal and child health

**We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.**

Author Response 13 May 2021

**Kumanan Wilson**, Ottawa Hospital Research Institute, Ottawa, Canada

*Thank you for your review. We have carefully responded to your concerns. We are updating the manuscript accordingly and it will be posted shortly.*

**Reviewer:**

In this manuscript by Hawken *et al.*, the authors have performed an external validation of newborn metabolomic markers for postnatal GA estimation in East and South-East Asian infants using an Elastic Net regression modelling approach.

**Main comments:**

- There is no doubt that an accurate estimate of GA is key. However, the authors propose a postnatal estimation of GA which does very little in advancing and encouraging determination of GA in early pregnancy. This is the recommendation of the WHO, The Brighton Collaboration GAIA definitions Prematurity and assessment of gestational age, The National Institute for Health and Care Excellence (NICE) Guideline for Routine Antenatal Care (2008), and International Society of Ultrasound in Obstetrics and Gynaecology (ISUOG).

**Response:**

*While the authors agree that accurate estimation of GA in early pregnancy is critical as the reviewer has outlined above, our objective was to provide a non-invasive alternative in low*

***resource settings where prenatal care involving GA dating ultrasound in the first trimester is not widely accessible.***

**Reviewer:**

There is a clear gap identified in the accuracy of determining GA especially in LMIC and I strongly doubt this approach will help in better characterising of the burden of vulnerable newborns and the potential impact towards achieving better estimates for population rates of preterm birth, low birth weight, small-for-gestational age, and combinations of these to identify other vulnerable newborn phenotypes.

**Response:**

***Post-natal estimation of GA has been identified as a priority of the Gates Foundation to facilitate population-based surveillance <https://www.linkedin.com/pulse/innovation-how-50-year-old-drop-blood-helps-solve-urgent-mundel/> . In many LMIC's, accurate early pregnancy estimations of GA are not accessible due to lack of ultrasound, pre-natal care and the unreliability of LMP. Our goal was to develop and refine one more potential tool to support maternal newborn care and surveillance in low resource settings. This external validation study has provided important information on the strengths and limitations of applying this method in different settings. In combination with the work of others, this may be a component of a broader solution that combines the strengths of different approaches. This may be an acceptable approach for a demographic surveillance site, for example. Ultimately, non-invasive 'omics methods may provide an alternative to first trimester ultrasound when not available. Furthermore, this approach could assist in distinguishing between SGA infants and pre-term infants to facilitate care at the individual infant level. However, the limitations of these approaches need to be recognized and our work identifies some of these challenges.***

**Reviewer:**

Overall reporting of results warrants improvement, the authors have decided to report overall mean agreement in GA between gold standard GA with model predicted GA and yet there are clearly large differences in precision and this differs according to GA (Figure 1). The authors should state explicitly what is meant by agreement within 7 days. For example, if on average, model estimate agrees within 7 days, this technically means  $\pm 7$  days and therefore for a given fetus say model GA estimate is 32 weeks + 0 days, this would mean that the true GA ranges between 31 weeks + 0 days and 33 weeks + 0 days which is effectively 2 weeks. Following this, the best model estimate (model 3) on average will be accurate to within 10 days at best. The authors should show a plot of true GA vs. predicted GA as this will evidently show the variability of the prediction as a function of GA as opposed to the aggregated estimates they have presented by GA in Figure 1.

**Response:**

***Thank you. We have clarified that "within" indicates  $\pm X$  days from the true GA throughout the manuscript. We have now also included residual plots of predicted - observed by ultrasound-assigned gestational age (new Figure 2 in manuscript).***

**Reviewer:**

Across all models, great discrepancies and perhaps unacceptable discrepancies are

observed for GA before 39 weeks. I am not convinced this approach offers any benefit/added value/utility compared to other methods in common use such as best obstetric methods for ascertaining GA.

**Response:**

***We agree that the model did not perform well in infants born before 39 weeks' gestation. However, results from our prospective validation studies are much more promising (1,2). In our previous prospective Bangladesh cohort, the model accurately estimated gestational age +/- 6 days (2). The current paper represents one component of our external validation strategy that included: a) validation in retrospective cohorts from established newborn screening programs and b) prospective validation in low-resource settings with primary data collection and better control over the quality of the ascertainment of the gold standard GA measurement.***

1. ***Murphy M, Hawken S, Cheng W, Wilson L, Lamoureux M, Henderson M et al. Postnatal gestational age estimation using newborn metabolic profiles: A validation study in Matlab, Bangladesh. Elife. 2019 Mar 19;8. pii: e42627. doi: 10.7554/eLife.42627.***

1. ***Hawken S, Ducharme R, Murphy MSQ, Olibris B, Bota AB, Wilson LA, et al. Development and external validation of machine learning algorithms for postnatal gestational age estimation using clinical data and metabolomic markers. BMC Medical Informatics and Decision Making (under review). Preprint: medRxiv 2020.07.21.20158196; doi: <https://doi.org/10.1101/2020.07.21.20158196>***

***The purpose of this study was to evaluate the utility of a model derived in the Ontario dataset but deployed in an external setting. It is possible that deriving models for each external site using the external data may yield more robust results, and this is something to possibly explore in the future.***

**Reviewer:**

The team used blood spot samples collected within 48hrs of delivery – there is considerable extra effort involved, time, and cost for drawing blood spots and processing of analytes. I do not see how this would be a feasible alternative especially for LMIC where accurate estimation of GA is a key data gap. The merits of the proposed approach have to clearly outweigh the performance of other known methods for postnatal GA determination such as the Ballard Score.

**Response:**

***Our objective in this external validation study was not to address feasibility, but rather to assess the performance of models developed in a North American cohort in infants in other international settings. The BMGF has funded our group to assess the feasibility of implementing our GA estimation method in multiple LMIC settings, including comparative accuracy/costs/burden versus other available methods so we will be able to address these important considerations when we publish.***

**Statistical modelling:**

**Reviewer:**

statistical modelling uses split-sampling for model derivation and model validation. Data splitting is an unstable method for validating models because if you were to split the data again, develop a new model on the training sample, and test it on the holdout sample, the results are likely to vary significantly. Recommended resampling approaches are cross-validation and bootstrapping and the authors should consider this.

**Response:**

***We agree with this comment, but in the context of much smaller databases. Although our modeling process did use bootstrapping and cross-validation in the model training phase (i.e. for optimizing ELASTICNET hyperparameters), our available sample was large enough to ensure that the test subset was large, stable and had a similar distribution of perinatal characteristics. Using a held-out test set also allowed us to exactly replicate the analysis pipeline that would be used in our external validations (i.e. data preprocessing, normalization etc. were executed separately, but using the same algorithm in the model training subset and in the internal validation subset, as well as in external validation settings, which would not be possible if a cross-validation approach was used). To reinforce this point, we conducted sensitivity analyses where we used different training and testing splits, and also a cross-validation approach, and these yielded nearly identical results.***

**Reviewer:** The authors should also report recommended metrics for evaluating model performance i.e., discrimination and calibration of the models.

**Response:**

***In this paper, we did not report results from a logistic regression model or other method meant to classify term vs. preterm birth, so discrimination is not relevant in the context of the models we are reporting. However, based on our model estimates, we have reported the preterm birth rate that is based on observed GA above and below 37 weeks and model-predicted GA above and below 37 weeks. This represents calibration in the large, which we have now clarified in the manuscript, and we have included plots of observed vs predicted GA (new Figure 2 in manuscript).***

**Reviewer:**

Interaction terms for models 1-3 – could the authors comment on the added value of the interaction parameters and how much improvement in model performance can be attributed to the inclusion of the interaction parameters?

**Response:**

***The inclusion of the interaction terms improved model performance appreciably, both overall and in important subgroups (<37 weeks and SGA10). The addition of the interaction terms reduced the MAE from 0.75 to 0.71 overall, from 1.14 to 1.03 in the <37 weeks subgroup, and from 1.39 to 1.13 in the SGA10 subgroup. We have now commented on the effect of the interactions in the methods section.***

**Reviewer:**

Reference for classifying SGA – the authors provide the reference INTERGROWTH-21st very preterm size at birth reference charts. The reference provided is only for infants born 24 to <33 weeks. What about for infants born after 33 weeks? Can the authors confirm that for

infants born <sup>33</sup> weeks they used the IG standards provided here: Villar *et al.* (2014<sup>1</sup>).

**Response:**

**Yes, we confirm that the reference provided by the reviewer was used for infants born 33 weeks and older. We have updated the text accordingly.**

**Results:**

**Reviewer:**

The authors should comment on the very low % preterm across the three cohorts. According to Blencowe *et al.* (2012<sup>2</sup>), in 2010, the preterm birth rate in Philippines was estimated to be 15% (vs 4.6% in current cohort) and was 7.1% in China (vs 4.8% in current cohort). Thank you we have commented on this in the discussion. The text now reads:

**Response: The preterm birth rate that we estimated in the current cohort (Philippines: 4.6%, China: 4.8%) was less than previously estimated (Philippines: 15%, China: 7.1%)(3). Although newborn screening initiatives are meant to be universal in these populations, some tests were paid for by the families. In the Philippines, all newborn screening has been covered since 2019, but prior to that only 4 of the tests were covered. Considering we only tested samples which the full panel was available this could have contributed to selection bias in our sample population where infants born in higher resource/urban areas were preferentially included where preterm birth rates could be substantially different. We also excluded infants in whom samples were collected later than 72 hours after birth as these are subject to a high level of heterogeneity. This had a similar effect as it did in Ontario, disproportionately excluding preterm infants, lowering the preterm birth rate from 6.9% to 4.8% in the China cohort (and thus accounting for the bulk of the discrepancy there) and from 5.6% to 4.62% in the Philippines cohort (thus accounting for only a very small part of the discrepancy).**

1. Blencowe H, Cousens S, Oestergaard M, Chou D, et al.: National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *The Lancet*. 2012; **379** (9832): 2162-2172

**Reviewer:**

In table 3, can the authors also include % low birth weight and % LGA?

**Response:**

**We have added low birthweight and LGA to Table 3.**

**Reviewer;**

**Minor comments:**

- It is unnecessary to have elastic net in the title – it distracts the main focus of the paper.

**Response:**

**Thank you, we have made this change. The title is now: *External validation of machine learning models including newborn metabolomic markers for postnatal gestational age***



**estimation in East and South-East Asian infants****Reviewer References**

1. Villar J, Ismail L, Victora C, Ohuma E, et al.: International standards for newborn weight, length, and head circumference by gestational age and sex: the Newborn Cross-Sectional Study of the INTERGROWTH-21st Project. *The Lancet*. 2014; **384** (9946): 857-868 [Publisher Full Text](#)
2. Blencowe H, Cousens S, Oestergaard M, Chou D, et al.: National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *The Lancet*. 2012; **379** (9832): 2162-2172 [Publisher Full Text](#)

**Competing Interests:** No competing interests were disclosed.

Reviewer Report 26 March 2021

<https://doi.org/10.21956/gatesopenres.14318.r30362>

© 2021 Sazawal S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Sunil Sazawal**

<sup>1</sup> Centre for Public Health Kinetics, New Delhi, Delhi, India

<sup>2</sup> Johns Hopkins School of Public health, Baltimore, MD, USA

**General Comments:**

This is an interesting study where the authors have collected unique datasets both from developed country and developing country settings. The paper addresses a very pertinent question related to the gestational age dating especially in the developing country settings where early ultrasound dating is missing due to unavailability of resources or because of the cost involved in getting an ultrasound done. But it is yet to be seen whether taking blood spots in on Whatman paper is feasible and getting the screen done in a lab with MS is possible in these settings. The paper is generally well written and structured with enough details, however, there are some queries which the authors need to address to make the paper clearer and more transparent.

Paper should be accepted for indexing with revisions.

1. Major Concern:

- Although not highlighted but proved in the original manuscript there were exclusions and imputations. A section providing details of these and how these may have affected the outcome or made improvements needs to be clearly stated in methods and discussion.

- The full model seems to include addition Hb ratio's and analyses not part of the newborn screening routinely, in terms of implications and discussion this needs better discussed. While in results a result that was obtained with metabolic screen with clinical variable routinely available as birth weight needs to be the key primary estimate and other estimates need to be provided as secondary exploratory results. The distinction as provided current is blurred and confuses the reader.

2. Other concerns:

- Reference GA assessment

*Statement:* In the Philippines cohort, mothers who delivered in private hospitals generally received gestational dating ultrasounds while other infants' GAs were generally measured using Ballard Scoring

*Q:* It is a discrepancy since Ontario based models trained data against USG confirmed GA. Then under **External validation** where Philippines samples were used: How can both ultrasound and Ballard scoring used under same bracket.

Internal validation of model performance in Ontario, Canada

*Q:* Was the internal validation performed with previously developed models including 47 analytes, birth weight, sex or the restricted model, needs clarification and discussed

*Q:* Restricted model definition

The proper definition of the restricted model is missing. Was a restricted model built separately for Manila and Shanghai or Separate models model were made for Manila and Shanghai

Statistical methods

*Statement:* In the Ontario cohort, all screen-positive results were excluded from analysis, which had the effect of removing a large proportion of extreme outliers and a typical metabolic profiles.

*Q:* What is the meaning of screen positive? Does this mean that all children who had a metabolic disorder which might have abnormal values for some metabolites were removed? If so, was it done in the other two datasets and will the model then be not applicable to children who show abnormal values for the metabolites.

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Partly

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

No

**Are the conclusions drawn adequately supported by the results?**

Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Infectious disease, pediatrics, epidemiology and statistics, global health

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 05 May 2021

**Kumanan Wilson**, Ottawa Hospital Research Institute, Ottawa, Canada

**Reviewer 1**

General Comments:

This is an interesting study where the authors have collected unique datasets both from developed country and developing country settings. The paper addresses a very pertinent question related to the gestational age dating especially in the developing country settings where early ultrasound dating is missing due to unavailability of resources or because of the cost involved in getting an ultrasound done. But it is yet to be seen whether taking blood spots in on Whatman paper is feasible and getting the screen done in a lab with MS is possible in these settings. The paper is generally well written and structured with enough details, however, there are some queries which the authors need to address to make the paper clearer and more transparent.

Paper should be accepted for indexing with revisions.

Major Concerns:

- Although not highlighted but proved in the original manuscript there were exclusions and imputations. A section providing details of these and how these may have affected the outcome or made improvements needs to be clearly stated in methods and discussion

**Response:** *No imputation of missing analyte/covariate values was undertaken either in the Ontario cohort or in the international cohorts given large sample sizes and very low occurrence of missing values. We have now provided further detail on the exclusion criteria applied in preparing the Ontario cohort for model development. The two criteria leading to the most exclusions were 1) requiring gold-standard GA measurement via 1st-trimester dating ultrasound, and 2) screening bloodspot collection within 48 hours of birth. The first exclusion criteria may have excluded infants born in rural or underserved areas of the province where access to comprehensive prenatal care was lower. In many cases however, this is more likely to be a data quality issue, where dating ultrasound was used but not recorded as*

*such. The second criteria led to the disproportionate exclusion of preterm infants who more often had delayed sample collection, despite this not being recommended practice. Although this exclusion biased the rate of preterm gestation observed in our Ontario study cohort downward, but it was unlikely to have had any important impact on GA model development, as we still had a large sample size across the full spectrum of gestational ages at birth to allow robust model development and performance evaluation. Further, the inclusion of samples collected later than 48 hours would introduce a large amount of heterogeneity in analyte levels which had to be balanced against the impact of exclusions. We have added more details to the methods and discussion reflecting these considerations.*

**Reviewer:**

The full model seems to include addition Hb ratio's and analyses not part of the newborn screening routinely, in terms of implications and discussion this needs better discussed. While in results a result that was obtained with metabolic screen with clinical variable routinely available as birth weight needs to be the key primary estimate and other estimates need to be provided as secondary exploratory results. The distinction as provided current is blurred and confuses the reader.

**Response:**

*Hemoglobin types (eg HbF HbA - fetal and adult hemoglobin types) are measured during routine NBS in Ontario and the Philippines, in the course of identifying mutant types associated with hemoglobinopathies. These are reported as "peak percentages" with respect to total hemoglobin. We have used the peak percentages for normal HbF, HbF1 and HbA to construct a ratio of fetal to adult hemoglobin by calculating  $(HbF+HbF1)/(HbF+HbF1+HbA)$ , to measure the proportion of normal fetal hemoglobin, relative to the proportion of total normal fetal + adult hemoglobin types. This is strongly predictive of gestational age as the transition from fetal to adult hemoglobin occurs apace with fetal development. We have added these details to the Methods.*

**Reviewer:**

Other concerns:

Reference GA assessment

Statement: In the Philippines cohort, mothers who delivered in private hospitals generally received gestational dating ultrasounds while other infants' GAs were generally measured using Ballard Scoring

Q: It is a discrepancy since Ontario based models trained data against USG confirmed GA. Then under External validation where Philippines samples were used: How can both ultrasound and Ballard scoring used under same bracket.

**Response:**

*Although we knew general practice patterns of gestational dating method used in the Philippines, we did not have individual-level data on what method was used in each individual pregnancy, hence we accepted this as an additional source of validation error which would lead to*

*larger MAE/RMSE. This was presented as a limitation in the discussion and we have clarified the description in the methods. It now reads:*

*'In the Philippines cohort, mothers who delivered in private hospitals generally received gestational dating ultrasounds while other infants' GAs were generally measured using Ballard Scoring, however individual-level data identifying which GA measurement method was used was not available.'*

**Reviewer:**

Internal validation of model performance in Ontario, Canada

Q: Was the internal validation performed with previously developed models including 47 analytes, birth weight, sex or the restricted model, needs clarification and discussed

**Response:**

*The internal validation of models 1-3 was conducted on an independent test dataset of Ontario infants using all 47 analytes, multiple gestation, birthweight and sex. Since not all of the predictors available for the Ontario dataset were also available for the external datasets (multiple gestation and a small subset of analytes were absent), we tailored the models to include the maximum number of available predictors in each of the external datasets (which we called 'restricted models'). The list of analytes available at each site is presented in Table 1. The tailored models were fit in the Ontario dataset, and validated in the Ontario test set and external datasets. We have added a figure (Figure 1) to clarify the different models tested. The Methods section now states:*

*'A total of 47 newborn screening analytes, as well as sex, birth weight and multiple birth status, were used in the original Ontario model development. GA at birth (in weeks) determined by first trimester gestational dating ultrasound was the dependent variable. A subset of screening analytes, as well as multiple gestation status were not available in the external cohorts (Table 1). Three main models were developed and evaluated in the Ontario cohort (Table 2). For models 2 and 3, we also developed restricted models including only the covariates available in each of the two external cohorts (Figure 1). Restricted models were trained on the Ontario datasets but deployed in the external cohorts.'*

**Reviewer:**

Q: Restricted model definition

The proper definition of the restricted model is missing. Was a restricted model built separately for Manila and Shanghai or Separate models model were made for Manila and Shanghai

**Response:**

*Restricted models were built separately to be applied in Manila and Shanghai based on availability of screening analytes/predictors in each setting. These models were trained in the Ontario data and then deployed in the external cohorts. We have updated the text and the methods now read:*

*A total of 47 newborn screening analytes, as well as sex, birth weight and multiple birth status, were used in the original Ontario model development. GA at birth (in weeks) determined by first trimester gestational dating ultrasound was the dependent variable. A subset of screening analytes as well as multiple gestation were not available in the external cohorts (Table 1). Three main models were developed and evaluated in the Ontario cohort (Table 2). Model 1 was developed excluding multiple gestation status, and for models 2 and 3, we also developed restricted models including only the covariates available in each of the two external cohorts (Figure 1). All of the restricted models were trained on the Ontario datasets and deployed in the external cohorts.*

**Reviewer:**

Statistical methods

Statement: In the Ontario cohort, all screen-positive results were excluded from analysis, which had the effect of removing a large proportion of extreme outliers and atypical metabolic profiles.

Q: What is the meaning of screen positive? Does this mean that all children who had a metabolic disorder which might have abnormal values for some metabolites were removed? If so, was it done in the other two datasets and will the model then be not applicable to children who show abnormal values for the metabolites.

**Response:**

*Screen positive refers to infants who tested positive for a disorder in the screening panel. We have clarified this statement in the methods. These infants were excluded from the Ontario population as they tend to have extreme outlying values for some analytes which impact negatively on model development. Additionally, we employed a strategy of winsorizing extreme values that lay more than three IQRs above the third quartile or three IQRs below the first quartile. Winsorizing replaces these extreme outliers with the upper and lower boundary value for the analyte, which preserves the extremeness, but reduces the impact of the original value. The same winsorization algorithm was applied in the external cohorts. Screen positive data points were not explicitly removed from the Philippines and China datasets. The reviewer is correct that the model may not be as accurate for children with abnormal values, however in the external settings where this algorithm is being deployed, it would not be known whether infants had a disorder at birth, so the model would need to be as robust as possible in estimating GA under these conditions. Because of the approach we took, the impact of abnormal/extreme values would be attenuated by our data normalization strategy which included both log transformation and Winsorization of extreme outliers. The occurrence of extreme outliers for either screen positive infants or for other reasons was extremely low, so would only affect a small number infants, but our strategy allowed us to produce a GA estimate in these infants that was robust to extreme values and less likely to produce a wildly inaccurate estimate. We have clarified these details in the Methods.*

**Competing Interests:** No competing interests were disclosed.