Reliable prediction of childhood obesity using only routinely collected EHRs is possible

Mehak Gupta mehakg@smu.edu
Computer Science, Southern Methodist University, Dallas, TX, USA

Thao-Ly T. Phan thaoly.phan@nemours.org
Nemours Children's Health, Wilmington, DE, USA

Daniel Eckrich daniel.eckrich@nemours.org
Nemours Children's Health, Wilmington, DE, USA

H. Timothy Bunnell tim.bunnell@nemours.org
Nemours Children's Health, Wilmington, DE, USA

Rahmatollah Beheshti rbi@udel.edu
Computer & Info. Sciences, and Epidemiology, University of Delaware, Newark, DE, USA

Keywords: Childhood Obesity, Electronic Health Records, Deep Learning

Word Count: 3,932

Funding: Our study was supported by NIH awards, P20GM103446 and P20GM113125.

Disclosure: The authors declare no competing financial interests.

Author Contributions: MG conceived and carried out experiments. TP and RB conceived the experiments and helped in the study design. MG, DE, and HB worked on data extraction, cleaning, and linkage. All authors were involved in writing the paper and had final approval of the submitted and published versions.

Abbreviations

AUPRC - Area Under Precision-Recall (PR) Curve

AUROC - Area Under the Receiver Operating Characteristic

BMI - Body Mass Index

CDC - Centers for Disease Control and Prevention

EHR - Electronic Health Record

US - United States

WFL - Weight-for-length

WHO - World Health Organization

Abstract

Objective: Identifying children at high risk of developing obesity can offer a critical time to change the course of the disease before it establishes. Numerous studies have tried to achieve this; but practical limitations remain, including (i) relying on data not present in routinely available pediatric data (like prenatal data), (ii) focusing on a single age prediction (hence, not tested across ages), and (iii) not achieving good results or adequately validating those.

Methods: A customized sequential deep learning model was built to predict the risk of childhood obesity, focusing especially on capturing the temporal patterns. The model was trained only on routinely collected EHRs, containing a list of features identified by a group of clinical experts, and sourced from 36,191 diverse children aged 0 to 10. The model was evaluated using extensive discrimination, calibration, and utility analysis; and was validated temporally, geographically, and across various subgroups.

Results: Our results are mostly better (and never worse) than all previous studies, including those that focus on single-age predictions or link EHRs to external data. Specifically, the model consistently achieved an area under the curve (AUROC) of above 0.8 (with most cases around 0.9) for predicting obesity within the next 3 years for children 2 to 7. The validation results show the robustness of the model. Furthermore, the most influential predictors of the model match important risk factors of obesity.

Conclusions: Our model is able to predict the risk of obesity for young children using only routinely collected EHR data, greatly facilitating its integration with the periodicity schedule. The model can serve as an objective screening tool to inform prevention efforts, especially by helping with very delicate interactions between providers and families in primary care settings.

1. Introduction

Childhood obesity is a major public health problem across the globe and in the US. Obesity affects almost 1 in 5 and about 14.7 million children and adolescents, for children and adolescents of 2-19 years of age [1]. Despite evidence of potentially modifiable risk factors that contribute to the development of obesity in children, children are often referred for obesity interventions after their obesity is well established and when intervention is less likely to be successful [2]. Prior studies show that less than 30% of children with overweight or obesity are identified by their provider at primary care visits and less than 10% of children

have a diagnoses code for overweight or obesity placed during visits [3, 4, 5, 6]. While pediatric providers frequently use recommended CDC or WHO age and sex-specific BMI charts, they do not often recognize or address weight concerns until children cross overweight and obesity thresholds on these charts [7]. Primary barriers to effective diagnoses and management of obesity in pediatric systems are reported as lack of time, limited resources, and uncertainties about the level of risk [6, 8].

In such a context, reliable predictive models of childhood obesity integrated within the structure of the common well-child visits have the potential to provide timely risk alerts and inform more effective interventions to prevent and control this disease. Clinical predictive models for obesity, designed using artificial intelligence and machine learning (AI/ML) methods, are being considered to understand the contributing factors to the obesity epidemic and inform more effective interventions [9, 10, 11, 12, 13, 14, 15, 16, 17, 18].

One major limitation of existing obesity prediction models is that they use features that, despite their importance, are generally not available in EHRs or are difficult and time-consuming to collect, such as genetic background, parental data, and children's habits. Another limitation is that they mostly focus on predicting the risk of obesity at only one age point (e.g., at age five) and are not flexible in allowing obesity screening for patients at different age points. Additionally, large-scale and rigorously validated predictive models of obesity are rare.

In this work, we aim to demonstrate the possibility of achieving reliable estimates of the future risks of childhood obesity using commonly available (unaugmented) EHR elements across multiple age ranges. As children's weight status in early childhood is highly predictive of weight status into adulthood [19], we focus our study on children up until the age of 10 years. This age range also includes a few years after the "adiposity rebound" patterns observed in children's weight [20, 21]. We rigorously evaluate our model through an extensive series of temporal, geographic, and subgroup validations and explore the most important predictors of obesity in our model. The overarching aim of our study is to offer

a general (practice-agnostic) predictive model that can be integrated within any healthcare system's EHR to reliably predict a child's risk for obesity throughout early childhood to support clinical decision-making and obesity prevention at the point of care.

2. Materials and Methods

2.1 Data source and study cohort selection

The EHR data used in this analysis was extracted from Nemours Children's Health; a large pediatric healthcare network in the United States (US) primarily spanning the states of Delaware, Florida, Maryland, New Jersey, and Pennsylvania. The IRB panel at Nemours approved this research study. We used encounters between January 1, 2002, to December 31, 2019, obtained prior to the pandemic to avoid any short-term influence that the pandemic had on weight gain patterns[22]. A total of 68 029 children with 44 401 791 encounters were selected as having: (i) no evidence of type 1 diabetes, and (ii) no evidence of cancer, and sickle cell disease. Of them, 37 844 children were included having at least two routine infant checkups and at least one checkup between 2 to age 10 with recorded weight and length/height measurements. Of them, 1 653 children were excluded whose year of birth could not be verified, leaving 36 191 children. The final cohort was separately divided for temporal and geographic validation. More details about the Nemours EHR dataset are provided in Supplementary A. Figure 1 shows the steps we took to extract our cohort of 36 191 patients for the model construction.

[Figure 1 about here.]

2.2 Feature selection and data preprocessing

We extracted clinically-relevant features to childhood obesity, by using a data-driven approach coupled with input from clinical experts [23]. Details about this process are presented in Supplementary B. The final EHR features consisted of 138 diagnoses (patient conditions and family-history conditions associated with obesity), 84 ATC3 (Anatomical Therapeu-

tic Chemical Classification Level-3) medication groups, and 51 measurements (vitals and labs).

Following prior studies [24], the EHR features before age 2, were segmented into 5 windows. These 5 sliding windows correspond to 0–4 months, 4–8 months, 8–12 months, 12–18 months, and 18–24 months. All EHR features after age 2 were segmented into 1-year windows due to the lower frequency of medical visits (compared to before 2).

Weight and height data were converted to Weight-for-length (WFL) and BMI% as defined by WHO and CDC, respectively. these values were considered Missing Completely at Random (MCAR) and imputed with carry-forward of the most recent value. WFL and BMI% were categorized into underweight, normal, overweight, and obesity categories. We defined cutoffs for normal weight, overweight, and obesity in accordance with the CDC's standard thresholds of the 85th and 95th percentiles for overweight and obesity, respectively. Because WFL% and BMI% trajectories can be helpful for determining the risk of obesity [25], we also engineered a WFL% change feature that calculates the change in WFL% values between 5 windows defined above for model input. Table 1 shows the characteristics of the final cohort used.

The EHR data also included demographic information about sex (male or female), race (categorized as White, Black, Asian, and other), ethnicity (categorized as Hispanic or non-Hispanic), and payer (categorized as public and private insurance). We also included the Child Opportunity Index (COI) by geolocating the last address of patients before age 2 [26]. COI combines indicators of educational opportunity, health and environment, and socioeconomic opportunity for all US neighborhoods. COI ranges from 1-100, with higher numbers representing neighborhoods with more opportunities. Details about demographic feature generation are provided in Supplementary B.1.

[Table 1 about here.]

A binary representation (not MCAR) was used for all features, where the presence of a value for the variable was captured with a 1, and 0 otherwise. We used quintile binning for all numeric measurement features to divide each measurement into 5 categorical features. In total, we generated 506 binary features: 138 (diagnoses) + 84 (medications) + 51*5 (measurements in 5 percentile bins) + 10 (demographic categories) + 4 (underweight, normal, overweight, and obesity) + 5 (WFL% changes in 5 percentile bins) + 10 (COIs in 10 percentile bins). Please refer to Table S1 in Supplementary B.5 for a detailed list of the final features.

2.3 Descriptive analysis

To study the transition of obesity status, we analyzed the distribution of children with BMI% ≥ 95 at age 3 to 10 that transitioned from different WFL% status (WFL% < 5, $5 \leq$ WFL% < 85, $85 \leq$ WFL% < 95, and WFL% ≥ 95) at age 2. Figure 2 shows that 22% of children with obesity at age 3 had normal weight at infancy and this percentage increases with increasing age, where nearly half (49%) of children with obesity at age 10 had normal weight at infancy.

[Figure 2 about here.]

2.4 Deep learning model training

We adopted our encoder-decoder[1] deep neural network model and the training procedure presented in detail in prior work [27, 18]. The encoder part consists of long short-term memory (LSTM)[2] cells and the decoder consists of a feed-forward network with two fully-connected layers.

To train the model, we first trained the encoder network on the input EHR data from 0 to 2 years. EHR data for every year after age 2 until the age that the prediction was performed

1. An encoder-decoder refers to a design, where an initial network (encoder) receives input and maps (compresses) that to a lower dimension representation. Then a second network (decoder) learns from the encoder output to generate output.
2. LSTMs are a type of (deep) neural network used for processing sequential data types.

at (i.e., the end of the observation window) were then combined (concatenated) with the n-dimensional vector representation derived from the encoder for 0-2 years of EHR data. The number of years of data after age 2 depends on the length of the patient's recorded medical history. Using this design, all the prior data for each patient was used for model training to predict the risk of obesity for the next 3 years. A detailed description of the architecture design and parameter settings is provided in Supplementary C.

3. Experiments and Results

3.1 Setup

We extensively evaluated our predictive model by studying its discrimination power, calibration, and robustness. We validated our model temporally and geographically and studied the model's performance across subpopulations (which can also capture the model's fairness across these groups), using separate test datasets. When using the entire cohort, the data was split with an 80:20 train and test regime, with 5% of the training data as a validation set to fix the best model. Model performance was reported exclusively on the test dataset. The confidence intervals (CI) are calculated using 100 bootstrapped replicates.

Baseline comparison – Similar to prior work [28, 11], we consider only using the last WFL% (below 2 years) and BMI% (above 2 years) available in the observation window as a baseline. This scenario mimics what is generally used in clinical practice for screening children (i.e., only using present data).

Discrimination power – We report prediction performance using Area Under Receiver Operating Characteristic (AUROC), Area Under Precision-Recall Curve (AUPRC), sensitivity, and specificity. We provide sensitivity and specificity at different binary classification thresholds, demonstrating the trade-off between false negatives (patients who develop obesity but are not predicted to develop obesity by the model) and false positives (patients who do not develop obesity but are predicted to develop obesity by the model). Specifically, we

fixed either sensitivity or specificity at 90% and 95% and measured the value of the other metric.

Utility and calibration – We perform decision curve analysis [29] to demonstrate the tradeoffs (costs and benefits) of using the prediction model to analyze the clinical utility at various thresholds. Related to the above approach, we provide calibration metrics to help quantify how well the predicted probabilities of an outcome match the true probabilities observed in the data [30, 31].

Temporal validation – To study our model's robustness across time shifts, we divided our cohort according to the date of birth of the children. The data for 26 786 children who were born between January 1, 2002, and December 31 2009, were included as a training set; and the data for 9 405 children who were born between January 1, 2010, and December 31, 2015, were included in the test set. We report AUROC for the temporal validation in Table 2.

Geographic validation – We additionally validated our model across two different geographic regions in the US. We used 32 848 children seen in Delaware Valley sites, located in the northeastern US, as a training set and 3 343 children seen in Florida, located in the southeastern US, as a test set. We report AUROC for the geographic validation in Table 2.

Robustness across subpopulations – Robustness across subpopulations (group fairness) is evaluated by comparing model AUROCs in the test dataset across five groups determined by the: last WFL% category before age 2 (underweight, normal weight, overweight, obesity), race (Black, White, Asian, Other), ethnicity (Hispanic, Non-Hispanic), sex (female, male), and payer (private, public).

[Figure 3 about here.]

3.2 Evaluation results

We trained our deep learning model using 36 191 eligible children to predict the risk of obesity using 6 different lengths of observation windows from age 0-2 to 0-7. The observed AUROC, sensitivity, specificity, and net benefit for all prediction ages are presented in Table 2. Focusing on a popular setting that is heavily studied in the literature [11, 28, 10], we specifically present discrimination results for 0-2 years observation window and obesity prediction at age 5 in Figure 3. Notably, our model outperforms the presented baseline based on a child's last WFL% with an AUROC and AUPRC of 0.81 and 0.61 compared with 0.71 and 0.56, respectively. Our model also dominates over other strategies (baseline, intervention for all, and intervention for none) across various net benefit threshold probabilities, with significant margins above the 15% threshold[3] probability regime.

Figure 4 compares the AUROC of our model on different subpopulations of children for prediction at age 5, across the 5 groups mentioned in Section 3.1. AUROCs among each group show minimal deviation of 0.04, 0.04, 0.01, 0.001, and 0.008, respectively, showing the robustness of our model across each group.

[Table 2 about here.]

[Figure 4 about here.]

3.3 Analyzing model predictors

We investigated the risk factors identified by the model by analyzing which predictors most attribute to the model's prediction. We used the attention scores [32, 33] obtained from the LSTM cells, as a way to determine which features are given more attention (importance) by the model to predict the output (more details are provided in Supplementary D.1).

We present the mean of the importance scores of all features inside each of the 7 input feature categories (diagnoses, family-history diagnoses, medications, measurements, demographics,

3. Here, a threshold such as 10% means that for every 10 children evaluated for risk of obesity by the predictive model, the health service is willing to risk the cost of mistakenly prescribing 9 children [28].

last obesity status before age 2, and WFL% changes before age 2 for the entire cohort) in Figure 5.A. Clinical diagnoses and WFL% changes were the top 2 important predictor groups. We also present the ranking for the top 20 predictors based on the mean importance scores for the predictors in the entire cohort in Figure 5.B. Previous weight percentile measurements and weight gain patterns were among the top 5 features; in addition to weight-related diagnoses and diagnoses of elevated blood pressure, gastroesophageal reflux disease, developmental delay, hypothyroidism, and asthma. Among the family history diagnoses, a family history of hypertension, cardiac disorder, and depression was important. Our engineered feature (WFL% change before age 2) was also among the top 10 predictors.

We also evaluated partial dependency plots (PDP[4]) for the WFL% change before age 2, and child opportunity index, as a function of the predicted risk of (the probability of developing) obesity at age 5 for the test dataset. Specifically, Figure 5.C shows that an increase in the WFL% before age 2 increases the risk of obesity at age 5. Figure 5.D shows that the risk of obesity decreases with increasing COI score.

[Figure 5 about here.]

4. Discussion

In this study, we developed and extensively validated a model that can predict obesity reliably in early childhood using a large EHR dataset with over 36 000 children using deep learning methods. Our descriptive analysis in Section 2.3 underscores the importance of having a tool that includes other elements beyond a static WFL% or BMI% to identify children at high-risk for developing obesity before their WFL% or BMI% goes above the 95th percentile. Our model used around 500 clinically relevant features from the EHR and demonstrated strong performance across multiple age ranges, chronological time periods, geographies, and demographic subgroups. Because our model is able to leverage unaugmented EHR data collected as part of routine clinical care and provides flexibility in the

4. Demonstrates the marginal effect of the different values of the feature of interest on the predicted outcome

age at which the model is applied, there is potential for this model to be integrated into common EHRs to provide clinical decision support at multiple time points and support the implementation of preventive measures before a child develops obesity between 3 to age 10.

To date, there has been limited research on using deep learning methods with large longitudinal EHR data to predict childhood obesity [34, 12, 18]. There is a large body of studies that use logistic regression to predict the development of obesity at a certain age, most commonly 5 years. Three of the most comparable ones are presented by Redsell et al. [35], Robson et al. [36], and Hammond et al. [11] with AUROC of 0.67, 0.78, and 0.81, respectively. While these studies used a smaller set of features, they also used features like parental weight, parental smoking habits, prenatal history, and infant dietary habits which are not available in many EHRs. Several advanced ML-based models were compared by Dugan et al. [37] and Pang et al. [24] using 167 features (anthropometric measures and questionnaire data), and 102 features (maternal and EHR data), respectively for children before age 2. They reported an accuracy of 0.85 using a decision-tree method at age 3 and an AUROC of 0.81 b using XGBoost for ages 2-7.

Our study improves on prior studies in several ways. Our predictive model achieves an AUROC of 0.85 (0.84-0.86), 0.83 (0.82-0.84), and 0.81 (0.80-0.82) for predicting the risk of obesity at 3, 4, and age 5, respectively, using 0-2 years of data. Our predictive model also demonstrates high accuracy for estimating the risk of obesity for the next 3 years across a wide age range from 2 to 7 years of age (Table 2). Indeed, our model provides the flexibility of learning from as much data as available for patients before the time of screening, which is a method referred to as a "flexible window design" that our team developed in a prior paper [27]. Because of this flexibility, it provides a tool to screen children at different ages and with enough of a time window before the development of obesity for preventive interventions to be effective.

Beyond demonstrating the accuracy and flexibility of our model, we demonstrated the validity of our model across temporal, geographic shifts, multiple demographic strata and the last WFL% before age 2 and demonstrated our model's sensitivity in predicting obesity. As a screening tool, having a model that is highly sensitive is preferred to ensure that young children at risk for developing obesity are identified and preventive measures like parenting interventions, lifestyle behavior counseling, and weight checks can be implemented in a timely manner. Notably, the sensitivity of the model does mean that there is a possibility of false positives; however, because obesity prevention measures are generally low-cost and beneficial for all children no matter their weight status, we believe this is an acceptable trade-off as long as discussion about a child's risk for developing obesity is done in a family-centered, thoughtful, and positive way and prevent any weight-related stigma that could otherwise occur.

Another important strength of our model is that it utilizes only EHR data that is collected as part of routine clinical care. Our feature selection process can be applied to any EHR dataset that contains common data elements of basic demographics, diagnoses codes, medications, and measurements. This is advantageous compared to prior models which may require extensive data collection or data linkages that are not feasible outside of the research setting.

Similar to other studies [32, 38], our evaluation of the importance of the EHR features to our model also allows us to better understand what risk factors are important to a child's risk of obesity, which can facilitate the provision of more personalized interventions. For example, stressing healthy lifestyles for infants whose family has a strong cardiovascular history could be an effective intervention based on our model. In addition to certain diagnoses and family history, our results demonstrated that change in WFL% in infancy is an important feature to calculate from EHR measurements to determine a child's risk of obesity. Being able to learn from these longitudinal engineered and other EHR features is an advantage of our deep learning model.

We also found that the risk of obesity decreases with the increasing child opportunity index (COI), corroborating the importance of social determinants of health on childhood obesity [39, 40]. While COI may not be part of routine EHR data, it is easily calculated with information that is in the EHR (through ZIP codes) and our study demonstrates the importance of accounting for the influence of neighborhood environments on child health outcomes.

Our study does have some limitations. First, our study is retrospective and may be affected by changes in healthcare delivery over time. However, we were able to temporally validate the model to account for time shifts between children born between 2002-2009 and 2010-2015. Second, the dataset included information from a single healthcare system. Despite this, the dataset did include children from five different states in different geographic regions and we were able to validate our model between these geographic regions. Third, a large number of children were excluded due to inadequate or missing height and weight measurements before age 2 and this may have introduced sample bias. Finally, our dataset did not contain information on lifestyle behaviors, prenatal variables, or other sociodemographic variables, which we know are important to a child's risk for the development of obesity. However, our focus was to use an EHR dataset that included only data collected during routine clinical care, to increase the potential that the model could be integrated into existing EHR systems to provide clinical decision support for providers in identifying children at risk for obesity to facilitate early prevention efforts across many pediatric healthcare settings. Our team is actively working on creating a CDS (clinical decision support) tool on the SMART on FHIR (Fast Healthcare Interoperability Resources) platform [41], allowing its wider adoption and integration in clinical practice and increasing its interoperability.

5. Data availability

Our code, containing the model with parameter (weight) values, is publicly available on GitHub at `https://github.com/healthylaife/ObesityPrediction`. Interested scholars

can access the data by contacting Nemours Biomedical Research Informatics Center and signing a data use agreement.

References

[1] CDC. Centers for disease control and prevention, childhood obesity facts, 2023. URL `https://www.cdc.gov/obesity/data/childhood.html`. Accessed: June 2023.

[2] Katie A Loth, Jocelyn Lebow, Marc James Abrigo Uy, Samantha M Ngaw, Dianne Neumark-Sztainer, and Jerica M Berge. First, Do No Harm: Understanding Primary Care Providers' Perception of Risks Associated With Discussing Weight With Pediatric Patients. *Global Pediatric Health*, 8:2333794X211040979, 2021. doi: 10.1177/2333794x211040979.

[3] Jonathan D Klein, Tracy S Sesselberg, Mark S Johnson, Karen G O'Connor, Stephen Cook, Marian Coon, Charles Homer, Nancy Krebs, and Reginald Washington. Adoption of body mass index guidelines for screening and counseling in pediatric practice. *Pediatrics*, 125(2):265–272, 2010.

[4] Stefan Kuhle, Sara F L Kirk, Arto Ohinmaa, and Paul J Veugelers. Comparison of ICD code-based diagnosis of obesity with measured obesity in children and the implications for health care cost estimates. *BMC Medical Research Methodology*, 11(1):1–5, 2011.

[5] Lan Liang, Chad Meyerhoefer, and Justin Wang. Obesity counseling by pediatric health professionals: an assessment using nationally representative data. *Pediatrics*, 130(1):67–77, 2012.

[6] Kyung E Rhee, Stephanie Kessl, Sarah Lindback, Marshall Littman, and Robert E El-Kareh. Provider views on childhood obesity management in primary care settings: a mixed methods analysis. *BMC Health Services Research*, 18(1):55, 2018. doi: 10.1186/s12913-018-2870-y. URL `https://doi.org/10.1186/s12913-018-2870-y`.

[7] Eliana M Perrin, Joanne P Finkle, and John T Benjamin. Obesity prevention and the primary care pediatrician's office. *Current Opinion in Pediatrics*, 19(3):354–361, 2007. doi: 10.1097/MOP.0b013e328151c3e9. URL https://journals.lww.com/co-pediatrics/Fulltext/2007/06000/Obesity_prevention_and_the_primary_care.22.aspx.

[8] John Conrad Rausch, Emily Rothbaum Perito, and Patricia Hametz. Obesity prevention, screening, and treatment: practices of pediatric providers since the 2007 expert committee recommendations. *Clinical Pediatrics*, 50(5):434–441, 2011.

[9] P O A Monteiro and C G Victora. Rapid growth in infancy and childhood and obesity in later life–a systematic review. *Obesity Reviews: An Official Journal of the International Association for the Study of Obesity*, 6(2):143–154, 2005. doi: 10.1111/j.1467-789X.2005.00183.x. URL http://www.ncbi.nlm.nih.gov/pubmed/15836465.

[10] Nida Ziauddeen, Paul J Roderick, Nicholas S Macklon, and Nisreen A Alwan. Predicting childhood overweight and obesity using maternal and early life risk factors: a systematic review. *Obesity Reviews*, 19(3):302–312, 2018.

[11] Robert Hammond, Rodoniki Athanasiadou, Silvia Curado, Yindalon Aphinyanaphongs, Courtney Abrams, Mary Jo Messito, Rachel Gross, Michelle Katzow, Melanie Jay, Narges Razavian, and Brian Elbel. Predicting childhood obesity using electronic health records and publicly available data. *PLOS ONE*, 14(4):e0215571, 2019. doi: 10.1371/journal.pone.0215571. URL https://doi.org/10.1371/journal.pone.0215571.

[12] Gonzalo Colmenarejo. Machine Learning Models to Predict Childhood and Adolescent Obesity: A Review. *Nutrients*, 12(8):2466, 2020. URL https://www.mdpi.com/2072-6643/12/8/2466.

[13] Elizabeth A. Campbell, Ting Qian, Jeffrey M. Miller, Ellen J. Bass, and Aaron J. Masino. Identification of temporal condition patterns associated with pediatric obesity

incidence using sequence mining and big data. *International Journal of Obesity*, 44(8): 1753–1765, Aug 2020. ISSN 1476-5497. doi: 10.1038/s41366-020-0614-7.

[14] Hera Siddiqui, Ajita Rattani, Nikki K Woods, Laila Cure, Rhonda K Lewis, Janet Twomey, Betty Smith-Campbell, and Twyla J Hill. A survey on machine and deep learning models for childhood and adolescent obesity. *IEEE Access*, 9:157337–157360, 2021.

[15] K Nirmala Devi, N Krishnamoorthy, P Jayanthi, S Karthi, T Karthik, and K Kiranbharath. Machine learning based adult obesity prediction. In *2022 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–5. IEEE, 2022.

[16] Xiaobei Zhou, Lei Chen, and Hui-Xin Liu. Applications of machine learning models to predict and prevent obesity: A mini-review. *Frontiers in Nutrition*, 9, 2022.

[17] Antonio Ferreras, Sandra Sumalla-Cano, Rosmeri Martínez-Licort, Iñaka Elío, Kilian Tutusaus, Thomas Prola, Juan Luís Vidal-Mazón, Benjamín Sahelices, and Isabel de la Torre Díez. Systematic review of machine learning applied to the prediction of obesity and overweight. *Journal of Medical Systems*, 47(1):1–11, 2023.

[18] Mehak Gupta, Thao-Ly T Phan, H Timothy Bunnell, and Rahmatollah Beheshti. Obesity Prediction with EHR Data: A deep learning approach with interpretable elements. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(3):1–19, 2022.

[19] Adi Horesh, Avishai M Tsur, Aya Bardugo, and Gilad Twig. Adolescent and childhood obesity and excess morbidity and mortality in young adulthood—a systematic review. *Current obesity reports*, 10(3):301–310, 2021.

[20] Jixing Zhou, Fu Zhang, Xiaoyun Qin, Peixuan Li, Yuzhu Teng, Shanshan Zhang, Fangbiao Tao, and Kun Huang. Age at adiposity rebound and the relevance for obesity: a systematic review and meta-analysis. *International journal of obesity*, 46(8):1413–1424, 2022.

[21] David S Freedman, Amy J Goodwin-Davies, Lyudmyla Kompaniyets, Samantha J Lange, Alyson B Goodman, Thao-Ly Tam Phan, Suchitra Rao, Ihuoma Eneli, and Christopher B Forrest. Interrelationships among age at adiposity rebound, bmi during childhood, and bmi after age 14 years in an electronic health record database. *Obesity*, 30(1):201–208, 2022.

[22] Laura N Anderson, Yulika Yoshida-Montezuma, Nora Dewart, Ezza Jalil, Jayati Khattar, Vanessa De Rubeis, Sarah Carsley, Lauren E Griffith, and Lawrence Mbuagbaw. Obesity and weight change during the covid-19 pandemic in children and adults: A systematic review and meta-analysis. *Obesity Reviews*, page e13550, 2023.

[23] Erinn T Rhodes, Thao-Ly T Phan, Elizabeth R Earley, Ihuoma Eneli, Matthew A Haemer, Nikki C Highfield, Saba Khan, Grace Kim, Shelley Kirk, Elizabeth Monti Sullivan, et al. Patient-Reported Outcomes to Describe Global Health and Family Relationships in Pediatric Weight Management. *Childhood Obesity*, 2023.

[24] Xueqin Pang, Christopher B Forrest, Félice Lê-Scherban, and Aaron J Masino. Prediction of early childhood obesity with machine learning and electronic health record data. *International Journal of Medical Informatics*, 150:104454, 2021.

[25] David C Grossman, Kirsten Bibbins-Domingo, Susan J Curry, Michael J Barry, Karina W Davidson, Chyke A Doubeni, John W Epling, Alex R Kemper, Alex H Krist, Ann E Kurth, et al. Screening for obesity in children and adolescents: Us preventive services task force recommendation statement. *Jama*, 317(23):2417–2426, 2017.

[26] Institute for Child, Youth, and Family Policy. Child Opportunity Index, 2022. URL `https://www.diversitydatakids.org`. Heller School for Social Policy and Management, Brandeis University, Accessed: Nov 2022.

[27] Mehak Gupta, Raphael Poulain, Thao-Ly T. Phan, H. Timothy Bunnell, and Rahmatollah Beheshti. Flexible-Window Predictions on Electronic Health Records. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12510–12516, Jun.

2022. doi: 10.1609/aaai.v36i11.21520. URL https://ojs.aaai.org/index.php/AAAI/article/view/21520.

[28] Hagai Rossman, Smadar Shilo, Shiri Barbash-Hazan, Nitzan Shalom Artzi, Eran Hadar, Ran D Balicer, Becca Feldman, Arnon Wiznitzer, and Eran Segal. Prediction of childhood obesity from nationwide health records. *The Journal of Pediatrics*, 233:132–140, 2021.

[29] Andrew J Vickers and Elena B Elkin. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006.

[30] Michael Matheny, S Thadaney Israni, Mahnoor Ahmed, and Danielle Whicher. Artificial intelligence in health care: the hope, the hype, the promise, the peril. *Washington, DC: National Academy of Medicine*, 2019.

[31] Ben Van Calster, David J McLernon, Maarten Van Smeden, Laure Wynants, and Ewout W Steyerberg. Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):1–7, 2019.

[32] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.

[33] Juan Du, Dajian Zeng, Zhao Li, Jingxuan Liu, Mingqi Lv, Ling Chen, Dan Zhang, and Shouling Ji. An interpretable outcome prediction model based on electronic health records and hierarchical attention. *International Journal of Intelligent Systems*, 37(6): 3460–3479, 2022.

[34] Andreas Triantafyllidis, Eleftheria Polychronidou, Anastasios Alexiadis, Cleilton Lima Rocha, Douglas Nogueira Oliveira, Amanda S da Silva, Ananda Lima Freire, Crislanio Macedo, Igor Farias Sousa, Eriko Werbet, et al. Computerized decision support and

machine learning applications for the prevention and treatment of childhood obesity: A systematic review of the literature. *Artificial Intelligence In Medicine*, 104:101844, 2020.

[35] Sarah A Redsell, Stephen Weng, Judy A Swift, Dilip Nathan, and Cris Glazebrook. Validation, optimal threshold determination, and clinical utility of the infant risk of overweight checklist for early prevention of child overweight. *Childhood Obesity*, 12(3): 202–209, 2016.

[36] Jacob O Robson, Sofia G Verstraete, Stephen Shiboski, Melvin B Heyman, and Janet M Wojcicki. A risk score for childhood obesity in an urban latino cohort. *The Journal of pediatrics*, 172:29–34, 2016.

[37] Tamara M Dugan, S Mukhopadhyay, Aaron Carroll, and Stephen Downs. Machine learning techniques for prediction of early childhood obesity. *Applied clinical informatics*, 6(03):506–520, 2015.

[38] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1903–1911, 2017.

[39] Izzuddin M Aris, Wei Perng, Dana Dabelea, Amy M Padula, Akram Alshawabkeh, Carmen M Vélez-Vega, Judy L Aschner, Carlos A Camargo, Tamara J Sussman, Anne L Dunlop, et al. Associations of neighborhood opportunity and social vulnerability with trajectories of childhood body mass index and obesity among us children. *JAMA Network Open*, 5(12):e2247957–e2247957, 2022.

[40] Mehak Gupta, Thao-Ly T Phan, Félice Lê-Scherban, Daniel Eckrich, H Timothy Bunnell, and Rahmatollah Beheshti. Associations of longitudinal BMI percentile classification patterns in early childhood with neighborhood-level social determinants of health. *medRxiv*, pages 2023–06, 2023.

[41] Joshua C Mandel, David A Kreda, Kenneth D Mandl, Isaac S Kohane, and Rachel B Ramoni. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. *Journal of the American Medical Informatics Association*, 23(5):899–908, 02 2016. ISSN 1067-5027. doi: 10.1093/jamia/ocv189. URL https://doi.org/10.1093/jamia/ocv189.

List of Tables

1 Statistics for the study cohort (full), as well as the cohorts used for temporal and geographic validation. Counts with the percentage prevalence in the respective cohort for last obesity status before age 2, sex, race, ethnicity, and payer are shown. In the bottom part, the mean(standard deviation) for the number of times that weight, height, diagnoses, medications, and lab measurements were recorded before and after age 2 are shown. 22

2 Predictive performance across six different observation windows. Results are shown for prediction for the next 1, 2, and 3 years. The fixed values for sensitivity and specificity are highlighted. The net benefits are shown at three thresholds of 20, 40, and 60 percent. 23

Table 1: Statistics for the study cohort (full), as well as the cohorts used for temporal and geographic validation. Counts with the percentage prevalence in the respective cohort for last obesity status before age 2, sex, race, ethnicity, and payer are shown. In the bottom part, the mean(standard deviation) for the number of times that weight, height, diagnoses, medications, and lab measurements were recorded before and after age 2 are shown.

		Temporal Train(n=26 786)	**Temporal Test**(n=9 405)	**Geographic Train**(n=32 848)	**Geographic Test**(n=3 343)	**All** (n=36 191)
Last obesity Status before age 2						
Underweight (WFL< 5%)	Count(%)	376(1.4)	128(1.36)	399(1.2)	105(3.14)	504(1.39)
Normal (5%<=WFL< 85%)	Count(%)	13 781(51.44)	5 108(54.31)	17 067(51.9)	1 822(54.5)	18 889(52.19)
Overweight (85%<=WFL< 95%)	Count(%)	5 811(21.69)	1 979(21.04)	7 164(21.8)	626(18.72)	7 790(21.52)
Obesity (WFL>=95%)	Count(%)	6 818(25.45)	2 190(23.28)	8 218(25.01)	790(23.62)	9 008(24.89)
Sex:						
Female	Count(%)	1 2643(47.20)	4 355(46.30)	15 549(47.33)	1 449(43.34)	16 998(46.96)
Male	Count(%)	14 143(52.79)	5 050(53.69)	17 299(52.66)	1 894(56.65)	19 193(53.03)
Ethnicity:						
Hispanic	Count(%)	3 475(12.97)	1 323(14.06)	4 277(13.02)	521(15.58)	4 798(13.25)
Non-Hispanic	Count(%)	23 197(86.60)	8 012(85.18)	28 432(86.55)	2 777(83.06)	31 209(86.23)
Race:						
Asian	Count(%)	472(1.76)	191(2.03)	610(1.85)	53(1.58)	663(1.83)
Black	Count(%)	11 524(43.02)	3 721(39.56)	14 694(44.73)	551(16.48)	15 245(42.12)
White	Count(%)	11 352(42.38)	4 088(43.46)	13 496(41.08)	1 944(58.15)	15 440(42.66)
Other	Count(%)	3 048(11.37)	1 216(12.92)	3 812(11.60)	452(13.52)	4 264(11.78)
Payer:						
Private	Count(%)	11 114(41.49)	3 850(40.93)	13 483(41.04)	1 481(44.30)	14 964(41.34)
Public	Count(%)	15 624(58.32)	5 539(58.89)	19 314(58.79)	1 849(55.30)	21 163(58.47)
Weight and height measurements before age 2	Mean(SD)	9.80(2.30)	10.11(2.24)	10.04(2.20)	8.28(2.53)	9.88(2.29)
Weight and height measurements between age 2 and 10	Mean(SD)	22.10(22.23)	17.07(19.16)	21.15(20.56)	17.86(15.47)	20.83(21.36)
Diagnoses available before age 2	Mean(SD)	2.95(3.41)	3.32(3.59)	2.92(3.38)	4.46(4.0)	3.03(3.44)
Diagnoses available between age 2 and 10	Mean(SD)	4.82(7.52)	3.87(6.16)	4.49(7.18)	5.81(7.80)	4.58(7.20)
Medications available before age 2	Mean(SD)	7.66(5.40)	9.97(7.87)	8.51(6.25)	5.27(5.25)	8.27(6.23)
Medications available between age 2 and 10	Mean(SD)	11.33(8.19)	10.75(8.87)	11.45(8.35)	7.99(7.93)	11.18(9.36)
Labs available before age 2	Mean(SD)	25.04(29.54)	26.81(29.54)	24.77(28.67)	32.70(35.66)	25.50(29.47)
Labs available between age 2 and 10	Mean(SD)	60.00(45.56)	43.95(31.46)	54.69(40.41)	66.99(61.44)	55.83(42.93)

Table 2: Predictive performance across six different observation windows. Results are shown for prediction for the next 1, 2, and 3 years. The fixed values for sensitivity and specificity are highlighted. The net benefits are shown at three thresholds of 20, 40, and 60 percent.

Input age range		**Prediction Age**		
0-2		3 (n=7 224)	4 (n=8 875)	5 (n=8 982)
	AUROC(95%CI)	0.87(0.85-0.88)	0.84(0.81-0.86)	0.81(0.79-0.83)
	Temporal validation	0.85(0.84-0.86)	0.83(0.82-0.84)	0.81(0.80-0.82)
	Geographic validation	0.82(0.80-0.83)	0.80(0.79-0.82)	0.77(0.76-0.79)
	Baseline	0.77(0.75-0.78)	0.75(0.73-0.77)	0.71(0.68-0.73)
	Sensitivity/Specificity	0.90 /0.56	0.90 /0.52	0.90 /0.43
		0.95 /0.40	0.95 /0.37	0.95 /0.30
		0.62/ 0.90	0.57/ 0.90	0.48/ 0.90
		0.47/ 0.95	0.42/ 0.95	0.37/ 0.95
	Net Benefit(20/40/60)	0.14/0.09/0.05	0.12/0.07/0.04	0.12/0.08/0.04
0-3		4 (n=7 224)	5 (n=7 224)	6 (n=7 216)
	AUROC(95%CI)	0.90(0.87-0.92)	0.88(0.86-0.90)	0.84(0.82-0.86)
	Temporal validation	0.89(0.87-0.91)	0.86(0.84-0.88)	0.83(0.81-0.85)
	Geographic validation	0.85(0.84-0.87)	0.83(0.81-0.84)	0.79(0.77-0.80)
	Baseline	0.82(0.81-0.84)	0.78(0.76-0.80)	0.74(0.72-0.75)
	Sensitivity/Specificity	0.90 /0.70	0.90 /0.60	0.90 /0.55
		0.95 /0.55	0.95 /0.46	0.95 /0.40
		0.68/ 0.90	0.62/ 0.90	0.54/ 0.90
		0.54/ 0.95	0.46/ 0.95	0.40/ 0.95
	Net Benefit(20/40/60)	0.17/0.12/0.08	0.14/0.09/0.05	0.12/0.07/0.03
0-4		5 (n=8 875)	6 (n=8 867)	7 (n=5 158)
	AUROC(95%CI)	0.91(0.89-0.92)	0.88(0.86-0.90)	0.86(0.84-0.87)
	Temporal validation	0.90(0.88-0.92)	0.86(0.85-0.88)	0.84(0.82-0.86)
	Geographic validation	0.87(0.85-0.89)	0.84(0.82-0.85)	0.82(0.80-0.84)
	Baseline	0.84(0.82-0.83)	0.82(0.81-0.83)	0.78(0.75-0.80)
	Sensitivity/Specificity	0.90 /0.71	0.90 /0.59	0.90 /0.53
		0.95 /0.58	0.95 /0.43	0.95 /0.38
		0.71/ 0.90	0.62/ 0.90	0.57/ 0.90
		0.58/ 0.95	0.51/ 0.95	0.45/ 0.95
	Net Benefit(20/40/60)	0.16/0.11/0.08	0.12/0.09/0.05	0.12/0.08/0.05
0-5		6 (n=8 974)	7 (n=5 256)	8 (n=2 300)
	AUROC(95%CI)	0.92(0.90-0.93)	0.90(0.88-0.92)	0.87(0.85-0.88)
	Temporal validation	0.91(0.89-0.92)	0.88(0.87-0.89)	0.85(0.84-0.86)
	Geographic validation	0.89(0.87-0.91)	0.85(0.83-0.86)	0.83(0.82-0.84)
	Baseline	0.83(0.82-0.84)	0.80(0.79-0.81)	0.78(0.75-0.80)
	Sensitivity/Specificity	0.90 /0.71	0.90 /0.69	0.90 /0.54

		0.95 /0.57	0.95 /0.53	0.95 /0.39
		0.75/ 0.90	0.72/ 0.90	0.64/ 0.90
		0.64/ 0.95	0.59/ 0.95	0.52/ 0.95
	Net Benefit(20/40/60)	0.17/0.12/0.08	0.15/0.11/0.08	0.16/0.12/0.07
0-6		7 (n=5 272)	8 (n=2 314)	9 (n=1 249)
	AUROC(95%CI)	0.92(0.90-0.95)	0.90(0.88-0.91)	0.88(0.86-0.90)
	Temporal validation	0.92(0.91-0.94)	0.89(0.87-0.90)	0.86(0.85-0.88)
	Geographic validation	0.90(0.88-0.91)	0.87(0.85-0.89)	0.84(0.83-0.86)
	Baseline	0.84(0.82-0.85)	0.82(0.81-0.83)	0.79(0.78-0.80)
	Sensitivity/Specificity	0.90 /0.74	0.90 /0.61	0.90 /0.54
		0.95 /0.60	0.95 /0.46	0.95 /0.34
		0.79/ 0.90	0.70/ 0.90	0.56/ 0.90
		0.67/ 0.95	0.59/ 0.95	0.45/ 0.95
	Net Benefit(20/40/60)	0.16/0.11/0.08	0.17/0.11/0.08	0.16/0.09/0.07
0-7		8 (n=2 316)	9 (n=1 150)	10 (n=539)
	AUROC(95%CI)	0.93(0.92-0.94)	0.91(0.88-0.92)	0.87(0.85-0.89)
	Temporal validation	0.92(0.90-0.93)	0.90(0.88-0.91)	0.85(0.82-0.87)
	Geographic validation	0.90(0.89-0.92)	0.88(0.86-0.90)	0.84(0.81-0.85)
	Baseline	0.84(0.82-0.85)	0.82(0.81-0.83)	0.79(0.78-0.80)
	Sensitivity/Specificity	0.90 /0.70	0.90 /0.63	0.90 /0.56
		0.95 /0.54	0.95 /0.47	0.95 /0.40
		0.75/ 0.90	0.67/ 0.90	0.59/ 0.90
		0.65/ 0.95	0.55/ 0.95	0.46/ 0.95
	Net Benefit(20/40/60)	0.16/0.11/0.09	0.16/0.09/0.09	0.17/0.09/0.08

List of Figures

1 The cohort selection steps. Temporal and geographic validation regimes are shown in gray. 26

2 Study of the distribution of children with obesity (BMI% $\geq$ 95) at age 3 to 10. We analyze what percentage of children with obesity at ages 3 to 10 had WFL% $<$ 5, 5 $\leq$ WFL% $<$ 85, 85 $\leq$ WFL% $<$ 95, and WFL% $\geq$ 95 at age 2. X-axis shows the distribution of the cohort with obesity at every age from 3 to age 10. Four categories of WFL% at age 2 are shown on the right. . . . 27

3 Evaluation of obesity prediction model to predict obesity at age 5 using 0-2 years of data: **A.** AUROC curve of the model (blue) and a baseline model based on the last available WFL% or BMI% measurement in the observation window (orange line), **B.** AUPRC curve of the model (blue) and a baseline model based on the last available WFL% measurement before age 2 (orange line), **C.** Decision curve analysis for different strategies of treatment, **D.** Calibration curve. The dotted line represents ideal calibration and the orange line for before calibration and the green line after calibration. 28

4 Evaluating the predictive model's robustness by comparing AUROCs (in the test dataset) across five groups (13 subgroups): last WFL before age 2 (3 categories), race (Asian, Black, White, Other), ethnicity (Hispanic, Non-Hispanic), sex (female, male), and payer (private, public). 29

5 Feature Importance Analysis for all samples to predict obesity at age 5 using 0-2 years of data. **(A)** Feature importance ranking of 7 feature groups, **(B)** Relative feature importance ranking of top 20 features, **C-D** partial dependence plots between risk for obesity and **(C)** Change in WFL% before age 2 over 5 percentile categories (Dividing the numeric data (WFL%) into 5 bins (aka quintiles) such that there are an equal number of observations in each bin. This would produce a categorical object indicating quantile membership for each data point. The x-axis shows the range of each bin.), **(D)** child opportunity index (COI) at age 2 over COI-decile (Dividing the numeric data into 10 bins (aka deciles) such that there are an equal number of observations in each bin. This would produce a categorical object indicating decile membership for each data point. The x-axis shows the range of each bin.) categories. 30

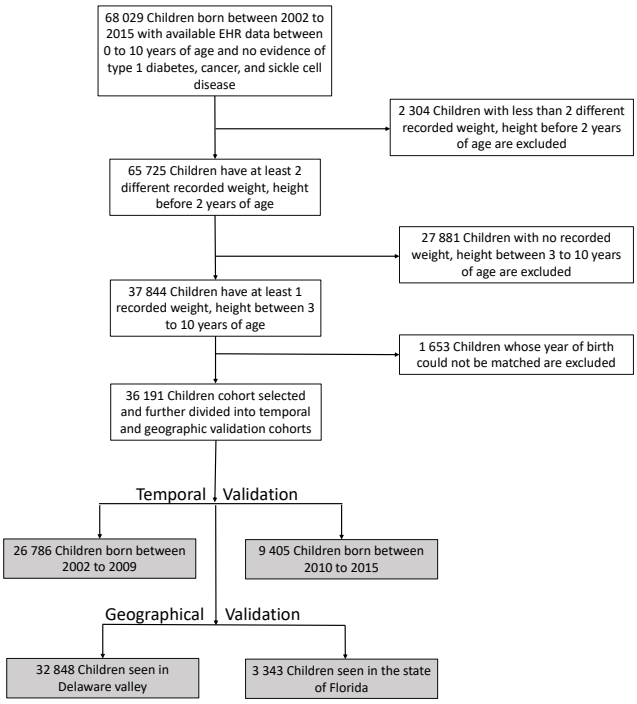

Figure 1: The cohort selection steps. Temporal and geographic validation regimes are shown in gray.

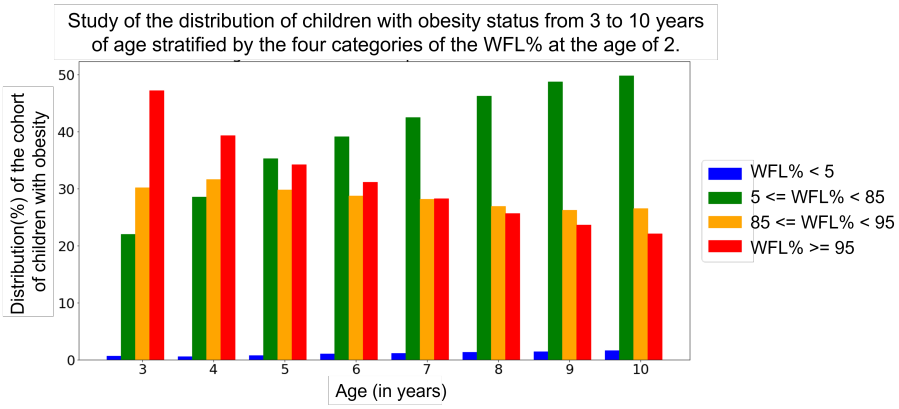

Figure 2: Study of the distribution of children with obesity (BMI% $\geq$ 95) at age 3 to 10. We analyze what percentage of children with obesity at ages 3 to 10 had WFL% $<$ 5, 5 $\leq$ WFL% $<$ 85, 85 $\leq$ WFL% $<$ 95, and WFL% $\geq$ 95 at age 2. X-axis shows the distribution of the cohort with obesity at every age from 3 to age 10. Four categories of WFL% at age 2 are shown on the right.

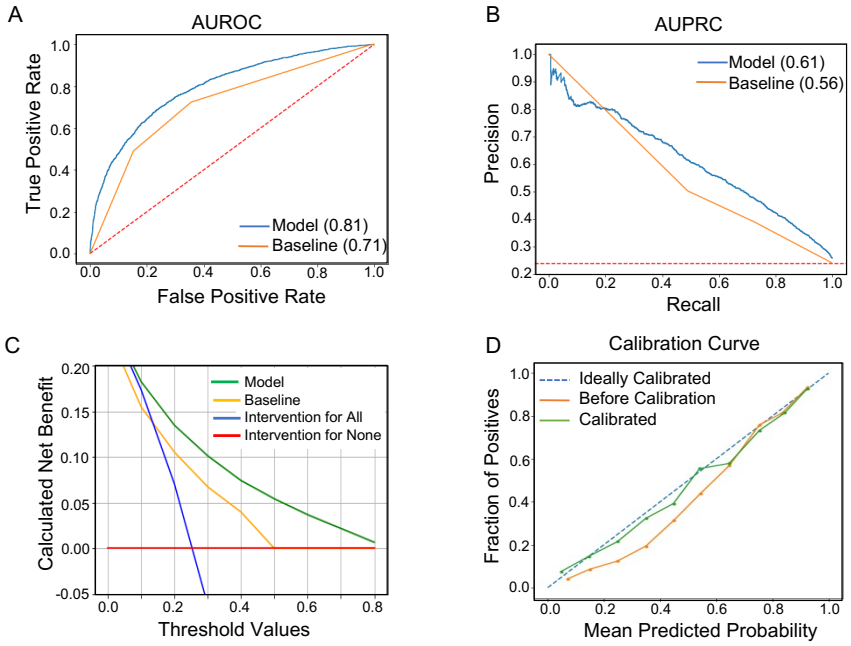

Figure 3: Evaluation of obesity prediction model to predict obesity at age 5 using 0-2 years of data: **A.** AUROC curve of the model (blue) and a baseline model based on the last available WFL% or BMI% measurement in the observation window (orange line), **B.** AUPRC curve of the model (blue) and a baseline model based on the last available WFL% measurement before age 2 (orange line), **C.** Decision curve analysis for different strategies of treatment, **D.** Calibration curve. The dotted line represents ideal calibration and the orange line for before calibration and the green line after calibration.

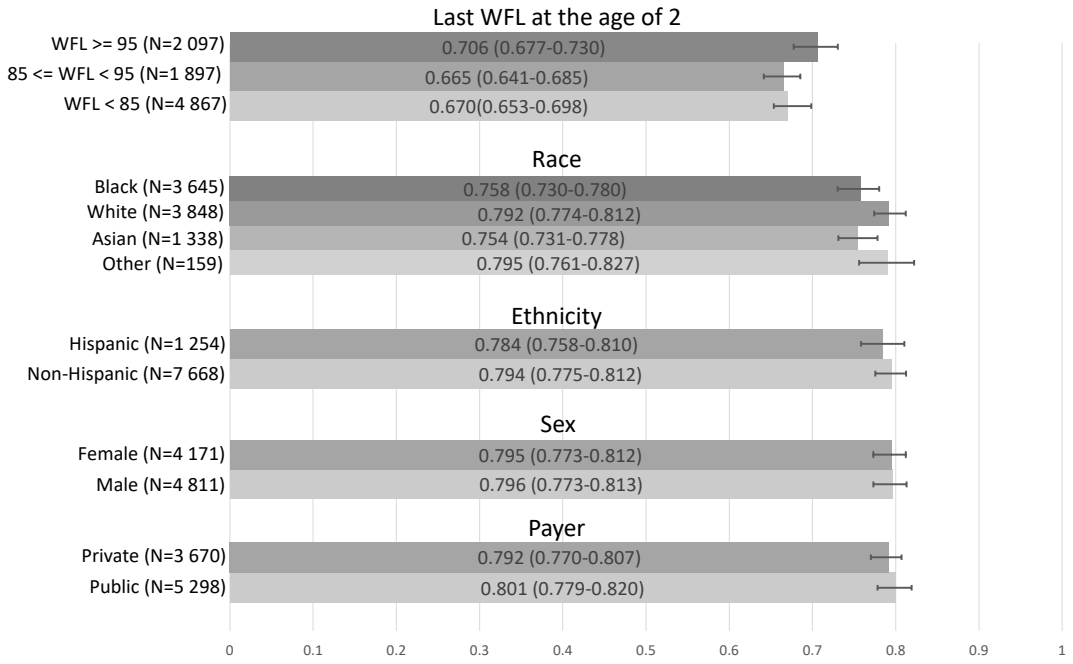

Figure 4: Evaluating the predictive model's robustness by comparing AUROCs (in the test dataset) across five groups (13 subgroups): last WFL before age 2 (3 categories), race (Asian, Black, White, Other), ethnicity (Hispanic, Non-Hispanic), sex (female, male), and payer (private, public).

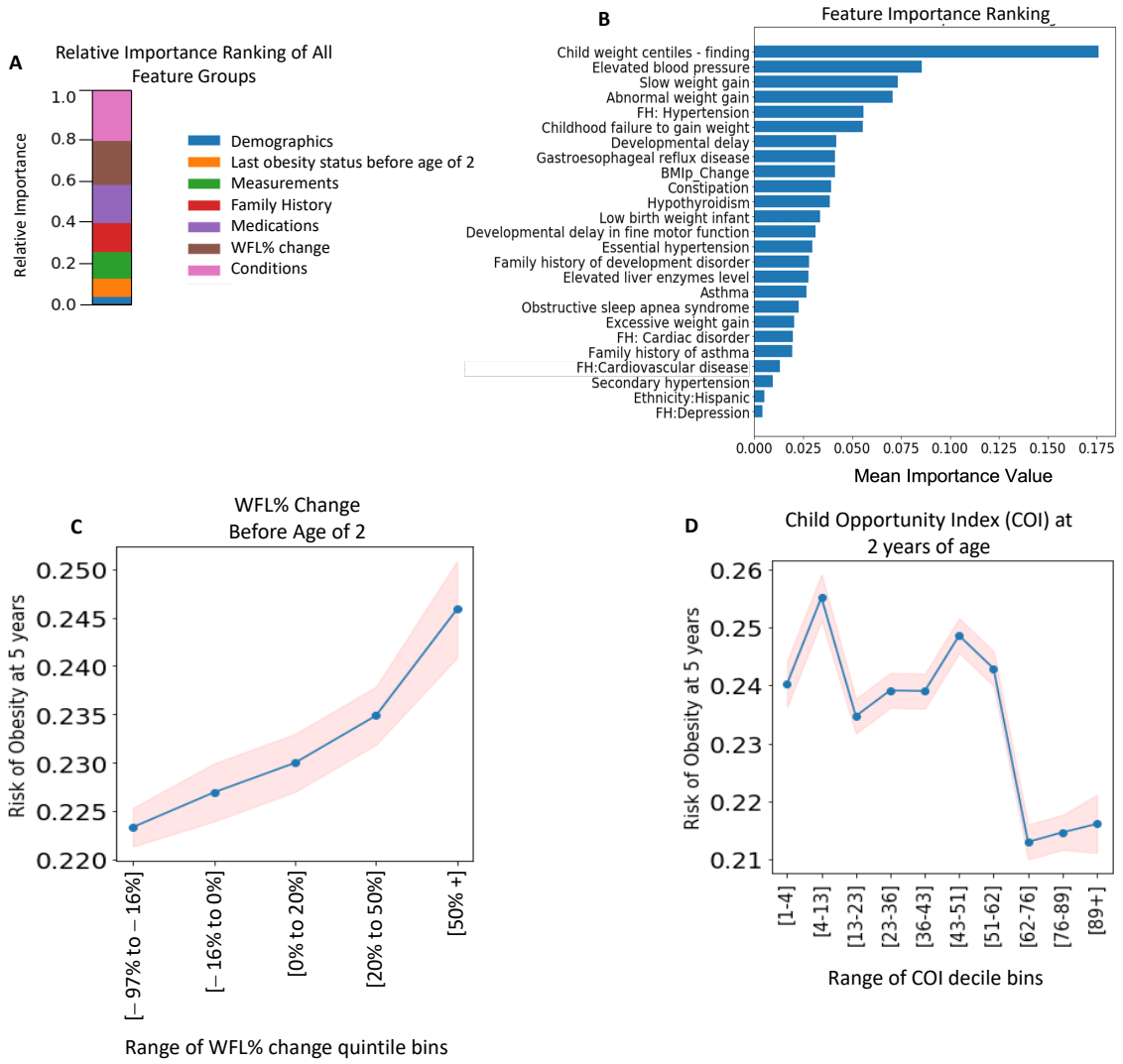

Figure 5: Feature Importance Analysis for all samples to predict obesity at age 5 using 0-2 years of data. **(A)** Feature importance ranking of 7 feature groups, **(B)** Relative feature importance ranking of top 20 features, **C-D** partial dependence plots between risk for obesity and **(C)** Change in WFL% before age 2 over 5 percentile categories (Dividing the numeric data (WFL%) into 5 bins (aka quintiles) such that there are an equal number of observations in each bin. This would produce a categorical object indicating quantile membership for each data point. The x-axis shows the range of each bin.), **(D)** child opportunity index (COI) at age 2 over COI-decile (Dividing the numeric data into 10 bins (aka deciles) such that there are an equal number of observations in each bin. This would produce a categorical object indicating decile membership for each data point. The x-axis shows the range of each bin.) categories.