External validation of ELASTIC NET regression models including newborn metabolomic markers for postnatal gestational age estimation in East and South-East Asian infants [version 1; peer review: awaiting peer review]

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

\textsuperscript{1}Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada
\textsuperscript{2}School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada
\textsuperscript{3}Newborn Screening Reference Centre, University of the Philippines Manila, Manila, Philippines
\textsuperscript{4}Pediatric Endocrinology and Genetic Metabolism, XinHua Hospital, Shanghai, Shanghai, China
\textsuperscript{5}Better Outcomes Registry & Network, Ottawa, Canada
\textsuperscript{6}Department of Gynecology and Obstetrics, XinHua Hospital, Shanghai, Shanghai, China
\textsuperscript{7}MOE-Shanghai Key Lab of Children's Environmental Health, Xinhua Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
\textsuperscript{8}Department of Pediatrics, University of the Philippines Manila, Manila, Philippines
\textsuperscript{9}Institute of Human Genetics, National Institutes of Health, University of Philippines Manila, Manila, Philippines
\textsuperscript{10}Newborn Screening Ontario, Children's Hospital of Eastern Ontario, Ottawa, ON, Canada
\textsuperscript{11}Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

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 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 group of infants, and externally validated in cohorts of infants from the Philippines and China.

Open Peer Review

Reviewer Status: AWAITING PEER REVIEW

Any reports and responses or comments on the article can be found at the end of the article.
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 5 days of ultrasound values in the Canadian internal validation (mean absolute error (MAE) 0.71, 95% CI: 0.71, 0.72), and within 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

Corresponding author: Steven Hawken (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; Cheng 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 [OPP114535]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Hawken S et al. This is an open access article 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.

How to cite this article: Hawken S, Murphy MSQ, Ducharme R et al. External validation of ELASTIC NET regression models including newborn metabolomic markers for postnatal gestational age estimation in East and South-East Asian infants [version 1; peer review: awaiting peer review] Gates Open Research 2020, 4:164 https://doi.org/10.12688/gatesopenres.13131.1

First published: 29 Oct 2020, 4:164 https://doi.org/10.12688/gatesopenres.13131.1
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. As complications related to preterm birth continue to be the most common cause of mortality for children under five, 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, but under special circumstances (e.g., preterm birth, neonatal transfer) may be collected later. 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 immunosassays, and high-performance liquid chromatography. 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 1 to 2 weeks. Recent work to refine metabolic GA models, 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.

Published approaches to model development and validation to date have been carried out in cohorts of infants from in North American settings. Although internal validation of these models has been conducted among infants from diverse ethnic backgrounds, 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. 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. 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 database study that relied on secondary use 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 (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

Infants whose blood spot samples were collected more than 48 hours after birth were excluded from model development in the Ontario cohort. In the China and Philippines datasets, the age of infant at sample collection was only available to the nearest calendar day. Samples were excluded from analysis if they were collected more than 72 hours after birth as most samples would have been excluded if the >48-hour exclusion were applied to these validation cohorts.

Newborn Screening Ontario (NSO): a provincial newborn screening program that coordinates the screening of 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. Newborn screening data collected between January 2012 and December 2014 were used in model building and internal validation.

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, either a basic panel of six disorders or an expanded panel of 28 disorders. 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%) annually for between 4 and over 20 rare conditions including metabolic and endocrine diseases. Four screening tests - for phenylketonuria, congenital adrenal hyperplasia, hypothyroidism and Glucose-6-phosphate dehydrogenase deficiency - are funded by the government. Screening tests reliant on tandem mass spectrometry are funded by the newborn’s family or the Shanghai neonatal screening center. 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 and for whom tandem mass spectrometry data were available 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. GAs were 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. 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 cleaning and normalization.** Data from the Ontario cohort were used for model development. Many of the details of the data preparation, model building, and internal validation have been reported previously. A series of steps were taken to prepare the newborn screening analyte data for modeling:

1. 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. Further, samples used in model development were limited to those collected within 48 hours of birth given that GA estimation is intended to be applied in LMICs where samples are expected to be collected almost exclusively within the first few hours after birth.

**Statistical modelling.** The Ontario-derived dataset 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 as the overall cohort.

A total of 47 newborn screening analytes, as well as sex, birth weight and multiple birth status, were used in our original 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, therefore we developed restricted Ontario models including those covariates available in each of the two external cohorts. Three main models were derived and evaluated (with variations in the included analyte predictors based on availability in each external cohort) (Table 2):

Model 1: Baseline model containing only infant sex, birth weight (grams), and the interaction between these, sex and birth weight.

---

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

<table>
<thead>
<tr>
<th>Newborn Screening Ontario, Canada</th>
<th>Newborn Screening Reference Centre, the Philippines</th>
<th>Shanghai Newborn Screening Program, China</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth weight, Hemoglobins (F1, F, A), TSH, 17OHP, alanine, arginine, citruline, glycine, leucine, methionine, ornithine, phenylalanine, tyrosine, valine, C0, C2, C3, C4, C5, C6, C8, C10, C12, C14, C16, C18, C10:1, C12:1, C140H, C14:1, C14:2, C16:1OH, C16:10H, C18:1OH, C18:0H, C18:2OH, C18:1OH, C18:2OH, C18:3OH, C3DC, C4DC, C4OH, C5DC, C5OH, C5:1, C6DC, C8:1, IRF, GALT, BIO</td>
<td>Birth weight, Hemoglobins (F1, F, A), TSH, 17OHP, alanine, arginine, citruline, glycine, leucine, methionine, ornithine, phenylalanine, tyrosine, valine, C0, C2, C3, C4, C5, C6, C8, C10, C12, C14, C16, C18, C10:1, C12:1, C140H, C14:1, C14:2, C16:1OH, C16:10H, C18:1OH, C18:2OH, C18:3OH, C3DC, C4DC, C4OH, C5DC, C5OH, C5:1, C6DC, C8:1, IRF, BIO</td>
<td>Birth weight, TSH, 17OHP, alanine, arginine, citruline, glycine, leucine, methionine, ornithine, phenylalanine, tyrosine, valine, C0, C2, C3, C4, C5, C6, C8, C10, C12, C14, C16, C18, C10:1, C12:1, C140H, C14:1, C14:2, C16:1OH, C16:10H, C18:1OH, C18:2OH, C18:3OH, C3DC, C4DC, C4OH, C5DC, C5OH, C5:1, C6DC, C8:1</td>
</tr>
</tbody>
</table>

TSH, thyroid stimulating hormone; 17OHP, 17 hydroxyprogesterone; GALT, galactose-1-phosphate uridyl transferase; IRF, Immuno-reactive trypsinogen; BIO, biotidnase. Analytes in **bold italics** are not available in one or more of the external validation cohorts.

---

### Table 2. Summary of models tested.

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>Multivariable regression model including sex, birth weight and their interaction</td>
</tr>
<tr>
<td>Model 2</td>
<td>ELASTIC NET regression model including sex, analytes and pairwise interactions among predictors</td>
</tr>
<tr>
<td>Model 3</td>
<td>ELASTIC NET regression model including sex, birth weight, analytes and pairwise interactions among predictors</td>
</tr>
</tbody>
</table>
MSE, also referred to as root mean square error (RMSE), is also expressed in the same units as GA (weeks). Lower values of both MAE and standard error of estimation (RMSE) reflects more accurate model estimated GA. For example, a reported MAE of 1.0 weeks 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 7 and 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. Parametric standard error estimates were not readily calculable for all of our performance metrics, therefore we calculated 95% bootstrap percentile confidence intervals for all validation performance metrics, based on the 2.5th and 97.5th percentiles for performance metrics over 1000 bootstrap replicates for 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, Table 3. Cohort Characteristics.

<table>
<thead>
<tr>
<th></th>
<th>Canada n=39,666 (Ontario test cohort)</th>
<th>Philippines n=82,909</th>
<th>China n=4,448</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>19,536 (49.3%)</td>
<td>42,867 (51.7%)</td>
<td>2351 (52.9%)</td>
</tr>
<tr>
<td>Female</td>
<td>20,130 (50.5%)</td>
<td>40,042 (48.3%)</td>
<td>2097 (47.1%)</td>
</tr>
<tr>
<td>Birth weight (g), mean±SD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>3379 ± 530.2</td>
<td>3007.7 ± 452.0</td>
<td>3337 ± 437.3</td>
</tr>
<tr>
<td>Term infants only</td>
<td>3430.6 ± 476.1</td>
<td>3044.4 ± 413.5</td>
<td>3369.1 ± 406.7</td>
</tr>
<tr>
<td>Preterm infants only</td>
<td>2504.1 ± 622.8</td>
<td>2250.1 ± 538.8</td>
<td>2709.6 ± 535.5</td>
</tr>
<tr>
<td>SGA (&lt;10th Centile), n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1561 (3.94%)</td>
<td>11,295 (13.6%)</td>
<td>123 (2.8%)</td>
<td></td>
</tr>
<tr>
<td>SGA (&lt;3rd Centile), n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>363 (0.92%)</td>
<td>3407 (4.1%)</td>
<td>19 (0.4%)</td>
<td></td>
</tr>
<tr>
<td>Completed gestational age (wks), mean±SD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Term (≥37 wks), n (%)</td>
<td>37,440 (94.4%)</td>
<td>79,077 (95.4%)</td>
<td>4233 (95.2%)</td>
</tr>
<tr>
<td>Late Preterm (32–36 wks), n (%)</td>
<td>2049 (5.2%)</td>
<td>3566 (4.3%)</td>
<td>197 (4.4%)</td>
</tr>
<tr>
<td>Very Preterm (28–31 wks), n (%)</td>
<td>126 (0.3%)</td>
<td>233 (0.3%)</td>
<td>11 (0.3%)</td>
</tr>
<tr>
<td>Extremely Preterm (&lt;28 wks), n (%)</td>
<td>51 (0.1%)</td>
<td>33 (0.0%)</td>
<td>7 (0.2%)</td>
</tr>
</tbody>
</table>

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).
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 2226/39666 (5.6%) of the Ontario cohort, 3832/82909 (4.6%) of the Philippines cohort, and 215/4448 (4.8%) of the China cohort.

Internal validation of model performance in Ontario, Canada

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 7 days of reference GA. Model 2, (including sex and metabolomic markers), was accurate within an average of 6 days (MAE 0.79 (0.79, 0.80) weeks) and Model 3, which included sex, birth weight and metabolomic markers was the most accurate, estimating GA within about 5 days of ultrasound-assigned GA (MAE 0.71 (0.71, 0.72) weeks), and estimated GA within ± 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).

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 2 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 rate was 4.2% (95% CI: 4.1%, 4.4%), compared to 4.3% using the observed reference GA in the Philippines cohort. Both Model 1 and Model 2 overestimated the preterm birth rate, at 5.1% and 5.0%, 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 rate 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%) compared to 4.8% based on the observed reference GA in the China cohort. Model 2, the least accurate of the three in the China cohort, underestimated the preterm birth rate to be 3.6%.

Model performance across spectrum of GA

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 (Figure 1). Across the spectrum of ultrasound-assigned GA, Model 3 provided the most accurate estimates overall.

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

<table>
<thead>
<tr>
<th>Models</th>
<th>Ontario Cohort</th>
<th>Philippines Cohort</th>
<th>China Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall, N=39,666</td>
<td>&lt;37 weeks, N=2226</td>
<td>SGA10, N=1561</td>
</tr>
<tr>
<td>Model 1: Sex and Birth weight</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAE (C)</td>
<td>0.96 (0.96, 0.97)</td>
<td>1.78 (1.73, 1.82)</td>
<td>2.71 (2.66, 2.76)</td>
</tr>
<tr>
<td>RMSE (C)</td>
<td>1.23 (1.22, 1.24)</td>
<td>2.13 (2.07, 2.18)</td>
<td>2.86 (2.80, 2.91)</td>
</tr>
<tr>
<td>% +/-1 wk</td>
<td>60.0 (59.5, 60.5)</td>
<td>30.2 (28.5, 32.0)</td>
<td>0.4 (0.1, 0.7)</td>
</tr>
<tr>
<td>% +/- 2 wks</td>
<td>90.9 (90.6, 91.2)</td>
<td>60.5 (58.4, 62.6)</td>
<td>21.7 (19.5, 23.8)</td>
</tr>
<tr>
<td>Model 2: Sex and Analytes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAE (C)</td>
<td>0.79 (0.79, 0.80)</td>
<td>1.25 (1.21, 1.29)</td>
<td>0.90 (0.85, 0.94)</td>
</tr>
<tr>
<td>RMSE (C)</td>
<td>1.02 (1.01, 1.03)</td>
<td>1.57 (1.51, 1.62)</td>
<td>1.19 (1.13, 1.26)</td>
</tr>
<tr>
<td>% +/-1 wk</td>
<td>69.4 (69.0, 69.9)</td>
<td>46.9 (45.0, 48.9)</td>
<td>66.6 (64.0, 69.0)</td>
</tr>
<tr>
<td>% +/- 2 wks</td>
<td>95.1 (94.9, 95.4)</td>
<td>80.3 (78.8, 82.0)</td>
<td>91.7 (90.3, 93.1)</td>
</tr>
<tr>
<td>Model 3: Sex, Birth weight and Analytes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAE (C)</td>
<td>0.87 (0.87, 0.88)</td>
<td>1.48 (1.44, 1.52)</td>
<td>0.98 (0.94, 1.03)</td>
</tr>
<tr>
<td>RMSE (C)</td>
<td>1.12 (1.11, 1.13)</td>
<td>1.80 (1.75, 1.85)</td>
<td>1.31 (1.25, 1.38)</td>
</tr>
<tr>
<td>% +/-1 wk</td>
<td>64.9 (64.5, 65.5)</td>
<td>39.8 (37.9, 42.0)</td>
<td>60.9 (58.3, 63.4)</td>
</tr>
<tr>
<td>% +/- 2 wks</td>
<td>93.0 (92.8, 93.3)</td>
<td>71.8 (70.1, 73.8)</td>
<td>91.2 (89.7, 92.6)</td>
</tr>
</tbody>
</table>

Data are presented as the mean and 2.5th and 97.5th bootstrap percentiles for MAE, RMSE and the percentage of model estimates within 1 and 2 weeks of ultrasound GA for 1,000 bootstrap samples generated from each cohort.

Ontario, MAE (C) 0.96 (0.95, 0.97), RMSE (C) 1.23 (1.22, 1.24), % +/-1 wk 60.0 (59.5, 60.5), % +/- 2 wks 90.9 (90.6, 91.2).

Philippines-restricted, MAE (C) 0.87 (0.87, 0.88), RMSE (C) 1.12 (1.11, 1.13), % +/-1 wk 64.9 (64.5, 65.5), % +/- 2 wks 93.0 (92.8, 93.3).

China-restricted, MAE (C) 0.71 (0.71, 0.72), RMSE (C) 0.92 (0.91, 0.93), % +/-1 wk 74.6 (74.2, 75.1), % +/- 2 wks 97.0 (96.8, 97.2).

Gates Open Research 2020, 4:164 Last updated: 30 OCT 2020
Figure 1. 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).

Previous validation of our models among infants born 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, but in this study, as well as in an external validation of previously developed GA estimation models in a prospective cohort from South Asia, differences were more pronounced. Third, variations in the clinical measures of GA used across the cohorts may have impeded the accuracy of our algorithms. Our GA models were originally 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. 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 after birth 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

<table>
<thead>
<tr>
<th></th>
<th>Average MAE (95% CI)</th>
<th>Average MAE (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Philippines, n=82909</td>
<td>China, n=4448</td>
</tr>
<tr>
<td>Model 1: Birth Weight Model</td>
<td>0.96 (0.95, 0.97)</td>
<td>0.90 (0.87, 0.92)</td>
</tr>
<tr>
<td>Model 2: Analyte Model</td>
<td>1.02 (1.02, 1.03)</td>
<td>1.07 (1.04, 1.10)</td>
</tr>
<tr>
<td>Model 3: Full Model</td>
<td>0.90 (0.90, 0.91)</td>
<td>0.89 (0.86, 0.91)</td>
</tr>
</tbody>
</table>

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 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. Second, metabolic profiles may be influenced by the differences in genetic and environmental exposures experienced by each cohort, as well as non-biological heterogeneity attributable to different laboratories conducting the screening assays.
also likely relied on different equipment, assays and reagents to quantify the measured analytes. 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\textsuperscript{16,17}.

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.

While there are numerous options currently available to health care providers to determine postnatal GA, none are as accurate as first trimester dating ultrasound\textsuperscript{14}. 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\textsuperscript{16,17}. Metabolic GA estimation models in particular have proven particularly promising\textsuperscript{7}, and we continue to refine and evaluate these models in a variety of populations\textsuperscript{7,20} 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\textsuperscript{7}. 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\textsuperscript{20}, with validation of previously developed models already completed in Bangladesh\textsuperscript{9}. 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, 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. A copy of the study protocol approved by a technical and ethics review board that includes methods and statistical analysis plans;

b. Full name, designation, affiliation of the person with whom the data will be shared; and,

c. 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.

**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), 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\textsuperscript{14}.

License: GNU General Public License v3.
References