2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)

Understanding Early Childhood Obesity via Interpretation of Machine Learning Model Predictions

Xueqin Pang
Department of Biomedical and Health Informatics
Children's Hopspital of Philadelphia
Philadelphia, USA
pangx@email.chop.edu

Christopher B. Forrest
Applied Clinical Research Center
Children's Hospital of Philadelphia
Philadelphia, USA
forrestc@email.chop.edu

Félice Lê-Scherban
Department of Epidemiology and Biostatistics
Dornsife School of Public Health, Drexel University
Philadelphia, USA
f.lescherban@drexel.edu

Aaron J. Masino
Department of Anesthesiology & Critical Care Medicine
Perelman School of Medicine
University of Pennsylvania
Philadelphia, United States
masinoa@email.chop.edu

***Abstract*—Obesity, as an independent risk factor for increased morbidity and mortality throughout the lifecycle, is a major health issue in the United States. Pediatric obesity is a strong risk factor for adult obesity, as it tends to be stable and tracks into adulthood. Therefore, prevention of childhood obesity is urgently required for reduction in obesity prevalence and obesity related comorbidities. In this paper, the general pediatric obesity development pattern and the onset time period of early childhood obesity was identified via analysis of approximately 11 million pediatric clinical encounters of 860,510 unique individuals. XGBoost model was developed to predict at age 2 years if individuals would develop obesity in early childhood. The model is generalized to both males and females, and achieved an AUC of 81% (± 0.1%). Obesity associated risk factors were further analyzed via interpretation of the XGBoost model predictions. Besides known predictive factors such as weight, height, race, and ethnicity, new factors such as body temperature and respiratory rate were also identified. As body temperature and respiratory rate are related to human metabolism, novel physiologic mechanisms that cause these associations might be discovered in future research. We decomposed model recall to different age ranges when obesity incidence occurred. The model recall for individuals with obesity incidence between 24–36 months was 97.63%, while recall for obesity incidence between 72–84 months was 48.96%, suggesting obesity is less predictable further in the future. Since obesity is largely affected by evolving factors such as life style, diet, and living environment, it is possible that obesity prevention may be achieved via changes in adjustable factors.**

Keywords— pediatric obesity, obesity development pattern, machine learning, model interpretation, obesity associated factors

I. INTRODUCTION

Obesity is a major health issue in the United States. In 2015-2016, obesity prevalence reached 18.5% and 39.8% for individuals aged 2-19 years and 20-years and over, respectively [1]. Obesity prevalence for youth increased from 13.9% in 1999-2000 to 18.5% in 2015-2016, while the prevalence for adults rose from 30.5% to 39.6% over the same time period. Obesity is associated with significant physical and behavioral health concerns, such as heart and vascular diseases [2], type 2 diabetes [3], hypertension [4], and depression [5]. Furthermore, obesity raises the medical care costs of obese adults by an average of $3,429 [6]. From 2001 to 2015, the US national medical expenditures devoted to treating obesity related illness in adults rose from 6.13% to 7.91% (29% increase) [7]. Obesity tends to persist from childhood to adulthood [8] and indeed pediatric obesity is a strong risk factor of adult obesity [9]. Therefore, a reduction in future obesity prevalence and obesity-related comorbidities may be achieved with interventions directed at prevention in early childhood [10].

Machine learning models trained on electronic health records (EHRs) have been widely researched for the prediction and analysis of numerous health conditions [11]–[13]. However, there has been only limited research on machine learning methods to predict childhood obesity, especially in the early childhood period (<2 years of age). Dugan et al. compared several tree based models, a Naïve Bayes, and Bayes model using 167 clinical variables collected for 7,519 subjects aged 2-10 years [14]. The authors reported that the ID3 model provided the best performance (85% accuracy, and 89% sensitivity). However, these results may be overly optimistic as the reported study did not indicate that evaluation was on a held-out test set, suggesting that they are either for the training data or from cross validation. Hammond et al. developed separate machine learning models for boys and girls to predict obesity at age 5-years using EHRs of 3,449 patients from birth to age 2 years [15]. Their best performing models were able to predict obesity with an AUC of 81.7% for girls and 76.1% for boys. We utilized EHR data from 860,510 patients and 11,194,579 healthcare encounters to develop multiple machine learning models to predict future obesity incidence between 2-7 years of age given clinical information from birth to age 2 years (details on data processing, model optimization, and statistical comparisons further discussed in another paper [16]). Our best model (generalized to both genders) utilized the XGBoost algorithm and achieved a competitive AUC of 81% (± 0.1%).

Although machine learning models appear able to predict future childhood obesity incidence with good performance, to our knowledge, no research has been presented that explains the machine learning model predictions in detail. In general, although machine learning models have been successfully used to predict future health outcomes, the ability to explain model predictions, especially from complex models, is largely limited [17], [18]. As a result, clinical applications often rely on simpler interpretable models (e.g. linear models), possibly at the expense of decreased performance. This tradeoff between model performance and interpretability may be more acute with the availability of big data, where complex models often offer significant performance gains.

978-1-7281-4550-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMLA.2019.00235