Title: Predictive models of fecal microbial biomarkers for obesity trajectories in preschool children

Authors: Xin Zhao[1], Anita Kozyrskyj[1]

[1] Department of Pediatrics, University of Alberta, Edmonton, AB, Canada

Abstract

This study was aimed to devise machine learning models to predict the early-onset and sustained trajectory for young children on 16S rRNA amplicon-based gut microbiota profiles of infants aged three months. A multi-center, longitudinal cohort CHILD study provided data points of 1943 children in Toronto, Vancouver, Edmonton, and Winnipeg in Canada. We combined OTU independently sequenced by two laboratories into a final sample size of 2450 stool samples, with 507 repeat measures. The outcome variable, the high-risk and low-risk groups, were derived from the initial BMIz trajectory identified by the growth trajectory model.

We applied advanced machine learning algorithms, such as random forest, XGBoost, generalized linear mixed models to the gut microbiota dataset. Furthermore, we built a rigorous modeling pipeline to address real-world analytic issues, including repeat measures, imbalanced class classification, and the batch effect. The SMOTE algorithm during cross-validation was used to combat the class imbalance problem. The stratified, repeated, 5-fold cross-validation procedure was used to train and tune models. Generalized linear mixed models (GLMM) outperformed other machine learning models, achieving AUC-ROC 0.84 (90% bootstrap CI 0.70, 0.94) on the test set. The optimal GLMM model possessed a sensitivity of 0.9 and specificity of 0.6 under the best threshold. In addition, the microbiome-based machine learning models identified five microbial biomarkers at the genus level, including *Osillospira*, *Rikenellaceae genus*, *Blautia*, *Phascolarctobacterium*, and *Haemophilus*.

Our models manifest good predictive power when compared to existing microbiome-based ML models for childhood obesity prediction. Our study provides a robust microbiome-based model to facilitate precision interventions for clinicians and identify a group of biomarkers for the diagnosis and prognosis of childhood obesity.