Title: Development of Patient-Level Prediction Models for Preterm Birth.

PRESENTER: **Jill Hardin**

INTRO: Preterm birth, i.e., birth between 20 and 37 weeks of gestation, is the leading cause of disease burden for newborns and children in the world. Every year, an estimated 15 million babies are born preterm. Preterm birth complications are the leading cause of death among children under 5 years of age, responsible for approximately 1 million deaths in 2015. The preterm birth rate is rising in high income countries. There are three sub-categories of preterm birth, based on gestational age: extremely preterm (less than 28 weeks); very preterm (28 to 32 weeks); and moderate to late preterm (32 to 37 weeks). The causes of preterm birth remain unclear but the following maternal conditions are reported to be risk factors for spontaneous preterm labor and birth: 1) socioeconomic determinants including low education and income; 2) maternal body mass index (BMI), diabetes mellitus, hypertensive disorder; and 3) obstetric determinants including infection, parity, vaginal bleeding during pregnancy, history of abortion, cesarean section, placenta abruption, placenta previa and prior preterm birth. A prior prediction study attempted to inform this question, and showed modest performance, but had substantial methodological limitations. Prior studies have employed relative risk regression models, used a small subset of variables and assumed others were constant. Our study uses large observational datasets containing thousands of preterm birth outcomes and will be used to develop highly discriminative models using predictors that are recorded prior to pregnancy. This purpose of this study is to develop extreme preterm birth prediction models using three healthcare claim databases and Lasso logistic regression models. These models would be used at a time a women becomes pregnant to identify those women who may benefit from interventions to reduce negative preterm birth outcomes.

METHODS: The following databases were used in this study:

- IBM MarketScan® Commercial Claims and Encounters (IBM CCAE)
- Multi-State Medicaid Database (IBM MDCD)
- Optum© De-Identified Clinformatics® Data Mart Database – Socio-Economic Status (Optum SES)
- We included all patients aged > 18 years using data from 2006 to 2017.

NOTE: all databases were transformed to the Observational Medical Outcomes Partnership Common Data Model. The use of Optum and CCAE was reviewed by the New England Institutional Review Board and was determined to be exempt from broad Institutional Review Board approval as this project did not involve human subject research.

- Our study builds upon a previously developed algorithm inferring pregnancy episodes and their outcomes, which was validated in three US based claims databases and the UK Clinical Practice Research Datalink (CPRD).

Target cohort included:

- Subjects with first observed pregnancies (first) (excluding all pregnancy related concepts all days prior to first pregnancy)

Outcome cohort included: pregnant subjects with live birth between 20 and 28 weeks gestation (extreme preterm).

Prediction was performed using the PatientLevelPrediction package and LASSO logistic regression. Thousands of binary candidate predictors were created using records prior to conception and included demographics, condition occurrences, drug exposures, procedures performed, measurements, and visit counts in three time periods (30, 365, and all days prior to index pregnancy). Model discrimination was assessed by calculating the area under the Receiver Operating Characteristic curve (AUC) and model calibration was assessed using the calibration slope. Models were developed in three US databases and the highest performing model was externally validated in the other two databases. In this way we get information about the best performing models and also information on the robustness and transportability.

We can predict extreme preterm birth amongst newly pregnant women using administrative claims data.

Take a picture to download additional details

RESULTS: The AUCs ranged between 0.663-0.696 and the calibration slopes were consistently around 1.
Several previously identified predictors were included in our final models. The final model included 199 predictors.
The model with the highest internal AUC (0.696 in Optum) was applied to IBM CCAE and MDCD (AUROC: 0.666 and 0.553).

Table 1. The study populations and internal AUROC in three claims databases

Database	Internal AUC	Population Size (N)	Outcome Count (N)	Internal Calibration Slope	Incidence
IBM CCAE	0.693	651,384	7,791	1.162	1.196069
IBM MDCD	0.663	151,069	3,939	1.315	2.60742
OPTUM SES	0.696	298,216	3,427	1.221	1.149

Figure 1. ROC Plot – Optum SES internal AUC=0.696

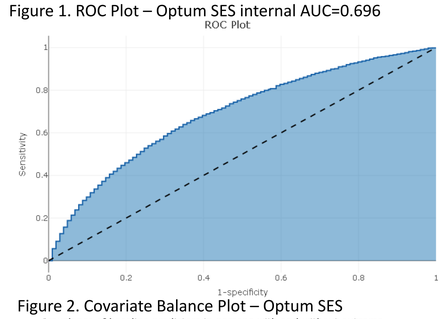

Figure 2. Covariate Balance Plot – Optum SES

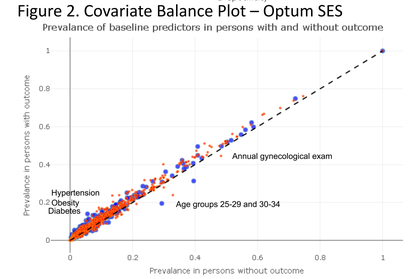

Table 2. External validation results

Validation Database	AUROC	Total Population Size (N)	Outcome count in test population (N)	Outcome % incidence in test
IBM CCAE	0.666	868,512	10,388	1.20%
IBM MDCD	0.553	201,425	5,252	2.61%

NOTE: Model validated uses development database = Optum SES; internal AUC = 0.696.

Please see the QR code for our shiny app with additional model details

Important Learnings:

1. The predictors identified in the final model included several previously suspected as positively associated: diabetes, obesity, age, and hypertension and several predictors negatively associated with the outcome: annual gynecological exam and multivitamin use.
2. When the target cohort did not exclude broad pregnancy related markers prior to index the final model included the index pregnancy markers model and revealed variables that shouldn't have been possible, but were present due to data inconsistencies with defining pregnancy start. Therefore we learned inferring 'pregnancy start' from data markers can result in noisy models that can be difficult to interpret.
3. The value in the model is in classifying subjects with certainty in the extremes (those with high and low probability) as having or not having the outcome; which when used in clinical setting could identify those women at highest and lowest risk.
4. Setting a risk threshold of 5%, we can prioritize review of 1% of the total pregnancies and find within that 'higher risk' group that we have a positive predictive value of 9% and can identify 4.5% of the preterm births. So, reviewing 1 in 100 pregnancies may help us anticipate 1 in 20 preterm births.

Jill H. Hardin, MBA, MS, PhD[1,3], Patrick B. Ryan, PhD[1,2,3], Jenna M. Reps, PhD[1,3]
[1] Janssen Research and Development, Raritan, NJ, USA; [2] Columbia University, New York, NY, USA; [3]Observational Health Data Sciences and Informatics (OHDSI), New York, NY