Machine Learning for Fetal Growth Prediction

Ashley I. Naimi,[a] *Robert W. Platt,*[b] *and Jacob C. Larkin*[c]

Abstract: Birthweight is often used as a proxy for fetal weight. Problems with this practice have recently been brought to light. We explore whether data available at birth can be used to predict estimated fetal weight using linear and quantile regression, random forests, Bayesian additive regression trees, and generalized boosted models. We train and validate each approach using 18,517 pregnancies (31,948 ultrasound visits) from the Magee-Womens Obstetric Maternal and Infant data and 240 pregnancies in a separate dataset of high-risk pregnancies. We also quantify the relation between smoking and small-for-gestational-age birth, defined as a birthweight in the lower 10th percentile of a population birthweight standard and estimated and predicted fetal weight standard. Using mean squared error and median absolute deviation criteria, quantile regression performed best among the regression-based approaches, but generalized boosted models performed best overall. Using the birthweight standard, smoking during pregnancy increased the risk of small-for-gestational-age 3.84-fold (95% CI: 2.70, 5.47). This ratio dropped to 1.65 (95% CI: 1.50, 1.81) when using the correct fetal weight standard, which was no different from the machine learning–based predicted standards, but higher than the regression-based predicted standards. Machine learning algorithms show promise in recovering missing fetal weight information. See video abstract at, http://links.lww.com/EDE/B314.
(*Epidemiology* 2018;29: 290–298)

Adequate growth from conception into early childhood is central to optimal health and survival.[1] Because of its predominant role in shaping the risk of adverse perinatal outcomes, fetal or intrauterine growth restriction is often itself considered to be an outcome of epidemiologic interest, usually quantified as small versus appropriate for gestational age. Additionally, there is a growing interest in quantifying the precise relation between the continuum of fetal growth and the occurrence of adverse perinatal and childhood outcomes. Because fetal (intrauterine) weight is typically not measured *in populo*, birthweight is often used as a proxy for fetal weight in population health studies. However, recent work shows that this practice can lead to potentially serious missing data or selection bias problems, particularly at early gestational ages.[2]

Attempts have been made to resolve the problems posed by missing fetal growth curves in clinical and population-level data. These include the use of population referents based solely on (1) birthweight,[3] (2) fetal weight estimates,[4,5] and (3) hybrid references that average growth curves generated from livebirth and intrauterine weights,[6] or switch from intrauterine to birth weight distributions at 37 weeks.[7] While these proposals partially overcome the missing data problem, they are based on the presupposition that fetal growth characteristics in a given sample of pregnancies can adequately serve as a referent for populations potentially far removed from the sample in which the referent curves were generated. Additionally, while fetal growth standards play an important clinical role in identifying high-risk patients, they cannot be used as a replacement for measured fetal weight data, particularly when interest lies in confounding adjustment or effect measure modification of an exposure-response relation by fetal weight history, or evaluating the gestational age–specific relation between the fetal weight continuum and the risk of adverse outcomes.

Here, we take a different approach to overcome the problem of missing fetal weights in population data. We focus on generating algorithms that take as input maternal and infant birth characteristics, and output predicted fetal weights. We then evaluate how predicted fetal weights compare with actual fetal weights using several different approaches: (1) by estimating the correlation between predicted and actual fetal weights; (2) by comparing the gestational week–specific quantiles of actual and predicted fetal weights; and (3) by examining the relation between smoking status and small-for-gestational-age birth, with the latter defined using actual and predicted data. In principle, correctly capturing features of birth data that accurately predict fetal weight across gestation would enable researchers to generate such predictions in a wide range of empirical settings.[8] We evaluate this principle in practice.

Submitted August 2, 2016; accepted November 14, 2017.
From the [a]Department of Epidemiology, University of Pittsburgh, Pittsburgh, PA; [b]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, QC, Canada; and [c]Magee Women's Research Institute, University of Pittsburgh; Pittsburgh, PA.
Supported, in part, by the University of Pittsburgh Center for Research Computing through the computing resources provided, and the assistance of Dr. Kim Wong.
Code/Data Availability: All software coded needed to reproduce these analyses is available on https://github.com/ainaimi.
The authors report no conflicts of interest.
SDC Supplemental digital content is available through direct URL citations in the HTML and PDF version of this article (www.epidem.com).
Correspondence: Ashley I. Naimi, Department of Epidemiology, University of Pittsburgh, 130 DeSoto Street, 503 Parran Hall, Pittsburgh, PA 15261. E-mail: ashley.naimi@pitt.edu.

ISSN: 1044-3983/18/2902-0290
DOI: 10.1097/EDE.0000000000000788

Specifically, we explore the potential of combining data available at birth with regression-based and machine learning algorithms to generate population-specific fetal weight curves for a given application. In particular, we: (1) assess whether machine learning algorithms can accurately predict estimated fetal weight over the course of gestation using readily available *ex utero* information (e.g., birthweight, race, gestational age at birth); and (2) assess the relationship between smoking and fetal/birthweight weight across gestation when the latter is quantified with: (1) clinically estimated and predicted fetal weight distributions and (2) actual birthweight distributions.

METHODS

Data

Data were obtained from the Magee Obstetric Medical and Infant (MOMI) database, which includes maternal, fetal, and neonatal demographic and clinical data from all deliveries at Magee-Womens Hospital, University of Pittsburgh Medical Center, Pittsburgh, PA. MOMI data were used with the approval of the Institutional Review Board of the University of Pittsburgh (IRB# PRO15020176, approved Feb 16 2015). Patients provided informed consent upon admission. The database is updated to incorporate new data every 3–6 months. With each update, data points included in the MOMI database are selected randomly and compared against prenatal records and delivery documentation to assure accuracy of variables including gestational age at delivery, neonatal sex, parity, Apgar scores. We considered all singleton pregnancies that resulted in live births between 1999 and 2013, our main cohort consisted of 18,517 live-born singleton pregnancies, with data on 31,948 clinical ultrasound visits from 20 weeks of gestation onward.

Because ultrasound is collected more regularly from high-risk women, our sample may be subject to selection bias. To evaluate the impact of the potentially higher proportion of high-risk women in our sample, we repeated all analyses in three subsets of the original MOMI sample: (1) pregnancies with 5-minute Apgar scores ≥ 8; (2) pregnancies with fewer than four ultrasounds; and (3) pregnancies with 5-minute Apgar scores ≥ 8 and fewer than four ultrasounds. Additional details are provided in eAppendix 1; http://links.lww.com/EDE/B301.

We also used data from 240 pregnancies from the Obstetrical Determinants of Neonatal Survival (ODNSR).[9] The study population consists of infants born in the National Institutes of Child Health and Human Development Maternal Fetal Medicine Units Network in 1992. The study sought to evaluate factors that facilitate predicting survival in extremely low birth weight infants. Eligible infants weighed less than 1,000 grams and were followed until death, discharge from the hospital, or 120 days of survival.

For both cohorts, ultrasound estimates of fetal weight (in grams) were our outcome, computed using the equation derived by Hadlock et al.[10] We also abstracted medical chart information on the gestational week at which the ultrasound was obtained, gestational age at birth, birthweight, race, maternal age at birth, 1- and 5-minute Apgar scores, infant gender, maternal smoking status, and the year of birth. Additional details are provided in the eAppendix; http://links.lww.com/EDE/B301.

Analytic Strategy

Fetal Growth Prediction

We examined the performance of regression-based and data mining techniques to predict fetal growth trajectories using routine information collected at delivery. Our analytic strategy, outlined in Figure 1, was to first split the MOMI data into algorithm development and algorithm validation subsamples. The validation subsample consisted of all pregnancies that occurred in 2010–2013 (inclusive, $N = 8{,}517$ pregnancies, $M = 13{,}354$ ultrasound visits). This allowed us to evaluate the performance of our prediction algorithms in a validation sub-sample temporally separated from our development sub-sample. Such temporal separation creates more difficult circumstances under which the predictions are made. The development subsample consisted of all available pregnancies prior to 2010 ($N = 16{,}156$ pregnancies, $M = 24{,}852$ ultrasound visits). The development subsample was then further split as required into training and testing datasets to

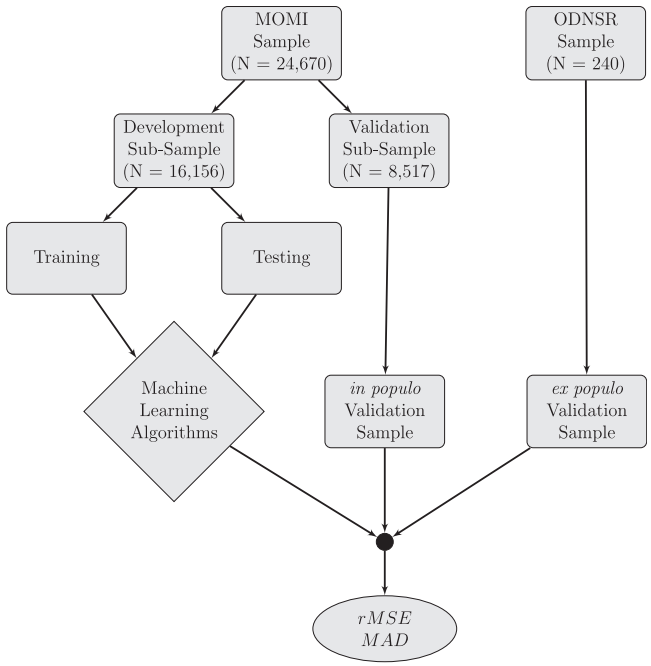

FIGURE 1. Analytic strategy used to train, test, and validate regression-based and machine learning algorithms to predict fetal growth. *MAD* indicates median absolute deviation; *rMSE*, root mean squared error.

optimize the selection of tuning parameters using K-fold cross validation.

Once generated, we input maternal and birth characteristics extracted from the MOMI validation subsample into our prediction algorithms to output fetal weight predictions for each pregnancy. Predictions were generated for each gestational week with an available fetal weight measurement. These predictions were then compared with actual fetal weight measurements obtained using standard ultrasound protocols. These comparisons were used to assess how well our algorithms performed at predicting fetal weight *in populo*.

We also carried out an *ex populo* assessment by inputting maternal and birth characteristics extracted from the ODNSR sample into our machine learning algorithms. Fetal weight predictions were obtained and compared with actual fetal weight estimates available in the ODNSR sample. This *ex populo* validation enabled us to assess how well our algorithms performed at predicting fetal weight for pregnancies in a population distinct from the one in which the algorithms were developed.

Regression-Based and Machine Learning Algorithms for Fetal Weight Prediction

We assessed how well two regression-based methods and three data adaptive ensemble learning algorithms predicted fetal weight. Regression-based methods consisted of standard linear regression[11] and quantile regression.[12] Machine learning algorithms consisted of random forests,[13] Bayesian additive regression trees,[14] and generalized boosted models.[15,16] Each approach (regression-based and machine learning) maps some component of the outcome's distribution (e.g., mean, median) to a function of the covariates $\mathbf{X}$, such as:

$$\gamma(Y \mid \mathbf{X} = \mathbf{x}) = h(\mathbf{x}). \quad (1)$$

Specific details on the variables in $\mathbf{X}$ for our analyses are included in the eAppendix; http://links.lww.com/EDE/B301. Here, we focus on a general overview of each approach. Standard linear regression takes $\gamma()$ as the expected value function, and $h(\mathbf{x}) = h(\mathbf{x};\beta)$ the linear predictor function in which categorical covariates were entered as disjoint indicator variables and continuous covariates were entered using restricted cubic splines. The parameters of this model were estimated by minimizing a quadratic loss function via ordinary least squares.

Quantile regression sets $\gamma()$ as the conditional quantile function $Q_y(\tau \mid \mathbf{X} = \mathbf{x})$, and $h()$ is the modified linear predictor in which parameters are made functionals of τ, where $\tau \in [0,1]$ indexes the τth quantile of the outcome's distribution. Again, categorical predictors were entered as disjoint indicator variables, and continuous predictors were entered using restricted cubic splines. The parameters of this model were estimated for $\tau = 0.5$ by minimizing an absolute loss function via least absolute deviations.[12]

Random forests, Bayesian additive regression trees, and generalized boosted models are based on classification and regression trees. For a continuous outcome, regression tree algorithms take $\gamma()$ to be the conditional expectation, and $h()$ a tree-based algorithm in which the sample of observations is recursively split into subsets defined by predictor values. Once sample-splitting is complete, the average outcome in each terminal node is estimated.

For random forests, many regression trees are fit to bootstrap resamples of the data and then their results are aggregated (averaged) into one tree, a process known as bagging.[13] Furthermore, the trees are built by choosing the best split in each bootstrap sample using a randomly selected subset of the predictors. Best splits that define each tree are determined by minimizing a residual sum of squares criterion.[17 (p233)] In our analysis, we fit 5,000 trees with a random selection of eight of 12 predictors.

For Bayesian additive regression trees, $h()$ is taken to be a sum of regression trees model, with the error assumed normally distributed with mean zero and variance σ^2. The parameters in each tree are solved by minimizing an expected loss function ($-$logLikelihood). The Bayesian additive regression tree also consists of a regularization prior on the joint distribution of the sum of trees model and the residual error variance. The posterior distribution of the sum of trees model is computed via the Markov Chain Monte Carlo algorithm, from which the averages over all trees are obtained. Tuning parameters for our Bayesian additive regression tree were chosen as the set of values that minimized cross-validation error in a preliminary run.

Finally, generalized boosted models treat $h()$ as a sequence of regression trees, iteratively combined to create an overall function for predicting average fetal weights. The trees in each iteration are fit to a randomly selected (but not bootstrapped) subsample of the data. Each subsequent tree is fit to the residuals of the model from the previous iteration, which results in "boosting" the performance of each subsequent fitting procedure. Predictions from each regression tree are combined using a shrinkage parameter that improves the overall fit of the sequence of models. A generalized cross validation criterion is used to avoid overfitting.[16,17 (p269)] We fit 5,000 trees with a shrinkage rate of 0.005, a subsampling rate of 50%, and fivefold cross-validation using a quadratic loss function.

Performance Measures

From each of these five approaches, we generated individual-level predictions for estimated fetal weight. The performance of each approach was then evaluated using the root mean squared error and median absolute deviation, defined as:

$$rMSE = \sqrt{\frac{1}{N}\Sigma_{ij}(y_{ij} - \hat{y}_{ij})^2}, \quad (2)$$

$$MAD = m(|\, y_{ij} - \hat{y}_{ij} \,|), \quad (3)$$

where y_{ij} and $\hat{y}_{ij}$ denote the actual and predicted fetal weight estimates for pregnancy i at visit j, and where $m()$ denotes the median function. These functions quantify the squared and absolute distance between the actual and predicted fetal weights on the original scale of interest (grams), and thus provide a summary measure of how well each method performs at predicting fetal weight. Variability estimates for root mean squared error and median absolute deviation were obtained using the nonparametric percentile bootstrap with 2,000 resamples.

Small for Gestational Age Analysis

We examined the impact of using different criteria to define small-for-gestational-age birth in the MOMI data. Small-for-gestational-age was defined as a birthweight in: (1) the lower 10th percentile of a gestational age–specific population referent birthweight standard (SGA_B);[18] (2) the lower 10th percentile of the gestational age–specific distribution of actual fetal weight + birthweight in the MOMI data (SGA_{FB}); or (3) the lower 10th percentile of the gestational age–specific distribution of predicted fetal weight + birthweight in the MOMI data (SGA_{PB}). Hereafter, we refer to these as the birthweight standard, the actual fetal weight standard, and the predicted fetal weight standard, respectively. For each approach, we quantified the proportion of babies classified as small-for-gestational-age and estimated risk differences and ratios for the association between maternal smoking status (1 = any smoking during pregnancy, 0 otherwise) and small-for-gestational-age birth under each definition using a generalized linear regression model.[19] These models were adjusted for maternal race, infant gender, maternal age, and their two- and three-way interactions.

We also evaluated how smoking status related to the entire distribution of fetal/birthweight, rather than just the proportion of births classified as small-for-gestational-age. This step was important to evaluate whether our algorithms were able to accurately predict the entire conditional distribution of fetal weight, rather than just the first moment of the conditional distribution. To do this, we fit regression quantiles at the 25th, 50th, and 75th percentiles of the actual and predicted fetal weight + birthweight distribution, smoothed across gestation, among smokers and nonsmokers. We then took the differences in these actual and predicted fetal weight + birthweight percentiles at each gestational week and used a locally weighted scatterplot smoother (loess) to evaluate the quantile differences in weight between smokers and nonsmokers, smoothed across gestational age.

All analyses were conducted using R version 3.2.3 (Vienna, Austria).[20] Quantile regression was implemented using the quantreg package.[21] Random forests, Bayesian additive regression trees, and generalized boosted models were implemented using the quantregForest package,[22] bartMachine package,[23] and gbm package,[24] respectively.

RESULTS

Over the entire MOMI sample, 25%, 46%, 19%, and 10% had one, two, three, or four or more clinical ultrasound visits during the course of follow-up, respectively. The proportion of non-Hispanic black, non-Hispanic white, and Hispanic women was 18%, 73%, and 2%, respectively. Table 1 shows the median and interquartile values of several characteristics in both the MOMI validation subsample and the ODNSR sample. This table shows that infants in the ODNSR sample were very different than those in the MOMI cohort, with lower fetal weights, birth weights, gestational age at birth. Women in the ODNSR sample were also younger than women in the MOMI sample.

eFigure 1 in the eAppendix; http://links.lww.com/EDE/B301 shows the estimated fetal weight and birth weight distributions over the course of gestation, with solid and dashed lines representing smoothed regression quantiles for the median of birthweight alone and birthweight + fetal weight. These lines show that, at earlier gestational ages, the median birthweight is lower than the combined birth and fetal weights, in line with results from previous research.[2]

After fitting the regression-based and machine learning algorithms to the development subsample of the MOMI cohort, we generated corresponding predictions for each estimated fetal weight in the validation subsample of the MOMI

TABLE 1. Descriptive Statistics for 24,670 Pregnancies (38,206 Ultrasound Visits) at the Magee-Womens Hospital, University of Pittsburgh Medical Center, Pittsburgh, PA, 1999–2013, and 240 Pregnancies in the ODNSR Data from the National Institute of Child Health and Human Development Maternal Fetal Medicine Units Network, 1992

	MOMI			ODNSR		
Characteristics	**Median**	**Q1**	**Q3**	**Median**	**Q1**	**Q3**
EFW	1.86	0.89	2.76	0.71	0.58	0.87
Birthweight	3.27	3.64	3.64	0.71	0.59	0.85
Maternal age	29	24	33	24	21	30
Gestational age at birth	39	38	40	25	23	27
Year of birth	2008	2005	2010	1992	1992	1992

EFW indicates estimated fetal weight; ODNSR, obstetrical determinants of neonatal survival; Q1, first quartile; Q3, third quartile.

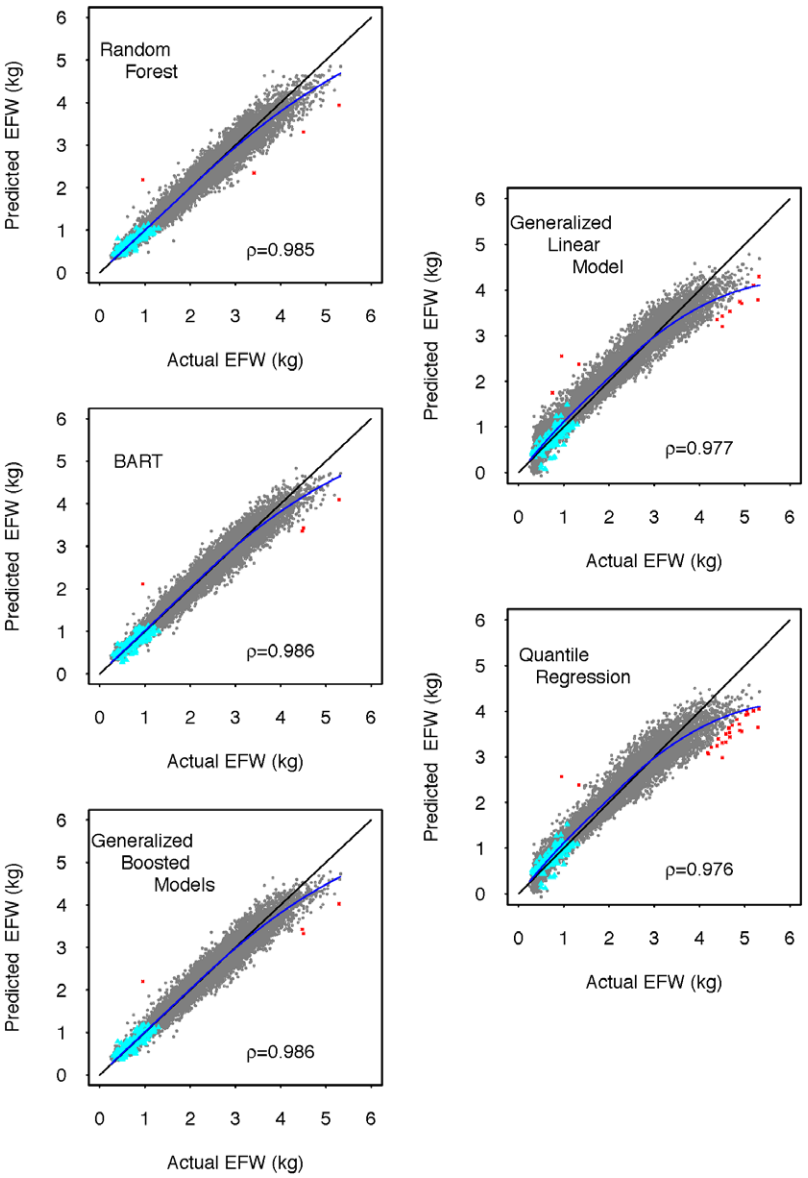

FIGURE 2. Comparison of actual estimated fetal weight to predicted estimated fetal weight for regression-based (right column) and machine learning (left column) approaches in a validation subsample of 8,517 pregnancies (13,354 ultrasound visits) at the Magee-Womens Hospital, University of Pittsburgh Medical Center, Pittsburgh, PA, 2010–2013. Dashed diagonal line represents perfect prediction. Blue lines represent a kernel regression fit to the scatterplots. Red ×'s represent predictions greater than |±1| kg of the actual EFW. Cyan ▲'s represent corresponding predictions from 240 pregnancies in the ODNSR sample from the National Institute of Child Health and Human Development Maternal Fetal Medicine Units Network, 1992.

cohort and the ODNSR data. Figure 2 shows scatterplots comparing the actual and predicted fetal weights from these data with each approach, including a kernel regression smoother demonstrating deviations from the diagonal (perfect prediction). These figures suggest overall better performance of machine learning algorithms compared with standard regression approaches. All methods systematically under-predicted estimated fetal weights greater than 3 kg, with larger deviations observed for the regression-based methods. However, regression-based methods systematically overpredicted estimated fetal weights less than 3 kg (as evidenced by the off-diagonal kernel density smoother), whereas the machine learning algorithms yielded unbiased predictions in this region. Predictions in the ODNSR data suggest better out-of-sample performance for machine learning over regression-based prediction algorithms.

eTable 1; http://links.lww.com/EDE/B301 summarizes the results displayed in Figure 2 using the root mean squared error and median absolute deviation. This table shows that the machine learning algorithms yielded predictions with systematically lower root mean squared error and median absolute deviation values than regression-based approaches. In particular, the lowest median of the absolute difference between predicted and actual fetal weights among the regression-based approaches was the quantile regression method, with median absolute deviation = 137.75 grams (95% CIs: 135.70, 140.29). By contrast, Bayesian additive regression trees yielded the lowest overall median absolute deviation = 88.18 g (95% CIs: 86.25, 90.28), which was not statistically different from generalized boosted model predictions.

A similar pattern was observed in the ODNSR sample predictions (eTable 1; http://links.lww.com/EDE/B301), with

TABLE 2. Risk Ratios and Differences for the Relation Between Smoking Any Time During Pregnancy (Relative to No Smoking) and Small for Gestational Age Birth Defined as Birthweight Less Than the 10th Percentile of a National Birthweight Standard (SGA_B), Estimated Fetal Weight + Birthweight Standard (SGA_{FB}), and Five Different Predicted Fetal Weights + Birthweight (SGA_{PB}) Standard in a Validation Subsample of 8,517 Pregnancies (13,354 Ultrasound Visits) at the Magee-Womens Hospital, University of Pittsburgh Medical Center, Pittsburgh, PA, 2010–2013

Method	Risk Ratio	95% Robust CI		Risk Difference per 100 Live Births	95% Robust CI
SGA_B	3.84	2.70	5.47	3.18	(2.23, 4.17)
SGA_{FB}	1.72	1.51	1.78	13.30	(10.32, 16.28)
SGA_{PB}					
GLM	1.34	1.23	1.44	14.34	(11.26, 17.34)
QR	1.25	1.16	1.34	13.96	(10.87, 17.04)
RF	1.81	1.62	1.88	13.45	(10.87, 15.89)
BART	1.65	1.50	1.81	13.01	(10.12, 16.10)
GBM	1.72	1.61	1.93	13.87	(10.77, 16.83)

BART indicates Bayesian additive regression trees; CI, confidence interval; GBM, generalized boosted models; GLM, generalized linear model; RF, random forest; QR, quantile regression.

one exception: the generalized linear model performed similarly to machine learning algorithms in this small sample of pregnancies, yielding median absolute deviation values that were not statistically distinguishable from the machine-learning approaches. However, the same is not true for the corresponding root mean squared error values, which indicates that the generalized linear model yielded several outlying predictions, as confirmed by visual inspection of Figure 2.

The small for gestational age analysis results varied strongly by choice of method. Table e2; http://links.lww.com/EDE/B301 shows the number of live-births from the MOMI data that were classified as small for gestational age using a gestational age specific population birthweight standard, and an actual or predicted fetal weight + birthweight standard. As displayed in the first two rows, the number of small for gestational age births identified using the birthweight standard was 1,184 of 8,517, lower than the actual fetal + birthweight standard of 1,815. A predicted fetal weight standard from all three machine learning approaches yielded results similar to the actual fetal weight birthweight standards. However, both regression-based predicted fetal weight standards yielded a much higher number of pregnancies identified as small for gestational age.

A similar pattern was observed when we estimated the association between smoking status during pregnancy and the risk of small for gestational birth identified using different standards (Table 2). On the risk ratio scale, using the birthweight standard, smoking anytime during pregnancy increased the risk of small for gestational age birth 3.84-fold (95% robust CI: 2.70, 5.47). However, this risk ratio dropped to 1.65 (95% robust CI: 1.50, 1.81) when using the actual fetal + birthweight standard, no different from those obtained using the machine learning-based predicted fetal + birthweight standard. When evaluated using the regression-based predicted fetal + birthweight standard, the risk ratio dropped even further. The same overall pattern was observed on the risk difference scale in terms of the grouping of magnitudes. However, the direction of these associations was reversed, with the birthweight standard yielding the lowest overall association, and the regression-based predicted fetal + birthweight standard yielding the highest association. This reversal is attributable to the variation across method in the baseline risk of small for gestational age birth. Still, the results from the machine-learning based predicted fetal + birthweight standard were no different than those from the actual fetal + birthweight standard.

Figure 3 shows quantile differences at the 25th, 50th, and 75th percentiles of the distribution of estimated fetal weight across gestation between smokers and nonsmokers. Compared with the smoothed quantile differences in the actual fetal weight + birthweight data (upper left panel, grey background), both generalized linear model and quantile regression model predictions yielded data that showed no differences between smokers and nonsmokers at the 25th, 50th, and 75th percentiles. On the other hand, predictions from Bayesian additive regression trees, random forests, and generalized boosted models showed quantile differences that were similar to those obtained in the empirical data. These smoothed quantile differences suggested that: (1) nonsmoker fetal weight + birthweight began to deviate from smoker fetal weight + birthweight at roughly 32 weeks of gestation, and (2) the differences are larger at the 75th percentile of the fetal weight + birthweight distribution, relative to similar differences at the 50th and 25th percentiles.

Sensitivity analyses were conducted in which development algorithms were trained in various low-risk subsets of the MOMI sample. These results suggest no material changes in these results. Details are provided in the eAppendix; http://links.lww.com/EDE/B301.

DISCUSSION

The absence of fetal weight information in population birth records poses several challenges in reproductive

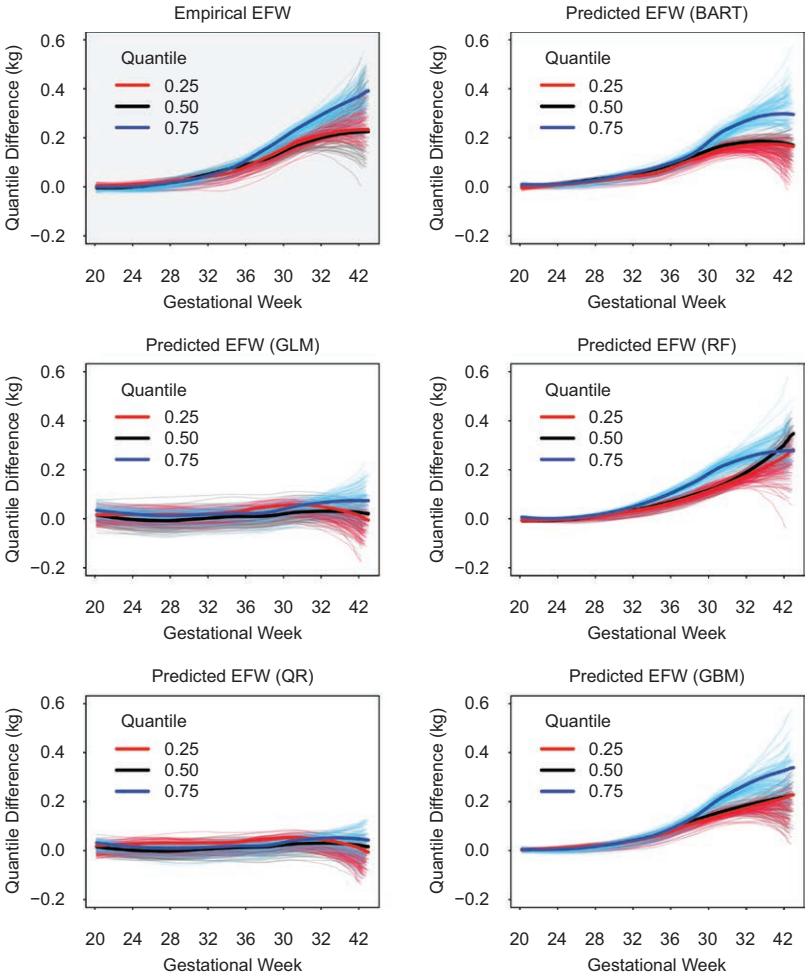

FIGURE 3. Quantile differences in estimated fetal weight across gestational weeks between smokers and nonsmokers at the 25th, 50th, and 75th percentiles in a validation subsample of 8,517 pregnancies (13,354 ultrasound visits) at the Magee-Womens Hospital, University of Pittsburgh Medical Center, Pittsburgh, PA, 2010–2013. Plot with grey background represents quantile differences estimated using actual fetal weight estimates. Plots with white backgrounds represent quantile differences estimated using predicted fetal weight estimates. Translucent lines represent fits from 100 bootstrap resamples.

epidemiology. Our focus was on the use of machine learning algorithms that take as input maternal and infant birth characteristics, and output predicted fetal weights. We evaluated how predicted fetal weights compare with actual fetal weights by (1) estimating the correlation between predicted and actual fetal weights; (2) comparing the gestational week–specific quantiles of actual and predicted fetal weights; and (3) examining the relation between smoking status and small-for-gestational-age birth, with the latter defined using actual and predicted data. We found that machine learning algorithms generated fetal weight predictions that closely aligned with actual fetal weight estimates in a sample of future observations from the same population, and in a sample of high-risk pregnancies from an independent population. Furthermore, our analysis of the relation between smoking status and small for gestational age birth birth when the latter was defined using the combined distributions of birthweight and actual fetal weight were no different from those obtained when using the distributions of birthweight and predicted fetal weight. However, results differed substantially when using the 10th percentile of a national birthweight standard.[18] Overall, these results suggest that machine learning algorithms may be preferable to commonly used population birthweight standards to recover missing fetal weight information in population birth records for the purposes of epidemiologic research.

To evaluate practices commonly employed in reproductive epidemiology, we used a single birthweight standard to define small for gestational age.[18] Other standards are available, including international standards derived from the INTERGROWTH-21st Project.[25] We did not evaluate the INTERGROWTH birthweight standard for three reasons: (1) the standards are only available for gestational weeks 33–42, and thus could not be used to define small for gestational age for a large number of gestational weeks in our sample; (2) the 10th percentile birthweight values for both boys and girls did not meaningfully differ from the national referent values that we used; (3) previous work has shown that the INTERGROWTH standard underestimates the number of small for gestational age infants relative to a more local standard.[26]

The machine learning algorithms we employed are generally known as ensemble learning methods.[8] These methods combine (via bagging or boosting) many weak predictors with good local performance to yield an ensemble predictor with strong global performance. We could have additionally combined

each ensemble learner and regression-based algorithm into a "SuperLearner" algorithm,[27] comprised of a family of weighted combinations of each, with weights chosen to maximize performance. This would have enabled us to automate the evaluation of a much larger set of tuning parameters and model forms.

Our results should be considered in light of certain limitations. First, our approach effectively amounts to imputing missing fetal weight data by combining information available at birth in a given dataset with algorithms trained on data from an external population. As with typical imputation approaches, quantifying the variance of the parameter estimates from a given model should incorporate uncertainty that results from predicting unknown quantities.[30] In the case of nonparametric data mining and machine learning algorithms, this may be particularly challenging due to what is known as the curse of dimensionality,[31] and the fact that important regularity conditions do not hold.[32] While in our particular case, use of machine learning methods yielded results (including confidence intervals) similar to the true relation (SGA_{FB} in Table 2) we do not suggest that bootstrapping is a generally valid way of obtaining confidence intervals for data-adaptive techniques.

The regression-based and machine learning–based approaches represent only a set of a larger number of approaches that may, in principle, be used to recover missing fetal weight information. Alternative approaches could involve use of novel probability weights with efficient semiparametric estimators such as targeted minimum loss-based estimation,[28,33] or Bayesian data augmentation approaches.[29] These latter methods may not be subject to the same limitations resulting from the curse of dimensionality or nonregularity conditions, but their potential for addressing missing fetal weight data in reproductive epidemiology is yet to be explored.

While very different overall, the range of covariate values in the ODNSR data were bounded within the ranges of the corresponding covariates in the MOMI data. Less optimal performance may occur if the regression-based and machine learning algorithms are trained on data that do not overlap with data in which predictions are to be generated. In the MOMI data, many of the ultrasounds were clinically indicated, suggesting covariate distributions may not overlap with those in more general population birth registries. Such "off-support" predictions will likely be as problematic as in the inferential case,[34] and thus caution is warranted.

Despite these limitations, we have shown that machine learning algorithms trained on an external dataset hold promise in recovering missing fetal weight information. Such missing information has long been problematic in reproductive and perinatal epidemiology. While several proposals have been made to resolve the problem of missing fetal weight data,[3–7] few are general enough to apply in a wide array of empirical settings. Use of machine learning algorithms only requires the existence of a common set of covariates in both the training and prediction datasets. Additionally, we have shown that using information routinely available in birth cohorts leads to relatively accurate predictions.

REFERENCES

1. Pallotto EK, Kilbride HW. Perinatal outcome and later implications of intrauterine growth restriction. *Clin Obstet Gynecol*. 2006;49:257–269.
2. Hutcheon JA, Platt RW. The missing data problem in birth weight percentiles and thresholds for "small-for-gestational-age". *Am J Epidemiol*. 2008;167:786–792.
3. Hadlock FP, Harrist RB, Martinez-Poyer J. In utero analysis of fetal growth: a sonographic weight standard. *Radiology*. 1991;181:129–133.
4. Marsál K, Persson PH, Larsen T, et al. Intrauterine growth curves based on ultrasonically estimated foetal weights. *Acta Paediatr*. 1996;85:843–848.
5. Johnsen SL, Rasmussen S, Wilsgaard T, et al. Longitudinal reference ranges for estimated fetal weight. *Acta Obstet Gynecol Scand*. 2006;85:286–297.
6. Mongelli M, Biswas A. A fetal growth standard derived from multiple modalities. *Early Hum Dev*. 2001;60:171–177.
7. Bernstein IM, Mohs G, Rucquoi M, et al. Case for hybrid "fetal growth curves": a population-based estimation of normal fetal size across gestational age. *J Matern Fetal Med*. 1996;5:124–127.
8. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer. 2009.
9. Bottoms SF, Paul RH, Iams JD, et al. Obstetric determinants of neonatal survival: influence of willingness to perform cesarean delivery on survival of extremely low-birth-weight infants. National Institute of Child Health and Human Development Network of Maternal-Fetal Medicine Units. *Am J Obstet Gynecol*. 1997;176:960–966.
10. Hadlock FP, Harrist RB, Carpenter RJ, et al. Sonographic estimation of fetal weight. The value of femur length in addition to head and abdomen measurements. *Radiology*. 1984;150:535–540.
11. Rencher AC. *Linear Models in Statistics*. New York: Wiley. 2000.
12. Koenker R. *Quantile Regression*. Cambridge, NY: Cambridge University Press, 2005.
13. Breiman L. Random forests. *Machine Learning*. 2001;45:5–32.
14. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. *Ann Appl Stat*. 2010;4:266–298.
15. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Statist*. 2001;29:1189–1232.
16. Ridgeway G. The state of boosting. *Comput Sci Stat*. 1999;31:172–181.
17. Berk RA. *Statistical Learning from a Regression Perspective*. New York, NY: Springer, 2008.
18. Talge NM, Mudd LM, Sikorskii A, et al. United States birth weight reference corrected for implausible gestational age estimates. *Pediatrics*. 2014;133:844–853.
19. Nelder JA, Wedderburn RWM. Generalized linear models. *JRSS-A*. 1972;135:370–384.
20. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 2015. Available at: https://www.R-project.org/. Accessed July, 2016.
21. Koenker R. *quantreg: Quantile Regression*. 2016. Available at: https://CRAN.R-project.org/package=quantreg. R package version 5.21. Accessed July, 2016.
22. Meinshausen N, Schiesser L. *quantregForest: Quantile Regression Forests*. 2015. Available at: https://CRAN.R-project.org/package=quantregForest. R package version 1.1.
23. Kapelner A, Bleich J. bartMachine: Machine Learning with Bayesian Additive Regression Trees. *Journal of Statistical Software*. 2016;70:1–40.
24. Greg Ridgeway with contributions from others. *gbm: Generalized Boosted Regression Models*. 2015. Available at: https://CRAN.R-project.org/package=gbm. R package version 2.1.1. Accessed July, 2016.
25. Villar J, Cheikh Ismail L, Victora CG, et al; International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st). International standards for newborn weight, length, and head circumference by gestational age and sex: the Newborn Cross-Sectional Study of the INTERGROWTH-21st Project. *Lancet*. 2014;384:857–868.
26. Anderson NH, Sadler LC, McKinlay CJD, et al. INTERGROWTH-21st vs customized birthweight standards for identification of perinatal mortality and morbidity. *Am J Obstet Gynecol*. 2016;214:509.e1–509.e7.

27. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6:Article25.
28. van der Laan MJ. Targeted maximum likelihood learning. *Int J Biostat*. 2006;2:Article 11.
29. Greenland S, Christensen R. Data augmentation priors for Bayesian and semi-Bayes analyses of conditional-logistic and proportional-hazards regression. *Stat Med*. 2001;20:2421–2428.
30. Little RJA. *Statistical Analysis with Missing Data. Wiley Series in Probability and Statistics*. Hoboken, N.J.: Wiley, 2002.
31. Robins JM, Ritov Y. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Stat Med*. 1997;16:285–319.
32. Horowitz JL. The bootstrap. In: Heckman JJ and Leamer E, eds. *Handbook of Econometrics*, Volume 5. Hoboken, NJ: Elsevier, 2001.
33. Rudolph KE, van der Laan MJ. Robust estimation of encouragement design intervention effects transported across sites. *JRSS-B*. 2017; 79:1509–1525.
34. Manski CF. Identification problems in the social sciences. *Sociol Methodol*. 1993;23:1–56.