## Introduction

This tutorial provides a structured, hands-on introduction to developing predictive machine learning models for childhood obesity using structured pediatric health data. It is intended as a companion to the guidelines paper, “A Primer for Machine Learning Applications in Childhood Obesity,” and demonstrates how key modeling decisions outlined in the main paper can be applied in practice. The focus is on translating conceptual recommendations into concrete, executable steps, designed to support researchers and clinicians who may not have prior experience with machine learning.

The dataset used in this tutorial is a synthesized version based on the Food and Brain Study ([www.clinicaltrials.gov](http://www.clinicaltrials.gov) : NCT03341247). While modeled on real longitudinal pediatric data, it has been generated to reflect realistic measurement complexity and missingness patterns while preserving participant anonymity. It includes individual and caregiver-reported features covering eating behavior, physical activity, psychological traits, and anthropometry, with a follow-up BMI percentile collected approximately one year later.

This tutorial is designed primarily as an educational exercise. It prioritizes conceptual understanding of modeling logic and decision-making over achieving optimal predictive performance. Users are guided through the end-to-end ML pipeline (including missing data handling, age-sensitive feature engineering, sex-stratified data splitting, model training, and interpretation) with explicit attention to the developmental and measurement challenges unique to child health data. The tutorial emphasizes how and why each step matters, particularly when working with caregiver-reported, behaviorally rich, and developmentally variable inputs.

The modeling approach aligns with a socio-ecological framework, focusing on proximal influences on child weight gain at the individual and family levels. Compared to adult health data, modeling pediatric outcomes requires special attention to age-linked variation, proxy-reported data, and smaller sample sizes. These considerations are addressed throughout the tutorial, with explanations provided for how they inform model design, data preprocessing, and interpretation.

Finally, while the dataset and code presented here are intended to support learning and reproducibility, they are not intended for clinical deployment. The models trained are based on synthesized data and are meant to illustrate methodological principles rather than provide deployable prediction tools. By the end of the tutorial, users will understand how to structure a pediatric ML pipeline, apply child-specific modeling decisions, and interpret outputs with an awareness of both methodological and developmental constraints.

## Why This Tutorial?

### 1. Educational Purpose

This tutorial was created to meet the needs of pediatric health researchers and clinicians who are beginning to explore the use of ML in their work but may lack technical training. It uses realistic pediatric data to guide users through practical decisions required to develop a working predictive model. All code and workflow steps are explicitly paired with explanations of why each modeling decision matters, particularly in pediatric datasets where developmental variability and proxy-reported data introduce complexity.

The tutorial introduces key preprocessing tasks such as outlier handling, variable-specific imputation, and feature engineering that are often overlooked in general ML tutorials. Feature selection is demonstrated using methods such as Recursive Feature Elimination (RFE), which help reduce overfitting in datasets with many correlated behavioral predictors. Special attention is given to strategies that reflect common constraints in pediatric datasets, including modest sample sizes, non-random missingness, and multiple sources of measurement noise.

### 2. Addressing Research Gaps

Many existing ML models in pediatric obesity rely narrowly on demographic and anthropometric predictors, often excluding behavioral and caregiver-reported variables due to complexity or concerns about data quality. This tutorial directly addresses those limitations by including behaviorally rich features and demonstrating how they can be prepared for ML applications.

In addition, the tutorial introduces modeling decisions such as sex-stratified data splitting to account for known biological and behavioral differences in obesity development. These decisions reflect real-world considerations for pediatric researchers and clinicians, and they are grounded in the challenges outlined in the companion review paper.

### 3. Practical Considerations

The tutorial makes use of synthesized data and is not designed for real-world prediction or clinical use. It is intended solely as a reproducible teaching tool. The pipeline, code, and rationale for each modeling step are designed to build transferable skills and improve understanding of how to approach ML in child health contexts.

Although modeled on a specific pediatric cohort, the approaches illustrated here are applicable to other childhood obesity datasets that include behavioral, psychological, or family-level data.

## Who Is This For?

This tutorial is intended for pediatric health researchers, clinicians, graduate students, and public health professionals who are beginning to apply ML methods in childhood obesity studies. No prior experience with ML frameworks is assumed, though a basic understanding of Python and structured data analysis will be helpful. The goal is to provide an accessible yet rigorous example of how structured ML modeling can be applied to developmentally sensitive health data.

### How to Use This Tutorial

Each section of the tutorial is organized around a real-world task in model development, with paired code and explanation. Key tasks include:

* Data structure review and variable types
* Missing data analysis and targeted imputation
* Outlier detection and handling
* Feature engineering
* Recursive feature selection
* Stratified train-test splitting
* Model training and evaluation
* Model interpretation using SHAP

Each step is presented with code, context, and rationale based on pediatric modeling needs. The overall goal is not only to build a working model but also to understand the process behind each modeling choice.

By completing the tutorial, users will learn a replicable approach to predictive modeling in childhood obesity that is consistent with the methodological and ethical guidance presented in the accompanying review paper.

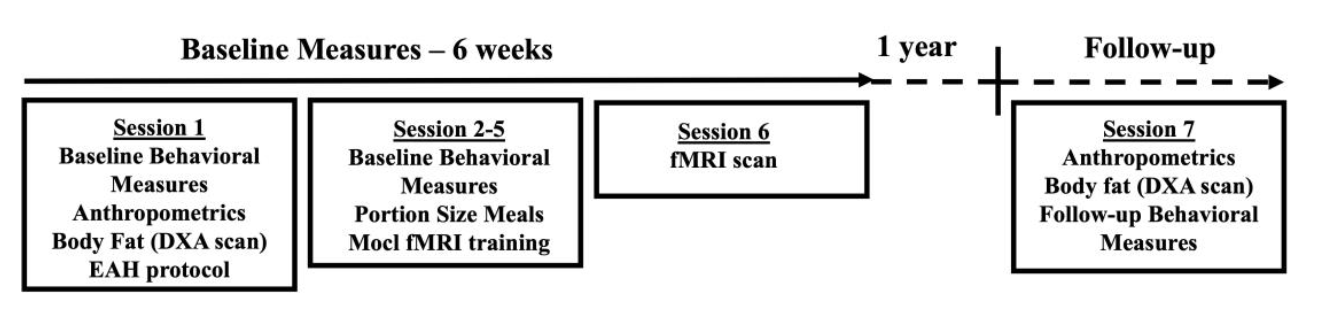
## Step 1: Dataset and Prediction Objective

### 1.1 Dataset Source and Design

The dataset used in this tutorial is a pseudo-synthesized version derived from the Food and Brain Study (NCT02855398). The original study enrolled children aged 7 to 9 years and followed them for approximately one to 1.5 years. The synthesized dataset maintains the structure and variable characteristics of the real study, including repeated meal assessments, psychological evaluations, and caregiver surveys, while protecting participant privacy. **Table 1** gives the predictors we considered for the study.

Key features of the original study design include (see **Figure 1** for timeline):

* Seven study visits (V1 to V7), with baseline data from Visits 1–6 and follow-up data at Visit 7.
* Stratification based on maternal weight status, distinguishing high and low obesity risk groups.
* Behavioral tasks measuring responses to varied portion sizes (100–200%).
* Objective measures of meal intake, eating in the absence of hunger (EAH), and structured caregiver-reported instruments.



**Figure 1.** Study visit timeline for Food and Brain Study

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 1. Variables considered for modeling** | | | | |
| **Variable Name in Code** | **Predictor** | **Explanation** |  | |
| Sex (sex\_1: one-hot encoded) | Sex | Biological sex of the child | |  |
| age\_yr | Age | Child’s age at baseline |  | |
| parent\_ed | Parent Education | Highest education level of parent |  | |
| risk\_status\_mom | Maternal BMI Risk Category | Low risk (BMI < 25); high risk (BMI ≥30) | (1) | |
| bmi\_percentile, v7\_bmi\_percentile | Body-Mass Index Percentile | Age- and sex-standardized BMI rank based on population growth references. >95th% = obesity | (2) | |
| v1\_meal\_total\_kcal | Baseline Meal Intake | Calories consumed at baseline meal |  | |
| v1\_eah\_total\_kcal | Baseline EAH Intake | Calories eaten when not hungry | (3–5) | |
| ps\_response\_slope, ps\_response\_int | Portion Size Response | Slope = intake change per gram served; intercept = intake at smallest portion | (6,7) | |
| bite\_size\_slope, bite\_size\_int | Bite Size Response | Slope = g/bite change with portion; intercept = initial bite size | (8–11) | |
| percent\_active\_eating\_slope, percent\_active\_eating\_intercept | Active Eating Time | Slope = % active time change; intercept = % at smallest portion |
| fswitch\_nok\_mean | Food Switching | Average switching between foods during meals | (8,12,13) | |
| mean\_mvpa | Physical Activity (weekly average) | Average time in moderate-to-vigorous physical activity | (14) | |
| mean\_sedentary | Sedentary Activity (weekly average) | Average time spent inactive or sitting |
| avg\_sleep\_frag\_index | Sleep Fragmentation Index | Mean weekly index of sleep interruptions | (15) | |
| cebq\_fr | CEBQ: Food Responsiveness | Eating in response to food cues | (16) | |
| cebq\_avoid | CEBQ: Food Avoidance | Lack of interest in eating or tendency to avoid food |
| cebq\_approach | CEBQ: Food Approach | General interest or enthusiasm for food |
| cebq\_eue | CEBQ: Emotional Overeating | Eating more in response to emotional states |
| cebq\_eue | CEBQ: Emotional Undereating | Eating less in response to emotional states |
| cebq\_dd | CEBQ: Desire to Drink | Frequent requests for drinks, especially sweetened beverages |
| cebq\_se | CEBQ: Satiety Responsiveness | Ability to stop eating when full |
| cebq\_sr | CEBQ: Slowness in Eating | Pace of eating, reflecting low eating speed |
| cebq\_ff | CEBQ: Food Fussiness | Selective eating or rejection of unfamiliar foods |
| cebq\_ef | CEBQ: Enjoyment of Food | General pleasure and interest in food |
| cfq\_pcw | CFQ: Perceived Child Weight | Parental perception of child’s current weight status | (17) | |
| cfq\_cwc | CFQ: Concern About Child Weight | Parental concern about the child becoming overweight |
| cfq\_resp | CFQ: Perceived Responsibility | Parental responsibility over child’s eating |
| cfq\_pressure | CFQ: Pressure to Eat | Parental pressure to eat |
| cfq\_rest | CFQ: Restriction | Restricting child’s eating |
| cfq\_mon | CFQ: Monitoring | Monitoring child’s intake |
| tfeq\_disinhibition | TFEQ: Disinhibition | Overeating tendency | (18) | |
| tfeq\_hunger | TFEQ: Hunger | Perceived hunger |
| tfeq\_cogcontrol | TFEQ: Cognitive Restraint | Conscious intake control |
| pwlb\_healthy | PWLB: Healthy Weight Control Strategy | Use of healthy weight-loss methods | (19) | |
| pwlb\_unhealthy | PWLB: Unhealthy Weight Control Strategy | Use of unhealthy weight-loss methods |
| ffbs\_control | FFBS: Maternal Control | Maternal control over child’s eating | (20) | |
| ffbs\_presence | FFBS: Maternal Presence | Maternal presence during meals |
| ffbs\_ch\_choice | FFBS: Child Choice | Child autonomy in food decisions |
| ffbs\_org | FFBS: Organization | Structure of eating environment |
| brief2\_gec\_p | BRIEF-2: Global Executive Composite | Cognitive self-regulation capacity (Executive Function) | (21) | |
| bas\_drive | BAS Drive | Motivation for goals | (22) | |
| bas\_funseeking | BAS Fun Seeking | Desire for new experiences |
| bas\_rewardresp | BAS Reward Responsiveness | Reward sensitivity |
| bis | BIS | Sensitivity to punishment |
| pds\_score | Pubertal Development Score | Puberty stage assessment | (23) | |
| BMI: Body Mass Index; EAH: Eating in the Absence of Hunger; MVPA: Moderate-to-Vigorous Physical Activity; CEBQ: Children’s Eating Behavior Questionnaire; CFQ: Child Feeding Questionnaire; TFEQ: Three-Factor Eating Questionnaire; PWLB: Parental Weight Loss Behavior; FFBS: Family Food Behavior Survey; BRIEF-2: Behavior Rating Inventory of Executive Function ver2; BIS: Behavioral Inhibition System; **BAS**: Behavioral Activation System | | | | |

### 1.2 Prediction Task

The goal of the ML pipeline is to predict BMI percentile at follow-up (Visit 7) using baseline features collected at Visit 1. This reflects a clinically relevant objective: identifying future obesity risk early enough to inform prevention.

* Prediction type: Supervised regression
* Target variable: v7\_bmi\_percentile
* Input variables: Behavioral, psychological, demographic, caregiver-reported, and anthropometric features from Visit 1

The use of a continuous outcome (BMI percentile) enables modeling risk as a spectrum, avoiding dichotomous labels such as obese vs. not obese. This is particularly applicable in this dataset where most children fall below the obesity threshold but still vary in adiposity risk.

Modeling the change in BMI percentile between the two visits may not be advisable due to the short follow-up period, which results in limited variability in BMI change and can reduce model stability and predictive signal.

### 1.3 Socio-Ecological Framework

Predictors in the dataset reflect two proximal levels of influence:

* *Child-level*: physical activity, eating behavior, psychological traits
* *Family-level*: caregiver perception, feeding styles, home food environment

These domains align with socio-ecological models of childhood obesity and capture modifiable influences that may precede changes in weight status. See **Figure 2** for the variety of metrics collected across the Food and Brain Study.

A diagram of a study measures

AI-generated content may be incorrect.

**Figure 2**. Metrics collected in Food and Brain study

### 1.4 Why This Dataset Is Useful for ML Training

This dataset includes several properties that make it useful for ML tutorial purposes:

* *Behavioral richness:* multiple eating behavior measures are included, often overlooked in ML studies.
* *Structured missingness:* realistic missing data patterns offer hands-on experience with imputation.
* *Variable diversity:* both continuous and categorical predictors spanning psychosocial, behavioral, and physiological domains.
* *Moderate sample size:* approximately 300 cases, typical of child cohort studies, useful for discussing model complexity vs. overfitting.

This setup enables illustration of practical modeling challenges such as structured missing data, collinearity, feature selection, and interpretation in developmentally sensitive datasets.

## Step 2: Feature Engineering

### 2.1 Why Feature Engineering Is Critical in Pediatric Datasets

Feature engineering transforms raw variables into structured, informative inputs that better capture behavioral and developmental patterns. In pediatric obesity research, this process must account for the developmental context, age-related differences in behavior, and non-linear relationships that are common in caregiver-reported or task-derived variables.

Many features in this dataset were not directly used in their raw form but were transformed to reflect response patterns, temporal scaling, or age-adjusted relevance. Unlike standard demographic features, behavioral metrics often require derived representations to align with their theoretical meaning (e.g., slope of portion response, bite rate, active eating profiles).

### 2.2 Behavioral Response Slopes and Intercepts

Several features were engineered by modeling a child’s response across increasing portion sizes:

* **Bite size slope**: Rate of change in bite size as portion increases, computed as grams consumed per number of bites at each portion size level.
* **Percent active eating slope**: Change in the ratio of active eating time to total meal duration across portion sizes.
* **Intercepts**: Represented baseline behavior at the smallest portion size (100%).
* **Food switching**: Averaged across all four portion size conditions to reflect typical behavior.
* **Bite rate**: Calculated as the average number of bites per minute across the four portion sizes.

Slopes were derived using linear regression across the 100%, 133%, 167%, and 200% portion size conditions. These slopes capture reactivity to portion size increases, while intercepts indicate baseline tendencies.

### 2.3 Wear-Time Adjusted Averages for Activity and Sleep

Actigraphy-derived features were calculated only for children with sufficient wear time:

* Data were included if the child had at least **3 valid days** of wear data
* A day was considered valid if **wear time ≥ 600 minutes (10 hours)**

From these criteria, the following weekly averages were computed:

* **Mean MVPA**: Average minutes per day in moderate-to-vigorous physical activity
* **Mean sedentary time**: Average minutes per day spent inactive or sitting
* **Average sleep fragmentation index**: Mean of the nightly fragmentation scores across valid days

These thresholds follow established practices in pediatric actigraphy research and help ensure that derived features reflect typical behavior rather than sporadic or non-compliant tracking.

### 2.4 Age-Sensitive Adjustments

While most children in the cohort were of similar age (7–9 years), small variations in age or pubertal status can influence behavioral and cognitive measures. Some psychometric tools were interpreted with age-standardized scores (e.g., BRIEF-2), and sensitivity analyses were conducted to assess whether inclusion of raw age or pubertal score improved prediction.

Age was also included as a covariate in later modeling stages and considered when assessing whether engineered features were developmentally plausible (e.g., extreme bite rates in older vs. younger children). Refer back to **Table 1** for features and corresponding feature engineering.

These engineered features were added to the working dataset and used in both exploratory and model training phases.

Feature engineering decisions were guided by existing literature on child eating behaviors and domain expert input. Emphasis was placed on maintaining interpretability and aligning with theoretical constructs of appetite, responsiveness, satiety regulation, and movement behaviors in middle childhood.

## Step 3: Missing Data Analysis

### 3.1 Rationale

Missing data are a routine and often underestimated challenge in pediatric health datasets, especially when using behavioral or caregiver-reported features. In this tutorial, we begin with a structured approach to diagnosing missingness across all variables before applying appropriate imputation strategies. Understanding the extent and pattern of missingness is essential for minimizing bias and preserving the generalizability of predictive models.

### 3.2 Diagnostic Steps

We followed a multi-step diagnostic process:

* Calculated the proportion of missing data for each feature.
* Categorized features based on missingness levels (e.g., low <5%, moderate 5–15%, high >15%).
* Generated a bar plot of missingness percentage across features.
* Created a missingness heatmap to assess patterns across cases and features.

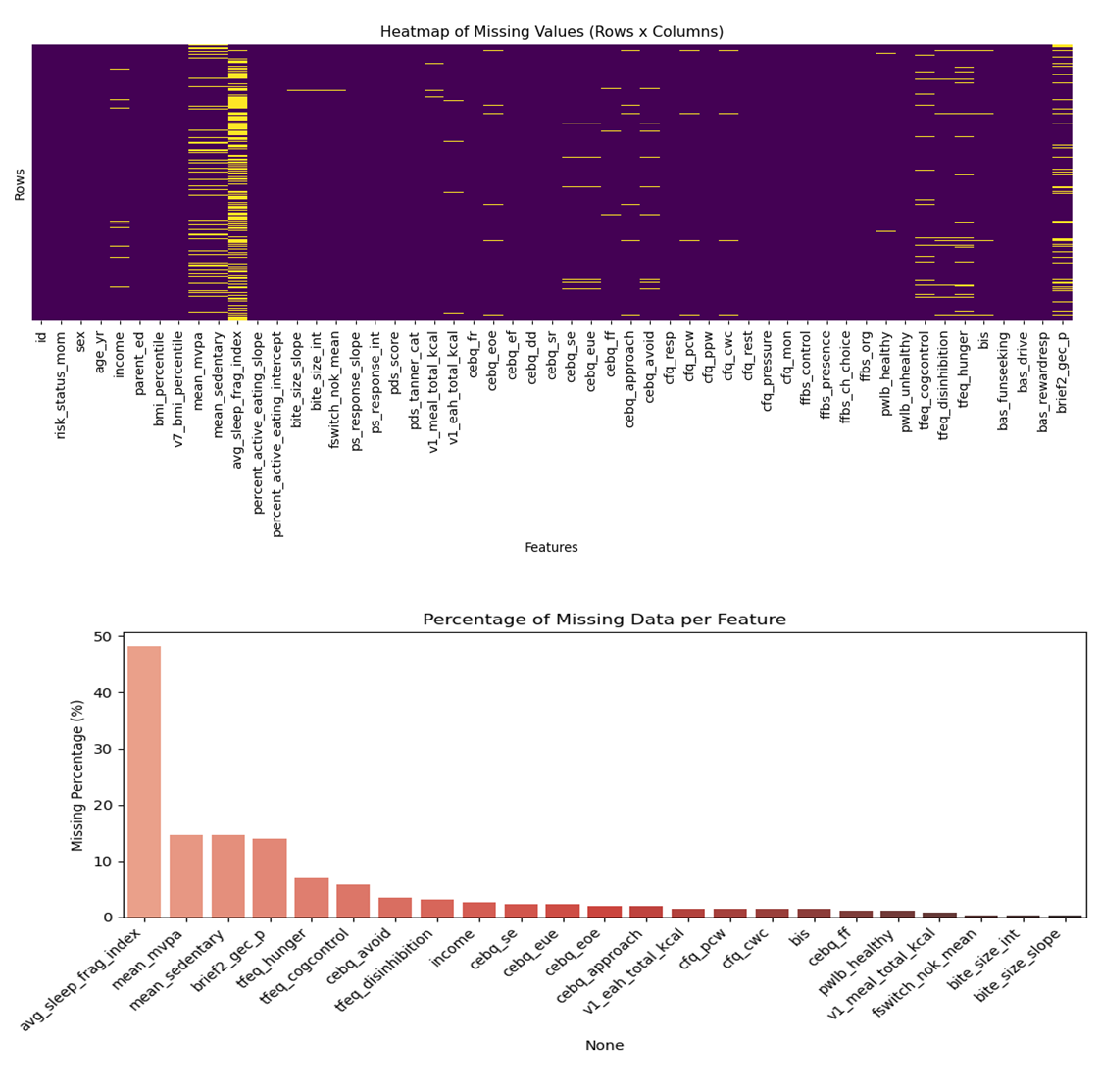
These steps enabled preliminary identification of features that may be missing completely at random (MCAR), at random (MAR), or not at random (MNAR).

### 3.3 Observed Patterns

* Most features had low (<5%) to moderate (5–15%) missingness.
* Some key variables, including avg\_sleep\_frag\_index, had >40% missingness, raising concerns about data reliability and potential exclusion.
* Behavioral and psychological measures showed clustered missingness within certain cases, consistent with incomplete survey responses.

These patterns suggest a combination of MCAR and MAR mechanisms, warranting variable-specific imputation strategies (see **Table 2** and **Figure 3**).

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 2.** Missingness Summary | | |  |
| Variable | Missing Count | Missing % | Strategy to address missingness |
| avg\_sleep\_frag\_index | 166 | 48.26 | Excluded for modeling |
| mean\_mvpa | 50 | 14.53 | kNN impute |
| mean\_sedentary | 50 | 14.53 | kNN impute |
| brief2\_gec\_p | 48 | 13.95 | kNN impute |
| tfeq\_hunger | 24 | 6.98 | kNN impute |
| tfeq\_cogcontrol | 20 | 5.81 | kNN impute |
| cebq\_avoid | 12 | 3.49 | median impute |
| tfeq\_disinhibition | 11 | 3.20 | median impute |
| income | 9 | 2.62 | mode impute |
| cebq\_se | 8 | 2.33 | median impute |
| cebq\_eue | 8 | 2.33 | median impute |
| cebq\_eoe | 7 | 2.04 | median impute |
| cebq\_approach | 7 | 2.04 | median impute |
| v1\_eah\_total\_kcal | 5 | 1.45 | median impute |
| cfq\_pcw | 5 | 1.45 | median impute |
| cfq\_cwc | 5 | 1.45 | median impute |
| bis | 5 | 1.45 | median impute |
| cebq\_ff | 4 | 1.16 | median impute |
| pwlb\_healthy | 4 | 1.16 | median impute |
| v1\_meal\_total\_kcal | 3 | 0.87 | median impute |
| fswitch\_nok\_mean | 1 | 0.29 | median impute |
| bite\_size\_int | 1 | 0.29 | median impute |
| bite\_size\_slope | 1 | 0.29 | median impute |



**Figure 3:** Missingness in variables of dataset

### 3.4 Strategy and Justification

Based on the diagnostic results, our imputation strategy was tiered by feature type and missingness level (see **Table 2**):

* **Low missingness (<5%)**: median or mode imputation, depending on variable type.
* **Moderate missingness (5–15%)**: K-nearest neighbors (KNN) imputation for continuous variables.
* **High missingness (>30%)**: flagged for review; some were excluded due to unreliability of imputation.

This approach balances model performance, transparency, and developmental plausibility. It is particularly suitable for pediatric datasets where blanket strategies (e.g., listwise deletion or mean imputation) can bias results or exclude meaningful behavioral data.

Each decision was grounded in theory and practicality, reinforcing the principle that preprocessing steps should be informed by both statistical logic and domain-specific understanding of the data.

## Step 4: Outlier Detection and Handling

### 4.1 Why Outlier Handling Matters in Pediatric Data

In pediatric datasets, outliers can arise from measurement error, data entry mistakes, or genuine but rare behaviors. Due to small-to-moderate sample sizes and age-sensitive variability in behavior and physiology, even a few extreme values can disproportionately affect model estimates. Therefore, outlier detection and mitigation are important for maintaining stability and interpretability in downstream ML models.

Unlike in large epidemiological datasets, pediatric studies often rely on proxy reports, structured tasks, and small sample-derived behavioral metrics (e.g., bite size slope, meal duration). These measures can show high variability and may not conform to population-based expectations. As such, identifying and managing outliers requires both distributional checks and domain understanding.

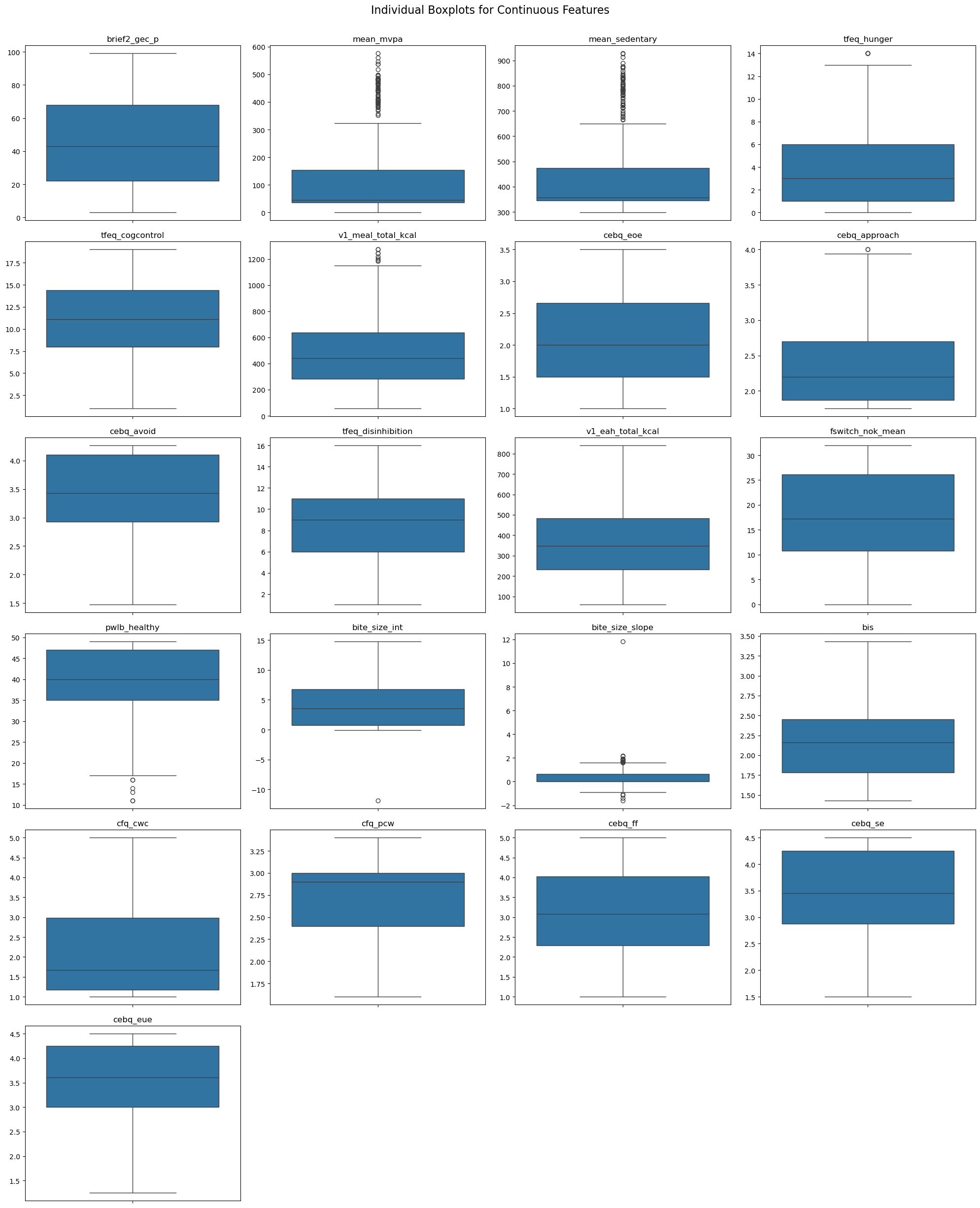
### 4.2 Visual Inspection as First-Line Tool

We used boxplots (**Figure 4**) to visually inspect the distribution of all continuous features. This approach allowed us to:

* Identify extreme values outside the whisker range (1.5 × IQR)
* Detect skewness that may not be biologically plausible
* Flag values that deviate markedly from the central tendency

Variables related to caloric intake, bite size, and physical activity (e.g., mean\_mvpa, v1\_meal\_total\_kcal, bite\_size\_slope) showed visible extreme values. These outliers may reflect true variability, measurement artifacts, or device wear-time issues (e.g., accelerometer compliance).

In contrast, most psychometric and questionnaire-derived scores (e.g., CEBQ, CFQ subscales) exhibited minimal outliers, likely due to bounded response scales and standardized instruments.



**Figure 4.** Boxplots of variables (continuous variables) to understand outliers

### 4.3 Winsorization Strategy

To manage outliers without excluding valuable data, we applied winsorization:

* Values below the 1st percentile and above the 99th percentile were capped at those percentile thresholds
* This was applied only to selected variables with high skew and outlier influence

Winsorization is appropriate in pediatric behavioral datasets where extreme values may arise from noise rather than signal. It preserves sample size while reducing the undue influence of extreme points on model training. We avoided listwise deletion to prevent loss of rare but meaningful cases.

|  |  |  |
| --- | --- | --- |
| **Table 3. Outlier addressing Strategy** | | |
| **Variable Category** | **Typical Outlier Presence** | **Action Taken** |
| Physical activity (mean\_mvpa) | High right skew | Winsorized 1%/99% |
| Eating behavior slopes | Bi-directional outliers | Winsorized 1%/99% |
| Caloric intake (meal or EAH kcal intake) | Extreme upper values | Winsorized 1%/99% |
| Questionnaire scales | Minimal | Retained as-is |

This approach ensured that the dataset remained robust against modeling distortion while preserving underlying behavioral signal.

Outlier handling decisions were based not only on statistical thresholds but also on whether values were developmentally implausible or inconsistent with measurement protocols. The balance of domain knowledge and visual diagnostics guided our choices.

## Step 5: Data Splitting: Train and Test Split

Before model feature selection and model training, we performed a **stratified train-test split** based on sex to ensure proportional representation of male and female participants in both sets. This is important in pediatric data given potential sex-related differences in growth trajectories, behavioral patterns, and BMI percentile distributions.

* **Split ratio**: 80% train, 20% test
* **Stratification variable**: sex
* **Random seed**: 42 (for reproducibility for split datasets)
* **Leakage prevention**: All feature selection, hyperparameter tuning, and imputation were performed **within** the training set only. The test set was held out and untouched until final model evaluation.

This procedure ensures an unbiased estimate of model performance and guards against information leakage across data partitions, especially when done before feature selection.

## Step 6: Feature Selection

### 6.1 Necessity of Feature Selection

Feature selection reduces dimensionality, improves model interpretability, and decreases the risk of overfitting especially important in pediatric datasets that often have small sample sizes and highly correlated behavioral variables. Many predictors in childhood obesity research are theoretically relevant but statistically redundant. Without proper selection, models risk learning noise rather than signal.

In this dataset, feature selection was especially important because of the overlap across multiple domains including behavioral slopes, caregiver reports, activity metrics, and psychological assessments. Many variables may individually show weak associations with the outcome but contribute meaningfully through interactions. Univariate filtering alone would be insufficient.

### 6.2 Stepwise Selection Strategy

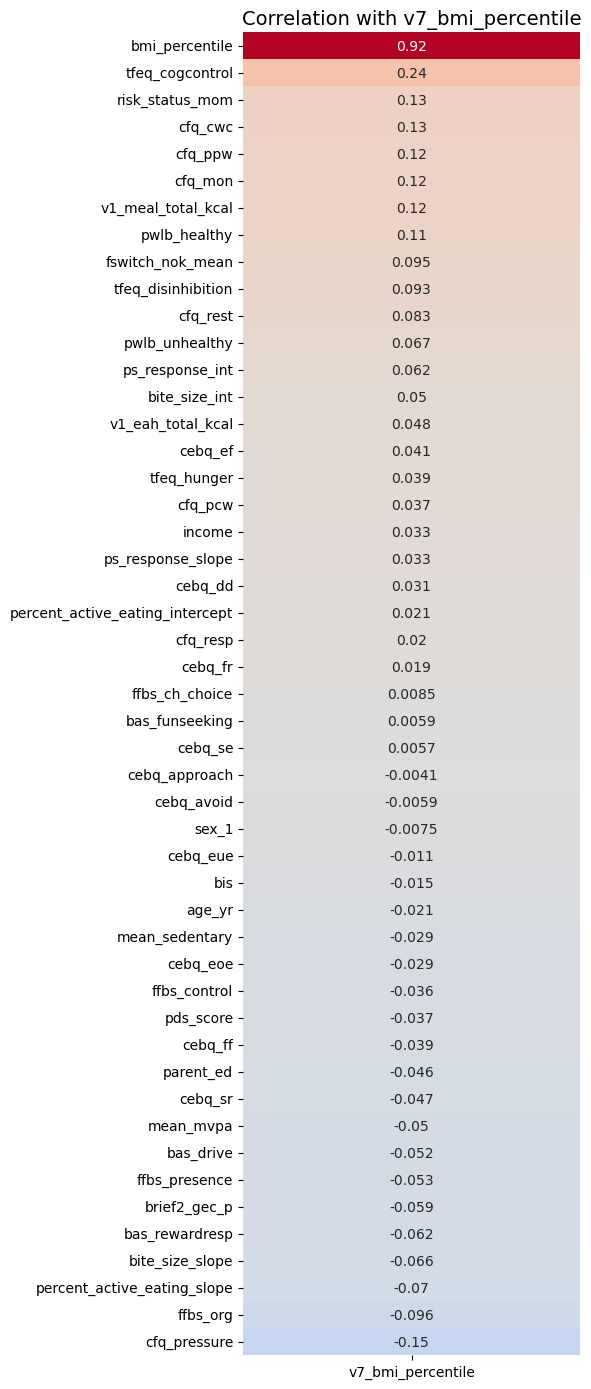
We implemented a two-stage process for identifying the most predictive features.

**Stage 1: Correlation Screening** We began with univariate correlation analysis to identify predictors with zero or minimal relationship to the outcome (v7\_bmi\_percentile). While not used for formal selection, this helped deprioritize clearly uninformative variables (see **Figure 5**). As expected, baseline BMI percentile showed the strongest correlation (r = 0.92), while most other features demonstrated weak or negligible linear associations, reinforcing the need for multivariate, model-based selection approaches.

**Stage 2: RFE-CV + Nested Cross-Validation** We then conducted model-dependent selection using Recursive Feature Elimination with Cross-Validation (RFE-CV), nested within an outer cross-validation loop. This allowed joint optimization of both the feature subset and the model hyperparameters. Each outer fold used the best-performing feature count and configuration from the inner CV.

This approach allowed the algorithm to:

* Reduce variance in feature selection
* Tune model complexity in alignment with predictor strength
* Avoid information leakage between tuning and evaluation phases



**Figure 5.** Univariate Correlations with outcome (follow-up BMI%)

### 6.3 Why RFE-CV vs Standard RFE?

Unlike standard RFE, which evaluates features on a single data split, **RFE-CV** ranks features based on cross-validated model performance. This reduces the likelihood of selecting predictors that perform well only by chance or within a specific fold. The nesting structure further allowed feature selection and hyperparameter tuning to be optimized jointly.

* **Inner loop**: Used RFE-CV to identify the most predictive subset of features within each fold.
* **Outer loop**: Evaluated model performance across a range of hyperparameter settings using the selected features.

This approach helped:

* Identify stable predictors that generalize across folds
* Avoid bias from tuning on the same data used for selection
* Capture interdependencies between model structure and feature informativeness

### 6.4 Summary of Feature Selection Results

Using nested cross-validation combining Recursive Feature Elimination with Cross-Validation (RFE-CV) and joint hyperparameter tuning, we identified distinct predictive subsets under two modeling scenarios: with and without baseline BMI percentile.

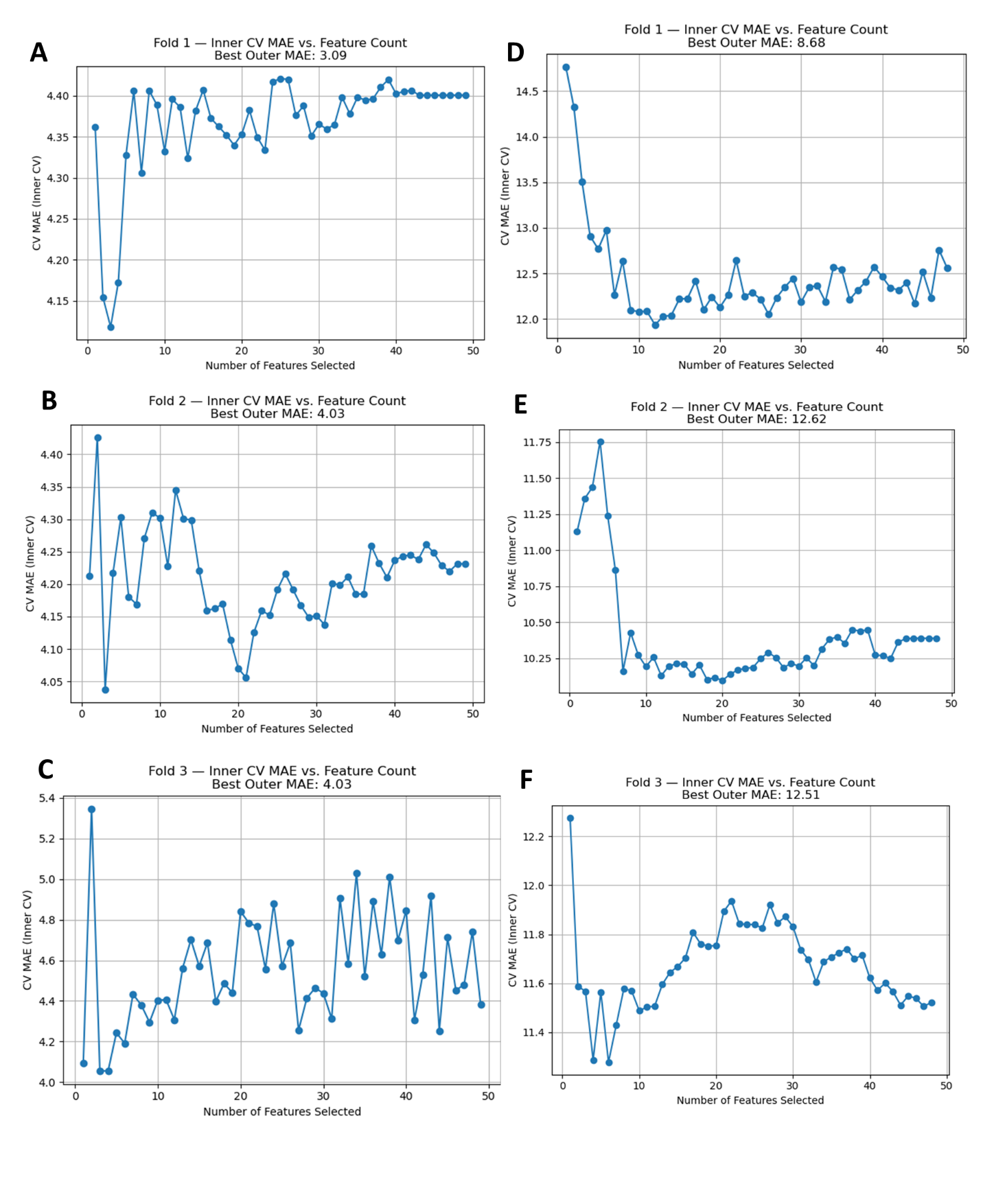
**With Baseline BMI Percentile**

Models achieved substantially lower mean absolute errors (MAEs), with the best outer fold MAE of 3.09. Across folds, bmi\_percentile and mean\_mvpa consistently appeared, underscoring their dominant role in short-term BMI prediction. A third feature, either bas\_drive, mean\_sedentary, or v1\_meal\_total\_kcal, varied by fold, suggesting smaller but potentially complementary effects. MAE curves plateaued after 3 to 5 features, supporting a compact model.

**Without Baseline BMI Percentile**

When baseline BMI was excluded to emphasize modifiable behavioral and contextual predictors, prediction error increased (best outer fold MAE = 8.68). However, selection revealed several interpretable risk factors, including mean\_mvpa, mean\_sedentary, percent\_active\_eating\_slope, caregiver-reported feeding practices (cfq\_ppw, cfq\_cwc), and child eating behavior traits (cebq\_dd, cebq\_approach). Feature stability was lower across folds, consistent with noisier signal in the absence of BMI anchoring.

These results indicate that while baseline BMI percentile drives predictive accuracy, the feature selection process without it highlights a broader set of behavioral and caregiver-report variables relevant to early risk profiling.



**Figure 6.** Inner Cross-Validation MAE vs. Number of Features Selected Across Folds (With and Without Baseline BMI Percentile). A–C show inner CV mean absolute error (MAE) as a function of feature count for each outer fold when baseline BMI percentile was included. D–F show the same for models excluding baseline BMI. Each point represents the average MAE from the inner loop at a given number of selected features. The lowest outer fold MAE and corresponding feature subset are annotated above each panel. These curves illustrate the diminishing returns of adding features and support parsimonious selection strategies, especially when baseline BMI is present.

## 7. Model Training

After completing feature and hyperparameter selection in the previous section, final models were trained using the selected features and corresponding optimal configurations. All training was conducted on the full training set, using only features and parameters identified through nested cross-validation to avoid information leakage and overfitting.

### 7.1 Separate Training Pipelines

To address different prediction goals and interpretability needs in pediatric obesity research, three distinct modeling pipelines were implemented:

1. **BMI-Based Model:** Included baseline BMI percentile
   * **Hypothesis:** Baseline BMI is the strongest available predictor of short-term weight status.
   * **Utility:** Optimized for **clinical risk tracking** in children already presenting with elevated or at-risk BMI. This model prioritizes predictive accuracy and is well-suited for near-term forecasting in clinical settings.
2. **Behavioral-Only Model:** Excluded baseline BMI percentile
   * **Hypothesis:** Modifiable behavioral, psychological, and environmental predictors alone can provide signal for early obesity risk, even without current BMI status.
   * **Utility:** Designed for **early identification and prevention**, particularly in children not yet exhibiting excess weight but potentially on a higher-risk trajectory based on their behavioral profiles. This model emphasizes **intervention-relevant features**.
3. **Combined Model:** Included baseline BMI percentile along with all selected behavioral and environmental features
   * **Hypothesis:** Adding behavioral context to baseline BMI improves prediction and interpretability.
   * **Utility:** Intended for use in **integrated clinical-preventive decision-making**, where both anthropometric status and modifiable risks are needed. This model may support tailored feedback or joint tracking of weight and behavior over time.

Each scenario was treated independently throughout model development. Separate pipelines ensured that the training process, including feature selection and parameter tuning, was appropriately tailored to each model’s structure and intended use case. This allowed for a fair comparison of model performance under different assumptions about available data and intervention goals.

## 8. Model Evaluation and Feature Evaluation

### 8.1 Model Evaluation Metrics

## MAE (Mean Absolute Error)

## RMSE (Root Mean Squared Error)

## R² (Coefficient of Determination)

### 8.2 Performance Results

Table 4presents evaluation of the three models on test set.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 4. Model evaluation** | | | |
| **Model Type** | **MAE** | **RMSE** | **R²** |
| BMI-Based Model | 2.82 | 4.65 | 0.87 |
| Behavioral-Only Model | 8.16 | 11.68 | 0.20 |
| Combined Model | 2.91 | 4.72 | 0.86 |

### 8.3 Interpretation and Practical Utility

1. **BMI-Based Model** demonstrated the highest predictive performance across all metrics, with an R² of 0.87 and the lowest error rates (MAE = 2.82, RMSE = 4.65). This is expected due to the inclusion of baseline BMI percentile, which is a strong statistical predictor of future BMI percentile.

**Practical utility**:

* + This model is well-suited for **clinical triage and risk stratification**, especially in settings where prior growth data is readily available.
  + It can be used to identify children at elevated risk for continued or worsening obesity, guiding early intervention decisions.
  + However, its heavy reliance on a single dominant predictor limits interpretability regarding why risk is elevated, which may be a barrier in family-centered care or behavioral counseling contexts.

1. **Behavioral-Only Model** produced substantially lower predictive accuracy (R² = 0.20), with wider prediction errors (MAE = 8.16, RMSE = 11.68). Despite this, it surfaced key behavioral features such as food responsiveness, bite rate, and active eating duration, which contributed meaningfully to variance explained.

**Practical utility**:

* + This model is most appropriate in settings where **baseline BMI data are unavailable or unreliable**, such as in mobile health tools, screening during community outreach, or retrospective risk assessment using behavioral surveys.
  + It may also serve as a **hypothesis-generating tool** for identifying behavioral phenotypes or informing the design of targeted behavioral interventions.
  + While not accurate enough for clinical prediction alone, it provides insight into potentially modifiable risk pathways that can be the focus of family-based education or coaching.

1. **Combined Model** achieved nearly equivalent performance to the BMI-Based Model (R² = 0.86, MAE = 2.91, RMSE = 4.72), while also incorporating behaviorally and contextually relevant predictors.

**Practical utility**:

* + This model provides a **balanced solution for interdisciplinary applications** where both predictive accuracy and interpretability are needed.
  + For instance, it could be embedded in **decision support systems** to inform clinicians of both risk level and contributing behavioral factors, prompting more personalized counseling.
  + It allows **caregivers to understand specific drivers of their child’s weight trajectory**, facilitating shared decision-making and engagement.
  + It is particularly useful in **intervention monitoring**, where clinicians may want to track behavioral changes over time and observe their influence on predicted risk even when BMI percentile remains stable.
  + In research settings, this model supports **mechanistic inquiry** into how behavioral patterns and baseline BMI interact to shape obesity risk over time.
  + Overall, the Combined Model offers the most clinically actionable format, maintaining strong performance while also enhancing transparency. It helps bridge the gap between statistical modeling and real-world pediatric care, where understanding "how" and "why" are often as critical as "what."

### 8.4 SHAP-Based Model Interpretation

To examine how individual features contributed to predictions, we used SHAP (SHapley Additive exPlanations), which assigns each prediction a set of feature attributions based on their marginal impact.

For each model, we generated both **SHAP bar plots (Figure 7)** and **beeswarm plots (Figure 8)**:

* The **bar plot** summarizes the average absolute contribution of each feature to model predictions across all participants.
* The **beeswarm plot** visualizes the full distribution of SHAP values per feature, with each point representing an individual observation. Point color reflects the original feature value (red = high, blue = low), and horizontal position indicates the direction and strength of the feature’s impact.

These plots are useful for identifying both **overall feature importance** and **how individual-level variation in predictors affects predictions**, which is particularly valuable in pediatric models where behavioral patterns and clinical indicators often vary substantially between individuals.

**Clinical and Research Utility**

* SHAP beeswarm plots offer **individual-level transparency** in how the model generates predictions.
* They enable **personalized explanation** of risk, which can support:
  + Tailored family feedback
  + Targeted behavioral interventions
  + Clinician confidence in model outputs
* These interpretations also support **mechanistic understanding** in research, informing future hypothesis testing or feature design.

### 1. BMI-Based Model

• **Key observations/Top Features**:

* bmi\_percentile
* mean\_mvpa
* bas\_drive

• **Interpretation from SHAP bar and beeswarm plots**:

* bmi\_percentile: As expected, dominant feature with a wide SHAP value range. Higher prior BMI strongly predicted future BMI percentile.
* mean\_mvpa: Maintains a modest negative effect, but impact is secondary to baseline BMI.
* bas\_drive: Positively associated with increased predicted BMI percentile, indicating that reward sensitivity remains an independent contributor.

• **Theoretical context**:  
This model captures trajectory-like prediction, emphasizing continuity in BMI. It mirrors traditional clinical practice, where BMI tracking is central to assessing growth risk.

• **Clinical utility and relevance**:  
This model provides the highest accuracy and is clinically intuitive for tracking growth trends. It is best suited for pediatric follow-up, EMR integration, and screening when growth data is readily available. However, it does not inform modifiable causes of weight change and offers limited behavioral interpretability.

### 2. Behavioral-Only Model

• **Key observations/Top Features**:

* mean\_mvpa (physical activity)
* mean\_sedentary (sedentary time)
* tfeq\_cogcontrol (cognitive restraint)
* cfq\_cwc (parental concern about child weight)
* percent\_active\_eating\_slope

• **Interpretation from SHAP bar and beeswarm plots**:

* mean\_mvpa: Higher physical activity (red) had mostly negative SHAP values, indicating lower predicted BMI, consistent with protective effects shown in prior pediatric literature (Janssen & LeBlanc, 2010).
* mean\_sedentary: Higher sedentary time was positively associated with predicted BMI.
* tfeq\_cogcontrol and cfq\_cwc: Mixed SHAP values suggest individual variability in how parental beliefs and control impact risk.
* percent\_active\_eating\_slope: Greater active eating behavior predicted higher BMI, consistent with associations between food approach and intake regulation.

• **Theoretical context**:  
This model isolates modifiable behavioral and psychological mechanisms, offering insight into obesity risk without reliance on baseline anthropometrics. It supports a behavior-first interpretation of pediatric obesity risk.

• **Clinical utility and relevance**:  
While less predictive overall, this model is useful in behavioral counseling or prevention contexts, particularly where anthropometric data is unavailable or where the goal is to target changeable lifestyle behaviors. It supports the identification of at-risk children based on habits and family dynamics, relevant for early intervention or community programs.

### 3. Combined Model

• **Key observations/Top Features**:

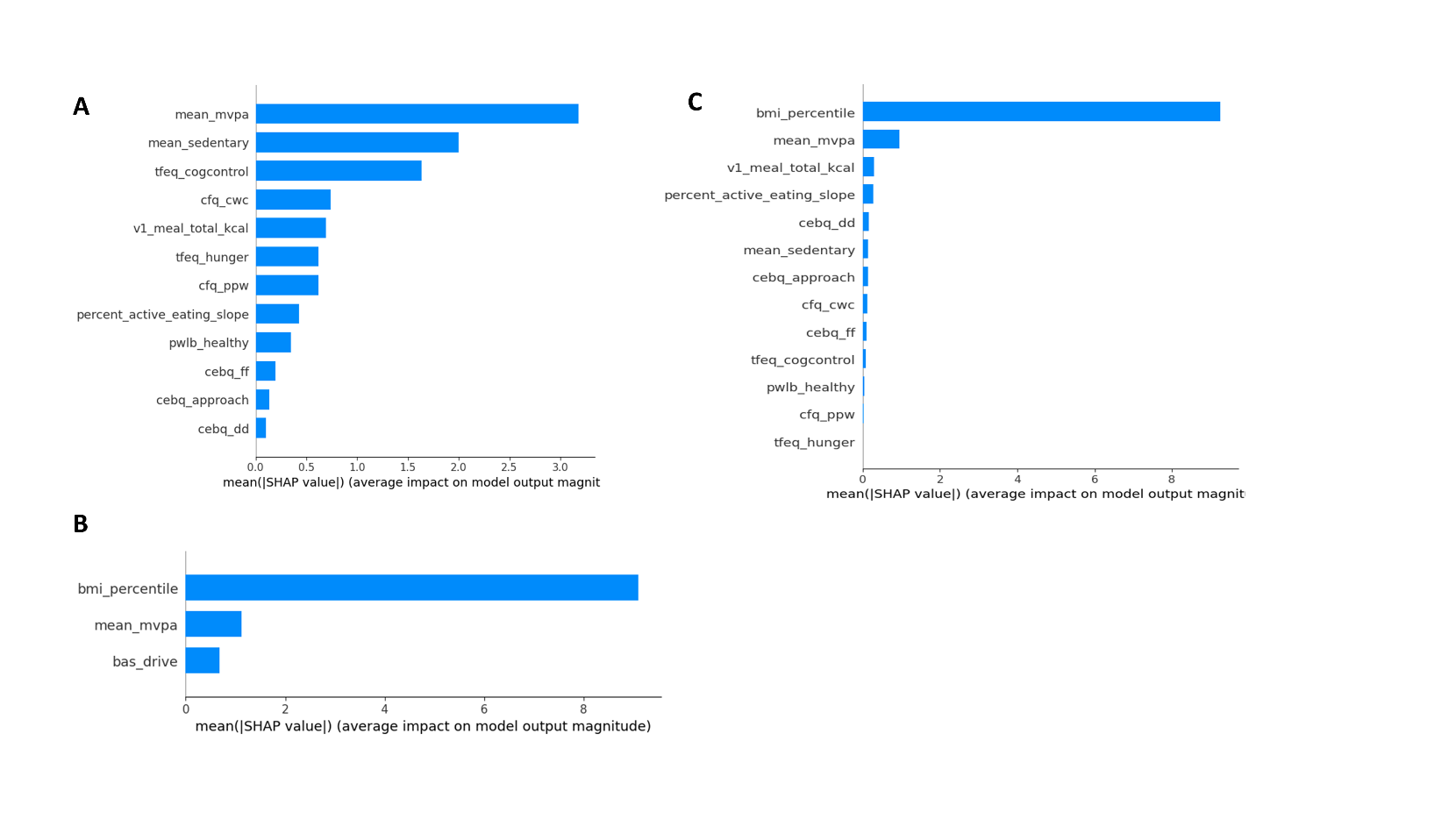
* bmi\_percentile remains dominant
* Behavioral contributors such as mean\_mvpa, v1\_meal\_total\_kcal, and percent\_active\_eating\_slope have visible SHAP contributions

• **Beeswarm interpretation**:

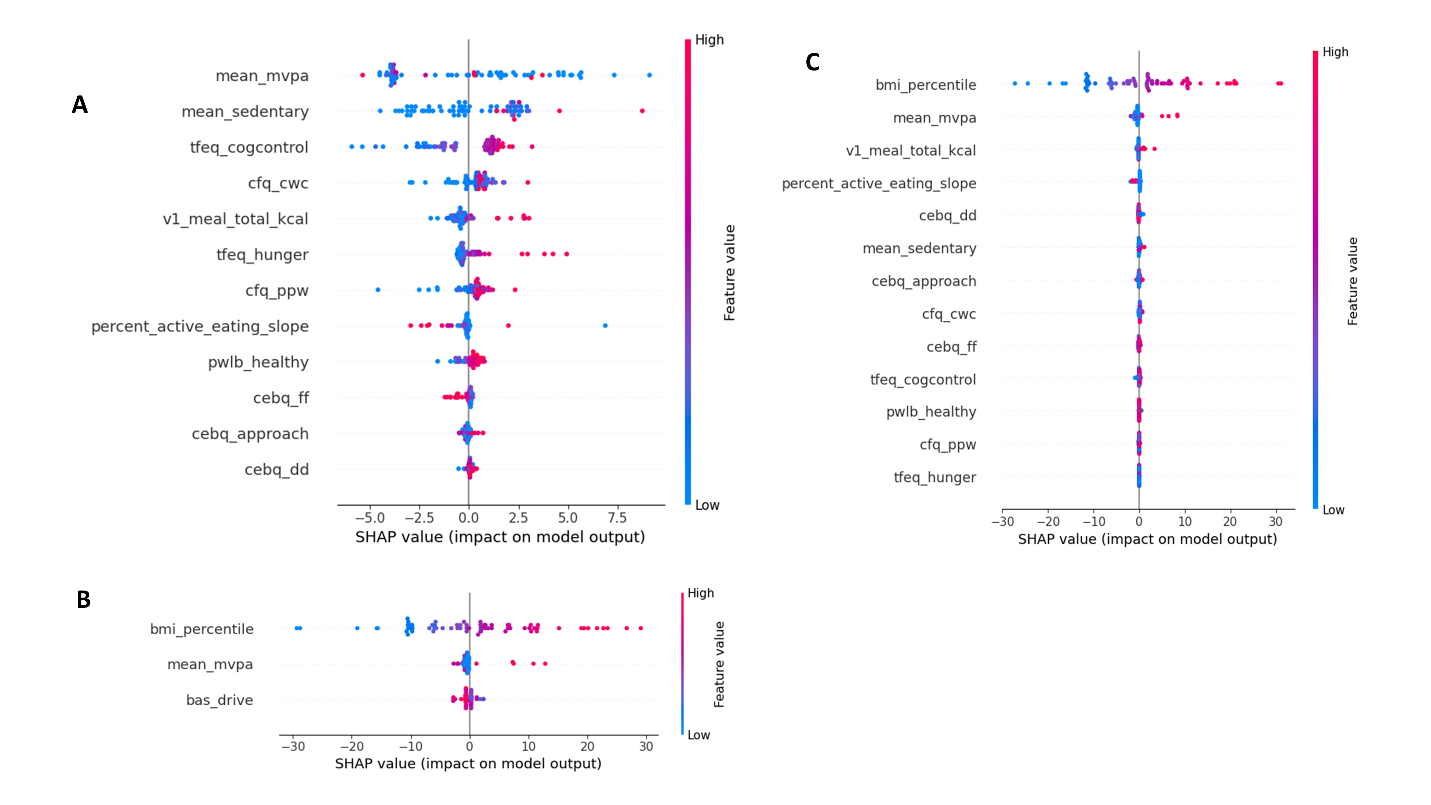
* bmi\_percentile retains wide SHAP spread but no longer monopolizes the prediction.
* mean\_mvpa and v1\_meal\_total\_kcal show moderate contributions, reflecting a balance between clinical history and behavioral risk.
* Shared influence allows more granular interpretation of why a child is at risk—past BMI or current behaviors.

• **Theoretical context**:  
This model aligns with socioecological frameworks by integrating structural (BMI trajectory) and behavioral inputs. It better reflects real-world pediatric assessments, which must synthesize multiple risk dimensions.

• **Clinical utility and relevance**:  
This model is most actionable for pediatric care teams who aim to provide both risk prediction and guidance on modifiable contributors. It supports personalized recommendations by allowing BMI to provide predictive power while behavioral data contextualizes why a child is at risk. Ideal for clinical counseling, family discussions, or integrated care plans.



**Figure 7. SHAP Summary Bar Plots of Feature Importance Across Model Variants.** Bar plots show the mean absolute SHAP value for each predictor, reflecting average contribution to model output magnitude. A) behavioral-only model, where mean\_mvpa, mean\_sedentary, and tfeq\_cogcontrol emerge as the strongest contributors. B) shows the BMI-based model, where baseline bmi\_percentile dominates prediction. C) displays the combined model, where bmi\_percentile remains influential but is accompanied by behavioral features such as mean\_mvpa and v1\_meal\_total\_kcal. This comparison highlights the relative dominance of clinical versus behavioral predictors across different modeling choices.



**Figure 8.** **SHAP Beeswarm Plots for Individual Feature Contributions.** Each point represents an individual prediction, with SHAP value (x-axis) indicating feature contribution and color reflecting feature value (red = high, blue = low). A) (behavioral-only) reveals interpretable trends such as higher mean\_mvpa lowering BMI predictions and higher mean\_sedentary increasing risk. B) (BMI-based) shows bmi\_percentile as the primary driver with wide variation. C) (combined) integrates both behavioral and clinical features, supporting nuanced interpretation of prediction sources. These visualizations demonstrate not only which features are important, but also how their values influence predicted BMI percentile.

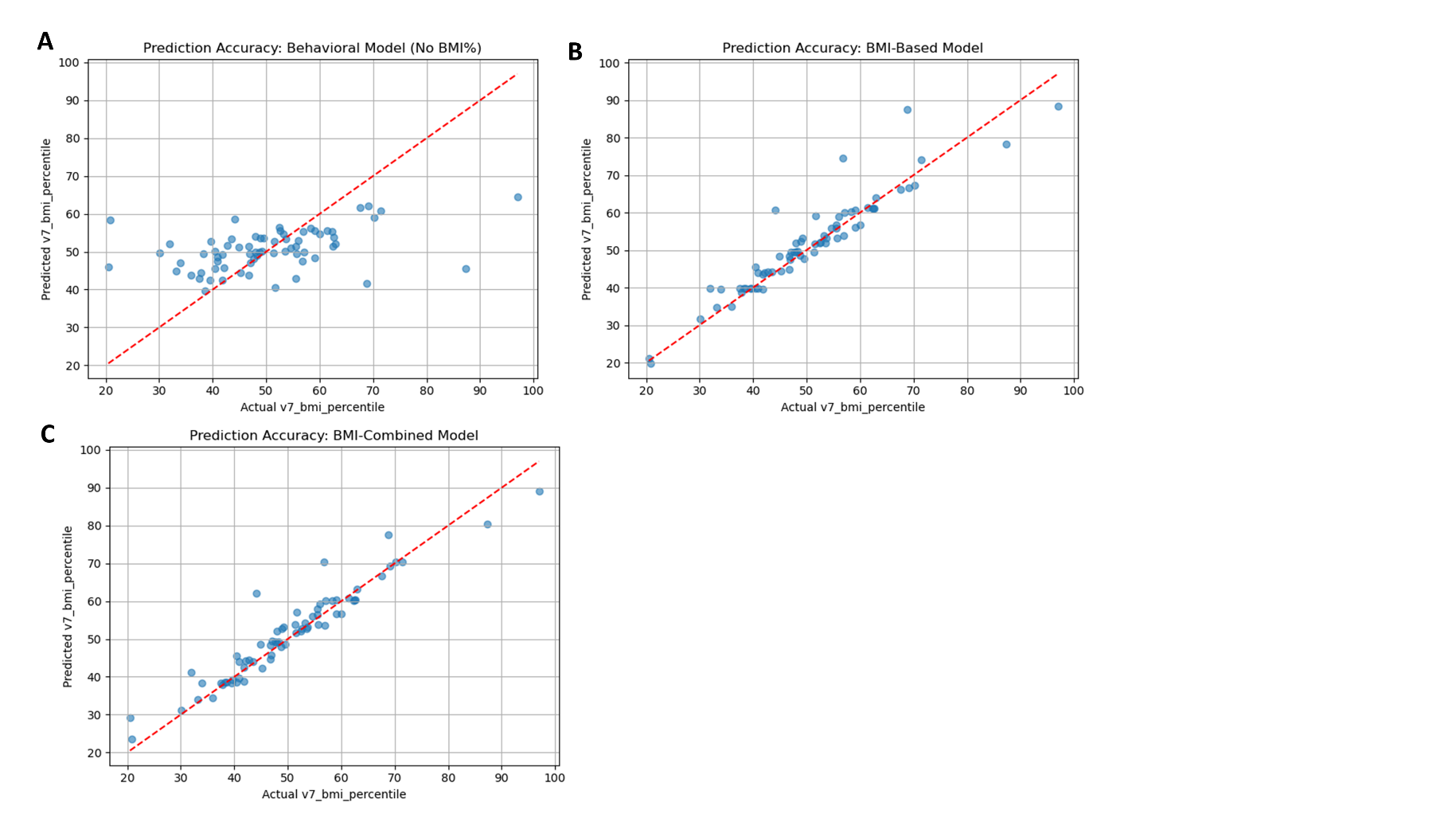
**Summary Comparison:**

| **Table 5. Summary of Model Performance** | | | |
| --- | --- | --- | --- |
| **Model** | **Strengths** | **Clinical Use Case** | **Limitations** |
| **Behavioral-Only** | Captures modifiable risk; aligns with behavioral theory | Prevention, behavior change, research trials | Lower predictive accuracy |
| **BMI-Based** | High accuracy; trajectory-based prediction | Growth screening, EMR use | Lacks behavioral interpretability and no possible intervention |
| **Combined** | Balanced accuracy and interpretability | Pediatric care, counseling, family engagement | Requires both clinical and behavioral data inputs |

### 8.5 Prediction Accuracy and Residual Patterns

## Prediction vs. Actual BMI Percentile Plots (Figure 9):

BMI-Based and Combined Models showed tight clustering around the identity line, indicating strong agreement with true outcomes. Behavioral-Only Model predictions were more dispersed, reflecting reduced precision, especially for tail -end – lower and higher end of the actual BMI percentiles -

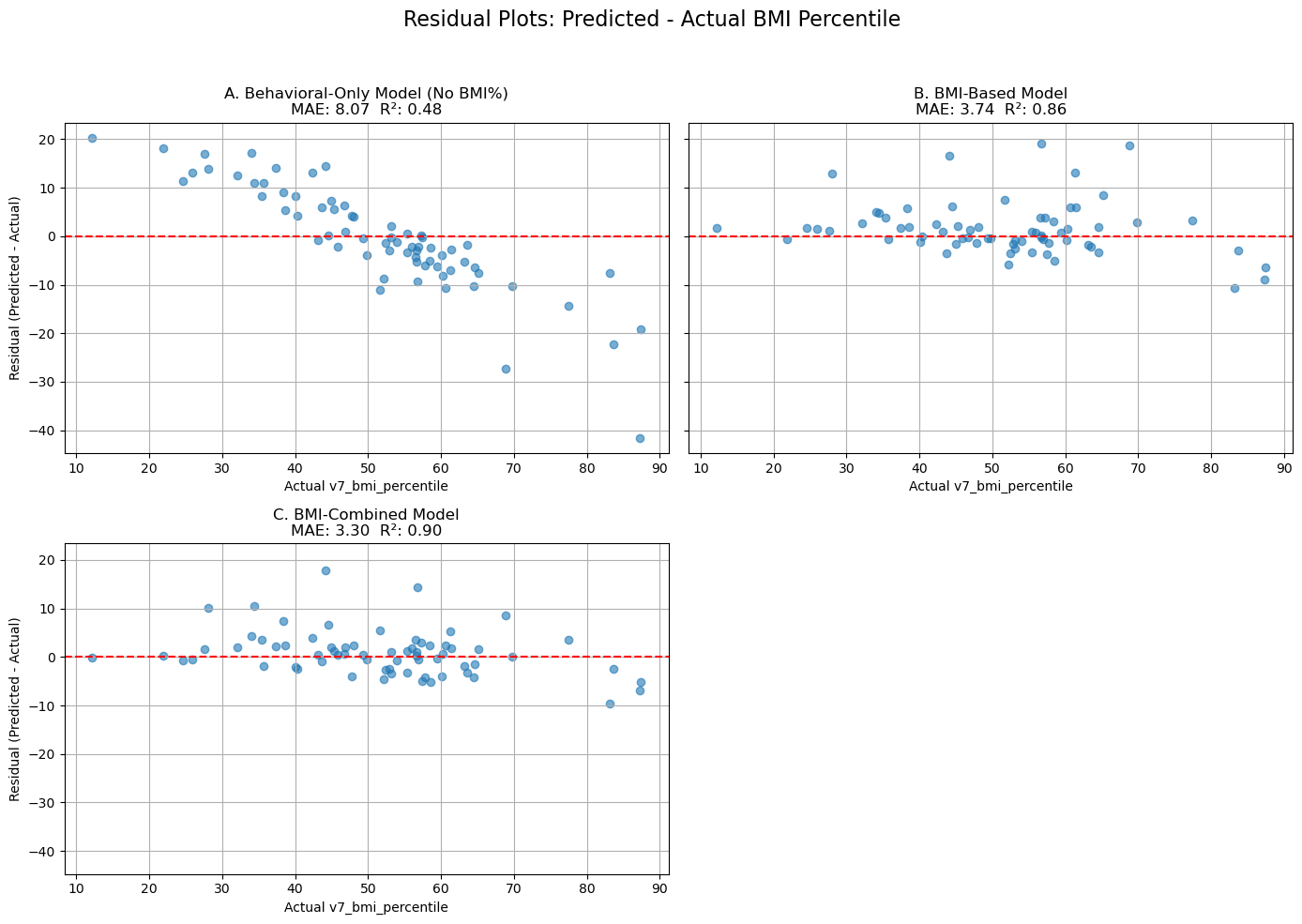


**Figure 9.** **Prediction Accuracy Across Model Variants (Actual vs Predicted BMI Percentile at Follow-Up).**

Scatter plots compare predicted versus actual BMI percentile at follow-up (v7) for each model. The red dashed line indicates perfect prediction (red line). **A)** (Behavioral-Only Model): Shows moderate correlation between predicted and actual values, with wider dispersion. This reflects limited predictive power when baseline BMI is excluded. **B)** (BMI-Based Model): Displays tight alignment along the diagonal, showing strong predictive performance largely driven by baseline BMI percentile. C) (Combined Model): Maintains strong predictive accuracy while also incorporating behavioral inputs, narrowing residuals and enhancing interpretability. Together, these highlight the trade-off between predictive strength and behavioral interpretability across modeling strategies.

**Residual Plots (Figure 10):**

* **Behavioral-Only Model:**  
  Residuals show substantial spread and a non-random pattern, especially across different BMI ranges. This implies that the model struggles to generalize, particularly for children at the extremes of the BMI distribution.
* **BMI-Based Model:**  
  Inclusion of baseline BMI percentile dramatically improves predictive stability. Residuals are smaller and distributed more evenly, with fewer extreme outliers. MAE is reduced (3.74), and model fit improves substantially (R² = 0.86).
* **Combined Model:**  
  Shows the best overall performance. Residuals are tightly clustered, indicating stable performance across BMI levels. It achieves the lowest MAE (3.30) and highest R² (0.90), confirming that behavioral features add incremental predictive value when layered on top of growth history.



**Figure 9.** **Residual Plots for All Models: Predicted - Actual BMI Percentile.** Each panel shows the residuals (Predicted - Actual BMI percentile at follow-up) plotted against the true BMI percentile values for each model. A) Behavioral-Only Model: Residuals display a clear pattern of underprediction at high BMI percentiles and overprediction at low percentiles, indicating systematic error and lower model calibration. B) BMI-Based Model: Residuals are smaller and more homoscedastic, suggesting improved model fit and calibration, largely due to the inclusion of baseline BMI. C) BMI-Combined Model: Residuals are smallest and most centered around zero, showing the best performance across the range of actual BMI percentiles. This model combines historical growth (BMI%) with behavioral data, improving both accuracy and generalizability.

## Step 9: Conclusion and Discussion

### 9.1 Summary of Key Findings

This tutorial demonstrated the application of machine learning models to predict follow-up BMI percentile in a pediatric cohort using behavioral, psychological, and anthropometric data. We trained and evaluated three models:

* A high-performing BMI-Based Model driven largely by baseline BMI percentile
* A behaviorally focused model that emphasized interpretability but had limited predictive strength
* A Combined Model that retained most of the accuracy of the BMI-Based Model while enhancing behavioral interpretability

Feature selection and hyperparameter tuning were conducted jointly to reflect the interdependence of variable importance and model structure. SHAP analysis provided insight into feature contributions and supported model transparency.

### 9.2 Interpretation and Use Cases

* The **BMI-Based Model** is best suited for clinical scenarios where baseline BMI is routinely available. Its high accuracy makes it ideal for tracking risk over short-term follow-up.
* The **Behavioral-Only Model** provides insight into modifiable factors that may drive future weight gain. It may be appropriate in community or early intervention contexts where prior anthropometric data are unavailable.
* The **Combined Model** offers a compromise between accuracy and interpretability, useful in interdisciplinary teams aiming to balance clinical utility with behaviorally actionable insights.

### 9.3 Cautions and Limitations

Several limitations should be acknowledged:

* **Short follow-up period**: Limits the variability in BMI change, reducing the signal for true behavioral predictors.
* **Sample size**: As is common in pediatric research, the moderate sample size constrains model complexity and may reduce generalizability.
* **Proxy-reported and task-derived data**: While ecologically valid, these variables may have inconsistent reliability across children and require cautious interpretation.
* **Synthetic dataset**: Although modeled after real data, the tutorial used a de-identified synthetic dataset. Performance on real clinical data may differ.

Models that include baseline BMI risk overstating their performance due to autocorrelation. Researchers must be cautious when interpreting models dominated by anthropometric history, particularly if the goal is to uncover modifiable predictors.

### 9.4 Broader Implications

This tutorial highlights the need for pediatric ML pipelines that integrate:

* Domain knowledge in feature engineering
* Rigorous preprocessing for missingness and outliers
* Joint consideration of interpretability and accuracy
* Transparent model interpretation using tools like SHAP

Future work should explore longer-term prediction, external validation, and integration of additional contextual variables such as food environment, social dynamics, or digital phenotyping.

In conclusion, ML can enhance our understanding of childhood obesity risk, but its application must be developmentally grounded, ethically cautious, and interpreted in partnership with pediatric and behavioral experts.