|  |  |  |
| --- | --- | --- |
| **eMERGE Network: External Collaborator Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT365 | |
| **Submission Date** | 10/31/2019 | |
| **Project Title** | Predicting Gestational Age from Electronic Health Records and Linked Genetic Biobanks using Boosted Decision Trees | |
| **Tentative Lead Investigator** *(first author)* | Abin Abraham | |
| **Tentative Senior Author**  *(last author)* | Tony Capra | |
| **eMERGE Site Sponsor & Contact** | VGER | |
| **All Other Authors** | Cosmin Bejan, Lea K. Davis, Peter Straub, Digna R. Velez-Edwards, Wei-qi Wei | |
| **Sites Participating** | We would like participation from any/all eMERGE sites. | |
| **Background / Significance** | Despite advances in neonatal medicine, preterm birth (PTB) leads to significant morbidity and mortality across the world. PTB presents with multiple comorbidities and the underlying mechanisms remains poorly understood. Furthermore, there are limited biomarkers that accurately predict PTB and quantify risk to inform clinical care. EHRs linked to genetic biobanks provide longitudinal and deep phenotyping of patients that have been under-utilized for studying adverse pregnancy outcomes. To our knowledge, there are no algorithms that can predict PTB risk early in pregnancy using features derived from EHRs. Additionally, new GWAS studies in multiple cohorts now enable us to model genetic risk for PTB. We will systematically evaluate the potential of diverse types of EHR data integrated with maternal genetics to predict PTB at clinically relevant timepoints.  The ability to predict the risk for PTB using genetic and existing clinical data is non-invasive, can be pre-calculated, and has the potential to improve medical management. This work takes important steps towards generating actionable models derived from EHRs to personalize treatment and reveal the major drivers PTB. | |
| **Outline of Project** | 1. Use timestamped ICD-9 and CPT codes to identify delivery date and type from EHR for women. Validated above algorithm with chart review. (Completed at Vanderbilt)  2. Develop machine learning models to predict preterm birth using total counts of ICD-9 and CPT codes.  3. Evaluate model performance when adding demographics, genetic risk, and clinical features.  4. Evaluate best performing model at clinically relevant time points.  5. Identify and interpret top features driving model performance | |
| **Desired Data - Common Variables\***  *(Available from the CC)* | Demographics  ICD9/10 codes indicating delivery with timestamps  CPT codes indicating delivery with timestamps Estimated Gestational Age  Phecodes  BMI | Common Variable Labs  Common Variable Meds  Other: Case/Control status on Phase I and Phase II phenotypes |
| **Other Desired Data *(Available from participating sites)*** | *All ICD-9 codes with timestamps*  *All CPT codes with timestamps*  *Obstetric Notes (optional) with timestamps*  *Genotyped Data (optional) with timestamps*  *Clinical Labs (optional) with timestamps* | |
| **Desired Genetic Data** | eMERGE I-III Merged set (HRC imputed, GWAS)  eMERGE PGx/PGRNseq data set  eMERGEseq data set (Phase III)  eMERGE Whole Genome sequencing data set  eMERGE Exome chip data set  eMERGE Whole Exome sequencing data set  Other (not listed above): | |
| **Does project pertain to an existing eMERGE Phenotype?** | Yes, if so please list  No | |
| **Planned Statistical Analyses** | Boosted decision tree will trained and optimized to predict preterm birth. Performance will be evaluated on a held-out dataset using ROC-AUC, PR-AUC, and F1. Model calibration for risk estimates will be evaluated using Brier Scores. The models will be developed at Vanderbilt. Reproducibility of the models will be assessed using external datasets from eMERGE institutions. | |
| **Ethical Considerations** | None | |
| **Available Funding or Resources** |  | |
| **Target Journal** | PLOS Medicine | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | - Analysis at Vanderbilt by Oct 2019 [Done]  - Reproducibility analysis by Jan 2019  - 1st draft of manuscript by Feb 2020  - 2nd draft of manuscript by March, 2020  Submit by April 2020 | |

**\*Common Variables available across all datasets:**

* Demographics: sex, year of birth, decade of birth, race, ethnicity
* Codes: (repeated values & age at event): ICD, CPT, Phecodes
* BMI: (repeated value & age at event) height, weight, BMI
* Labs: (lab name, repeated lab value & age at event) Serum total cholesterol, LDL, HDL, Triglycerides, Glucose fasting/non-fasting/unknown, & White Blood Cell count
* Medications: (medication name, repeated, & age at event) Cerivastatin sodium, Rosuvastatin, Simvastatin, Fluvastatin, Pravastatin, Lovastatin, Atorvastatin, & Pitavastatin
* Other: Case/Control status on Phase I and Phase II phenotype: only on GWAS dataset participants