Machine Learning Methods for Estimating Gestational Age at Birth from Electronic Health Records

Cosmin A. Bejan, PhD[1], Amelie Pham, MD[1], Leena Choi, PhD[1], Sarah Osmundson, MD[1], S. Trent Rosenbloom, MD MPH[1], Elizabeth J. Phillips, MD[1]

[1]Vanderbilt University Medical Center, Nashville, TN, USA

Introduction

Due to the strict regulations imposed for the participation of pregnant patients in drug development trials, leveraging real-world healthcare data such as electronic health records (EHRs) has emerged as an attractive strategy for pharmacoepidemiologic studies investigating the effect of maternal drug exposures on various clinical outcomes. Accurate identification of timing and duration of drug exposures in pregnancy is critical to determine the fetal effects by stage of development. However, precise pregnancy dating information can be difficult to obtain from the EHR data. To minimize drug exposure misclassification and reduce the risk of potential false discoveries in pharmacoepidemiologic studies of drug safety during pregnancy, we evaluate machine learning methods for gestational age (GA) estimation at birth using EHR data.

Methods

The study (IRB# 231322) included all mother-child dyads from the Research Derivative (RD), a repository of identified EHR data restructured for research at Vanderbilt University Medical Center, a large academic medical center that offers primary and specialty referral care for obstetrics and newborns. Regular expressions were developed to convert the clinical GA expressions available in the RD (eg, '*40 4/7 weeks'*, '*37.4 wks'*, '*38 w 2d*') into numeric format representing GA values in days. Each valid GA value was assigned to a mother-child dyad if its corresponding delivery date overlapped with the child's documented date of birth, within a range of ±2 days. The dyads corresponding to (1) infants with unknown sex or (2) mothers without any EHR data recorded during pregnancy or with unknown race/ethnicity were excluded from the study. The resulting mother-child dyad dataset was randomly split into training (80%) and test (20%) sets. For GA estimation, 3 simple models using baseline values (40 weeks, mean and median GA values computed from the training set) and 3 machine learning models (linear regression, random forest, and gradient boosting) were implemented. The features used to train the machine learning models included maternal age at delivery, maternal race and ethnicity, infant sex, and International Classification of Diseases (ICD) codes for preterm (644.2*, O42.0*, O42.1*, O42.9*, O60.1*), term (645.1*, 649.8*, 650, O48.0, O60.2*, O75.82, O80), and post-term (645.0*, 645.2*, O48.1) deliveries. The models' performance was evaluated using mean squared error (MSE), mean absolute error (MAE), and the coefficient of determination (R^2).

Results

79,326 mother-child dyads of 59,900 pregnant patients (56.7% White, 16.4% Hispanic, 16.3% Black, 4.4% Asian, and 6.2% Other/Unknown) were identified in the RD. Of these, 70,758 (89.2%) pairs were assigned to valid GA values. The remaining 8,568 (10.8%) pairs (a) did not have any GA data in the corresponding records (N=5,710), (b) did not have GA mentions within the delivery date range (N=2,113), or (c) had invalid GA expressions such as '*unknown*' and '*NA weeks*' (N=745). After exclusion criteria, the final dataset consisted of 65,467 pairs with valid GA values. The mean GA value at birth in the training set was 269.3 days while the median was 274 days (interquartile range, 265–280 days). The evaluation for GA estimation from the simple models using baseline values and machine learning models over the test set is shown in Table 1 and Figure 1. All machine learning models show consistent improvement over the baseline models.

Table 1 GA evaluation on the test set.

Model	MSE	MAE	R^2
40 weeks	519.12	13.57	-0.3029
Mean	398.53	12.89	-0.0003
Median	423.28	11.91	-0.0624
Linear Regression	211.98	8.77	0.4679
Random Forest	213.63	8.78	0.4638
Gradient Boosting	207.53	8.63	0.4791

MSE Mean Squared Error, MAE Mean Absolute Error, R^2 coeff. of determination.

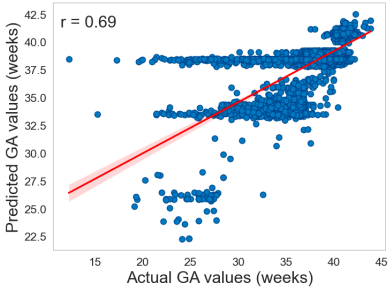

Figure 1 Scatter plot of actual vs. predicted GA values by gradient boosting.

Discussion and Conclusions

The preliminary results of this study show that machine learning models have the potential to improve GA estimation at birth using maternal demographic and clinical information associated with pregnancy care and delivery-specific ICD codes. Future directions for improving the GA estimation at birth include further refinement of the machine learning models using additional pregnancy-related ICD codes and Current Procedural Terminology (CPT) codes, as well as hyperparameter tuning of the machine learning models with k-fold cross validation.