Estimation of Postpartum Depression Risk from Electronic Health Records Using Machine Learning

Guy Amit[1], Irena Girshovitz[1], Karni Marcus[1], Pinchas Akiva[1], Vered Bar[2]

[1] KI Research Institute, Kfar Malal, Israel ; [2] Sheba Medical Center, Women Mental Health, Ramat Gan, Israel

Background

- Postpartum depression is one of the most common complications of pregnancy and childbirth, with estimated prevalence of 10-15%
- PPD risk is associated with biological, phycological and sociodemographic factors
- There are no quantitative tools for risk estimation, and screening is typically based on symptom questionnaires, such as the Edinburgh postnatal depression scale (EPDS)
- Early identification of PPD risk during or before pregnancy may enable effective early intervention
- **Suggested solution: A predictive model that uses electronic health records (EHR) for learning to identify patients at risk**
  - May enable early identification of patients at risk
  - May be used to augment current screening tools

Methods

- **Dataset**: Primary care EHR data from IQVIA Medical Research Data (IMRD), incorporating data from The Health Improvement Network (THIN, a Cegedim database). The dataset contains records of over 18M patients, approximately 5% of UK population
- **Population**: 266,544 women between the ages 18-45 who had their first live birth between 2000-2017

Patient characteristic	Value
N	266544
Age (years)	30.0±5.8
Deprivation index (quantile)	3.03±1.3
Pre-pregnancy BMI	25.0±5.4
Cesarean section	51151 (19.2%)
Smoking	64778 (24.3%)
History of depression	17384 (6.5%)

- **PPD outcome definition**: Record of depression diagnosis or new treatment for depression during the 12-month period after childbirth
- **Predictor variables**: (1) Demographic, socio-economic and personal measures;(2) Medical diagnoses during and prior-to pregnancy (3) Labor complications and infant-related measures; (4) Drug prescriptions during and prior-to pregnancy (6) Healthcare utilization measures
- **Machine learning model**: Data was split into training, testing and holdout sets. Prediction models were trained using gradient tree boosting algorithm.

Results

- The prevalence of PPD was 13.4%
- The EHR-based prediction model achieved area under the ROC curve (AUC) of 0.74 and sensitivity of 0.56 at specificity 0.8
- Early prediction, using only pre-pregnancy variables was just minorly less accurate (AUC=0.73)
- Combining the EHR-based prediction score with the EPDS score improved the AUC from 0.805 (EPDS alone) to 0.844 (combined, P<0.001) and the sensitivity from 0.72 to 0.76, at specificity 0.8

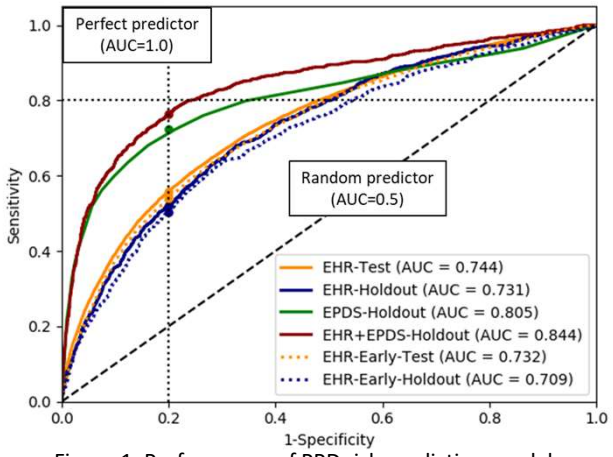

Figure 1: Performance of PPD risk prediction models

- The major variables in the model were history of antidepressant drugs, total number of any drug prescriptions, previous diagnoses of abdominal pain, premenstrual syndrome, depression and anxiety. Age, BMI, smoking status and deprivation index were also contributing variables.
- A prediction model without any mental health variables achieved fair AUC of 0.70.

Conclusions

- PPD risk can be predicted from EHR data with good accuracy, even before or during pregnancy
- EHR-based prediction can be combined with EPDS to improve the accuracy of PPD screening
- Integrating and quantifying the PPD risk factors using an automatic and objective predictive tool may support the clinical decision making and facilitate timely interventions and improved outcomes