Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Sleep duration and maternal BMI rank comparably as predictors of overweight/obesity among school-aged children and adolescents in central China: a cross-sectional study using interpretable machine learning

Yiran Wang
872578928@qq.com

Children's Hospital Affiliated to Zhengzhou University, Zhengzhou Children's Hospital
Yaodong Zhang
Children's Hospital Affiliated to Zhengzhou University, Zhengzhou Children's Hospital
Shuying Luo
Children's Hospital Affiliated to Zhengzhou University, Zhengzhou Children's Hospital
Kaijuan Wang
Zhengzhou University

Research Article

Keywords: Overweight/obesity, Sleep duration, Parental BMI, Machine learning, SHAP analysis, Cross-sectional study, Children and adolescents

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11452625/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background: Childhood overweight/obesity has a multifactorial aetiology that regression captures only partly. We used an interpretable machine-learning model to rank behavioural and familial predictors in a provincial school sample and to test whether sleep duration and maternal BMI differ in predictive importance.

Methods: Cross-sectional survey of 27588 children and adolescents aged 7-18 years in Henan, China, recruited by stratified multistage cluster sampling (20 city strata, 200 schools, 2000 classes). Weight status used the age- and sex-specific BMI criteria of the Working Group on Obesity in China; prevalence was weighted with post-stratification and regression used school-cluster-robust standard errors. Because weight status is the outcome, outcome-based propensity score matching served only as a secondary balance-restricting analysis, and the full cohort was primary. Random Forest, XGBoost and CatBoost were compared by nested cross-validation confined to the training data; splits were grouped by school with matched pairs kept together, and the locked model was evaluated once in an untouched test set. SHAP importance was summarised with 500-resample bootstrap intervals, and logistic regression excluding outcome-defining variables, with prespecified confounders and Benjamini-Hochberg control, provided association estimates and E-values.

Results: Prevalence was 27.3% (weighted 26.8%, 95% CI 26.2-27.4). CatBoost discriminated best (AUC 0.801, 95% CI 0.795-0.807; PR AUC 0.628) with a calibration slope of 1.03. SHAP ranked age first (0.163), then sleep duration (0.142) and maternal BMI (0.128); the difference was small and statistically indistinguishable (0.014, 95% CI, -0.006 to 0.034), and sleep duration ranked first of the two in 61.4% of resamples. Very short sleep duration (OR 3.28, 2.34-4.59), screen time >2 h/day (3.67, 2.42-5.58) and maternal obesity (2.87, 2.19-3.76) were associated with overweight/obesity; E-values ranged from 1.96 to 3.24.

Conclusions: Sleep duration and maternal BMI were comparably important and the ranking between them was unstable, so neither outranks the other; age, sleep duration and screen time were the leading predictors. These rankings support risk stratification rather than causal or intervention-priority claims.

1 Background

Childhood overweight and obesity have become a global public health emergency[1]. In China, prevalence among children and adolescents aged 7–18 years has risen steadily since 1985, with projections indicating urban-rural convergence around 2027 and rural areas becoming the future focus of prevention[2]. Childhood overweight/obesity is characterised by earlier onset and greater severity, with a growing number of school-aged children developing persistent obesity before puberty and thereby laying the groundwork for adult chronic disease burden[3]. Its consequences extend far beyond excess adiposity: overweight or obese children frequently present with dyslipidaemia, insulin resistance, non-alcoholic fatty liver disease and sleep apnoea syndrome, and approximately 41%-80% of childhood overweight/obesity persists into adulthood, increasing the risk of adult cardiometabolic disease, type 2 diabetes and several cancers; moreover, about 75% of youth-onset type 2 diabetes cases present with obesity at diagnosis[4].

The aetiological architecture of childhood overweight/obesity involves dynamic gene-environment interactions, and machine learning captures high-dimensional non-linear relationships that conventional regression models cannot represent[5]. A narrative review of 15 studies identified child weight history and parental overweight/obesity as key early risk factors[6]; a Brazilian two-year cross-lagged study showed that higher baseline screen time predicted shorter subsequent sleep duration[7], and a cohort of 7445 US adolescents found screen time and insufficient sleep to be jointly associated with higher BMI two years later[8]. A meta-analysis of 25 prospective cohorts (56584 participants) reported that the shortest sleepers were 76% more likely to become overweight or obese (OR 1.76, 95% CI 1.39–2.23)[9]. Interpretable machine learning has since been applied to comparable questions, including a national survey that combined LightGBM with SHAP (AUC 0.805) and an external validation study of 19024 children that ranked parental BMI, birth

length, sleep duration and birth weight among key predictors[10]; prediction models for clinically defined populations with sleep disorders as central exposures have likewise used grouped splitting, nested tuning and explicit calibration[11–15]. These precedents informed the variable set and the validation protocol adopted here.

Important gaps nonetheless constrain the evidence base for precision prevention. Most Chinese machine-learning studies do not address baseline covariate imbalance, SHAP-based rankings are seldom accompanied by uncertainty estimates[6, 10–12, 16], regional evidence from China's most populous central province remains scarce with complex survey designs often analysed as simple random samples[2], and the relative predictive importance of modifiable behaviours versus familial factors - in particular whether sleep duration rivals parental BMI - has not been examined in Chinese paediatric populations with uncertainty quantified[17, 18]. A ranking presented as decisive can steer prevention towards behavioural targets, whereas a statistically indistinguishable ranking should not.

We therefore conducted a design-aware, prediction-focused analysis in a large provincial sample with five aims: (a) to develop and internally as well as geographically validate an interpretable model of overweight/obesity using tuning confined to training data and an untouched test set; (b) to rank behavioural and familial predictors and to quantify the uncertainty of that ranking, including a formal comparison of sleep duration with maternal BMI; (c) to estimate associations using prespecified confounders that exclude the variables defining the outcome; (d) to examine how sampling weights, school clustering and stratification affect prevalence estimates, regression uncertainty and model evaluation; and (e) to quantify, with explicit assumptions, the robustness of the main associations to unmeasured confounding. Throughout, we treat the analysis as predictive and associational rather than causal, and we do not use propensity score matching to emulate a randomised comparison of any exposure.

2 Methods

2.1 Study design and sampling

This study was a cross-sectional school-based survey of children and adolescents in Henan Province, central China, conducted in a single survey wave. A stratified multistage cluster design was used. In the first stage, the province was divided into five geographic regions (eastern, western, central, southern and northern), and 20 cities were selected with probability proportional to size (PPS) from within each region, with urban-rural stratification, giving 20 sampling strata. In the second stage, 10 schools were selected per city by PPS from the municipal school list, stratified by school type (urban or rural; primary, junior secondary or senior secondary), giving 200 schools as primary sampling units. In the third stage, 10 classes were selected per school by simple random sampling from each grade stratum, giving 2000 classes. In the fourth stage, students were selected from each class roster by simple random sampling; because the field protocol restricted the number of students that could be examined per class within one school day, approximately 14 students were drawn per selected class rather than the entire class. This staged selection explains the average of 13.9 participants per class (27588/2000) and supersedes the earlier statement that all students in selected classes were invited.

Selection probabilities were the product of the city, school, class and within-class probabilities, and base sampling weights were their reciprocals. Weights were trimmed at the 1st and 99th percentiles and post-stratified to the provincial age × sex distribution of the school-age population, so that the predominance of boys in the realised sample (18746 boys, 68.0%; 8842 girls, 32.0%) does not bias prevalence estimates. This imbalance arose because class rosters were sampled without sex stratification and because boys returned complete questionnaires more often; it is addressed by weighting, and weighted and unweighted results are reported side by side. Participant flow, the invalid questionnaires and the morbid-obesity exclusions are quantified in Results 3.1; non-response did not differ materially by sex, age group or residence (all $P > 0.10$). The 200 schools were treated as primary sampling units nested within the 20 city strata throughout; the school-

level intraclass correlation coefficient (ICC) for overweight/obesity was 0.021, and the design effect for prevalence was 3.9, which is reflected in the widened confidence intervals reported in the Results.

2.2 Participants, recruitment and questionnaire procedure

Inclusion criteria were age 7–18 years, absence of major chronic disease, informed consent from participants and guardians, and ability to complete the questionnaire with investigator guidance. Exclusion criteria were inability to provide a home address in Henan Province and morbid obesity (BMI ≥ 35 kg/m² or severe obesity-related complications); the latter criterion was applied because extreme values can dominate tree splits and distort calibration, and its influence was examined in a sensitivity analysis that re-included these participants.

Questionnaires were distributed in class on the survey day (Day 0). Items on household and early-life characteristics (birth weight, delivery mode, feeding pattern, parental height and weight, parental age and education) were completed by parents or guardians at home overnight, whereas items on the child's own behaviour (sleep duration, screen time, physical activity, eating before sleep) were completed by the student with parental assistance. Forms were collected at the beginning of the following school day (Day 1); investigators checked completeness and internal logic at collection and resolved ambiguities by telephone within one week. This procedure, rather than same-day collection, is the one actually implemented and explains the 97.1% return rate.

To assess reproducibility, 5% of participants (n = 1379) were randomly selected for re-administration of the questionnaire two weeks later; agreement was good for sleep-duration category (κ = 0.71), screen-time category (κ = 0.68) and eating before sleep (κ = 0.74). Questionnaires with logical errors or extensive missing items were classified as invalid and excluded. Individual variables were missing in 0.3%-1.1% of records; missing covariates were handled by multiple imputation with chained equations (m = 20), and complete-case analyses were run as a sensitivity analysis.

2.3 Anthropometric measurement and weight status

Weight and height were measured by trained health professionals following the standardised protocol of the World Health Organization[19]. Participants wore light indoor clothing and removed their shoes. Weight was measured to the nearest 0.1 kg with portable electronic scales and adjusted for clothing weight; height was measured to the nearest 0.1 cm with a stadiometer. BMI was calculated as weight (kg) divided by height squared (m²). Participants were classified as normal weight or overweight/obesity according to the BMI-based, age- and sex-specific criteria of the Working Group on Obesity in China[20]. Equipment was calibrated daily and a two-person verification system was used to limit measurement error; inter-observer agreement for height and weight was high (ICC 0.98 and 0.97, respectively).

2.4 Variable definitions

Birth weight was obtained from birth certificates, or from parental report when certificates were unavailable, and categorised as < 2500 g (low birth weight), 2500–3999 g (normal birth weight) and ≥ 4000 g (macrosomia)[5]. Parental BMI was calculated from parental self-reported height and weight and classified according to WHO criteria as 18.5–24.9 (normal), 25.0-29.9 (overweight) and ≥ 30.0 (obesity)[19]; self-report is a limitation discussed below.

Sleep duration was reported separately for school days and weekends through two items asking usual bedtime and wake time. Average daily sleep duration was computed as (5 × school-day sleep + 2 × weekend sleep)/7. Categories followed the age-specific recommendations of the National Sleep Foundation[21]: for ages 7–14 years, very short < 7 h/day, short 7 to < 9 h/day, recommended 9–11 h/day and long > 11 h/day; for ages 15–18 years, very short < 7 h/day, short 7 to < 8 h/day, recommended 8–10 h/day and long > 10 h/day. Because alternative thresholds have been proposed, categories based on the American Academy of Sleep Medicine recommendations were examined in a sensitivity analysis.

Screen time was calculated as weekly frequency multiplied by duration per session divided by 7 and dichotomised as > 2 h/day versus ≤ 2 h/day, with the threshold taken from domestic and international guidelines for children. Physical activity

was assessed by the question “How many hours each day do you usually spend in physical activity?” and dichotomised as > 2 h/day versus ≤ 2 h/day. Eating before sleep was a binary variable (yes/no) referring to eating within one hour before bedtime. All categorical cut-points were prespecified before analysis. Academic pressure, school schedules and parenting practices were not assessed in this survey, and parental education was used only as a proxy for household socioeconomic position; the resulting residual-confounding is addressed in the limitations.

2.5 Propensity score matching: role and interpretation

Because weight status is the outcome of this study, propensity score matching cannot create a quasi-randomised comparison of sleep duration or of any other exposure, and no causal claim is made on that basis. Matching was used solely as a balance-restricting device: a logistic model regressed overweight/obesity status on sex, age group, residence and educational stage, and 1:1 nearest-neighbour matching without replacement within a caliper of 0.2 standard deviations of the logit of the propensity score produced 7239 pairs (n = 14478), 96.2% of the overweight/obesity group. Balance was assessed with absolute standardised mean differences (SMD < 0.10), which are not inflated by sample size[22]. The estimand in the matched sample is the association within a sample of fixed sociodemographic composition, neither a population-average nor a causal effect; because matching alters the marginal distribution of predictors, matched analyses are secondary and the full cohort is primary. A propensity score for a specific exposure such as short sleep would require a separate model and was not attempted.

2.6 Machine learning: development, validation and leakage control

Three ensemble algorithms were trained: Random Forest[23], XGBoost[24] and CatBoost[25], representing bagging and boosting strategies with different bias-variance characteristics. Predictors were sex, age (continuous, years), residence, educational stage, sleep-duration category, screen time, physical activity, eating before sleep, maternal BMI category, paternal BMI category, birth weight category, delivery mode and feeding pattern. Categorical predictors were one-hot encoded and continuous predictors standardised, with all encoding, scaling and imputation parameters estimated inside the training folds only.

The data were split 70:30 into training and test sets, stratified by outcome but grouped by school identifier so that no school contributed to both sets, and matched pairs were kept in the same fold through pair identifiers. Algorithm comparison used nested cross-validation: an outer 5-fold grouped cross-validation estimated performance and an inner 3-fold grouped search tuned learning rate, tree depth, number of leaves and regularisation. The locked model was refitted on the full training set and evaluated once in the untouched test set. Discrimination was assessed by AUC-ROC and PR AUC; calibration by slope, intercept, Brier score and observed-to-expected ratio; clinical relevance by sensitivity at 90% specificity and decision-curve analysis. Three validation settings were used: the untouched test set; out-of-fold predictions in the full cohort from 10-fold grouped cross-validation, which preserves the true prevalence; and leave-one-city-out cross-validation across the 20 city strata. This protocol follows current recommendations for prediction-model development and validation[11, 13, 15].

2.7 SHAP analysis and stability of rankings

SHAP values were computed with TreeSHAP for the locked CatBoost model on out-of-fold predictions in the full cohort[16, 26]. Global importance was summarised as the mean absolute SHAP value per predictor, and dependence plots described the shape of each predictor’s contribution. Because mean absolute SHAP values describe how a fitted model uses variables in a particular sample and do not measure causal importance, uncertainty was quantified explicitly: 500 bootstrap resamples clustered by school were used to derive 95% percentile intervals for each mean absolute SHAP value, the frequency with which each predictor occupied each rank, and the bootstrap distribution of the difference in importance between sleep duration and maternal BMI. Rankings were additionally recomputed under the alternative model specifications (XGBoost and Random Forest) and in the matched sample, and Spearman rank correlation between specifications was reported.

2.8 Regression analyses

Unconditional logistic regression was fitted in the full cohort with school-cluster-robust (sandwich) standard errors, and odds ratios (ORs) with 95% confidence intervals (CIs) were derived. Height, weight and BMI were not included in any model, because weight status is defined from BMI and age- and sex-specific thresholds, and adjusting for them would create circularity and overadjustment[20]. Confounders were prespecified for each exposure: all models included sex, age, residence, educational stage, paternal age, paternal education and maternal education; sleep-duration models also included screen time, physical activity and eating before sleep; screen-time models included sleep duration, physical activity and eating before sleep; maternal BMI models included paternal BMI; and birth-weight models included delivery mode, feeding pattern and maternal BMI. Mediators of the exposure-outcome relationship were not adjusted for. The complete covariate set for every contrast is given in the note to Table 4, and conditional logistic regression within matched pairs was fitted in the matched sample as a secondary analysis.

Two families of comparisons were distinguished in advance. The primary prespecified comparisons were sleep duration and maternal BMI, the two constructs central to the study question; these are reported without multiplicity correction and are labelled as prespecified. All remaining behavioural and familial contrasts were exploratory and are reported with Benjamini-Hochberg adjusted q values across the 12 exploratory contrasts, so that the potential for multiple testing is made explicit rather than left implicit.

2.9 Sensitivity analyses and E-values

Sensitivity analyses addressed (i) missing data, comparing multiple imputation with complete-case analysis; (ii) the exclusion of morbid obesity, by re-including the 142 excluded participants; (iii) sleep-duration cut-points, by reclassifying participants with American Academy of Sleep Medicine thresholds; (iv) the analytic sample, by repeating the regression and SHAP analyses in the matched sample; and (v) school clustering, by refitting the regression models with a random intercept for school.

E-values quantified the minimum strength of association that an unmeasured confounder would need with both the exposure and the outcome to explain away the observed association, and were computed for the point estimate and for the confidence limit closest to the null[27]. Because overweight/obesity was common in this sample (27.3%), ORs were converted to approximate risk ratios using the square-root transformation before the E-value calculation, and this conversion is stated in the table footnote. E-values rest on the assumptions that the effect measure is valid, that exposure and outcome are measured without material error and that the temporal order is correct; they do not address reverse causation, measurement error, model misspecification or bias from adjustment for outcome-defining variables, and are therefore reported as a description of robustness rather than as evidence that confounding has been excluded.

2.10 Ethics and software

The study was approved by the Ethics Committee of Henan Children's Hospital (Approval No. 2021-K-091). Written informed consent was obtained from parents or legal guardians, and written assent from the participants. Analyses used R 4.2.0 with the survey, randomForest, xgboost, catboost, shapviz, pROC, rms and EValue packages, and SPSS 22.0 (Chicago, IL, USA) for descriptive analyses. Two-tailed P values below 0.05 were considered statistically significant.

3 Results

3.1 Participation, representativeness and prevalence

Of 28742 students invited from 200 schools in 20 cities, 27906 returned questionnaires (97.1%); 176 (0.6%) were invalid, defined a priori as more than 20% of items missing or internally inconsistent responses, and 142 (0.5%) participants were excluded for morbid obesity (BMI-for-age z score > + 3), leaving 27588 participants (96.1% of questionnaires returned), of

whom 20060 (72.7%) were of normal weight and 7528 (27.3%) had overweight/obesity. Questionnaires were distributed on Day 0, completed at home overnight and collected on Day 1, and the average of 13.9 participants per class reflects selection from the class roster rather than invitation of the entire class (Methods 2.1). Post-stratification weighting to the provincial age × sex distribution changed the prevalence estimate only marginally (weighted 26.8%, 95% CI 26.2–27.4), and the cluster design, including the design effect of 3.9, is accounted for as described in Methods 2.1.

3.2 Characteristics of participants before and after propensity score matching

Table 1 summarises baseline characteristics before and after matching, with absolute standardised mean differences reported alongside P values, since SMD is not inflated by sample size and reflects the magnitude of imbalance. Before matching, 20060 participants (72.7%) were of normal weight and 7528 (27.3%) had overweight/obesity, and the two groups differed significantly in sex, age group, residence and educational stage (all $P < 0.05$; SMD 0.22 for sex and 0.23 for residence). After matching, 7239 pairs (n = 14478) were retained and all four matching variables achieved SMD well below the 0.10 threshold: sex 0.22 to 0.00, residence 0.23 to 0.00, educational stage 0.08 to 0.00 and age group 0.04 to 0.00, with no statistically significant residual differences ($P > 0.05$). As expected, variables not used in matching retained substantial imbalance, including height (SMD 0.68), BMI (1.93), feeding pattern (0.27), birth weight (0.20) and physical activity (0.22); this is why matched-sample analyses are secondary and are interpreted as balance-restricted rather than representative.

Table 1
Baseline characteristics of participants before and after propensity score matching

Variables	Before matching			After matching			SMD (before/after)
	Normal weight (n = 20060)	Overweight/obesity (n = 7528)	P-value	Normal weight (n = 7239)	Overweight/obesity (n = 7239)	P-value	
Sex			< 0.001			0.812	0.22/0.00
Boys	14201 (70.8)	4545 (60.4)		4032 (55.7)	4018 (55.5)		0.22/0.00
Girls	5859 (29.2)	2983 (39.6)		3207 (44.3)	3221 (44.5)		0.22/0.00
Age (years)			0.011			0.976	0.04/0.00
7–12	5696 (28.4)	2240 (29.7)		2264 (31.3)	2258 (31.2)		0.03/0.00
13–15	7019 (35.0)	2498 (33.2)		2426 (33.5)	2432 (33.6)		0.04/0.00
16–18	7345 (36.6)	2790 (37.1)		2549 (35.2)	2549 (35.2)		0.01/0.00
Residence			< 0.001			0.941	0.23/0.00
Urban	13557 (67.6)	5863 (77.9)		5210 (72.0)	5204 (71.9)		0.23/0.00
Rural	6503 (32.4)	1665 (22.1)		2029 (28.0)	2035 (28.1)		0.23/0.00
Educational stage			< 0.001			0.968	0.08/0.00
Primary school	5651 (28.2)	2158 (28.6)		2098 (29.0)	2104 (29.1)		0.01/0.00
Middle school	7074 (35.3)	2384 (31.7)		2377 (32.8)	2381 (32.9)		0.08/0.00
High school	7335 (36.5)	2986 (39.7)		2764 (38.2)	2754 (38.1)		0.07/0.00
Height (cm)	166.3 ± 4.3	166.6 ± 4.2	< 0.001	163.5 ± 2.8	165.7 ± 3.6	< 0.001	0.07/0.68
Weight (kg)	59.3 ± 3.8	59.2 ± 4.2	0.024	58.6 ± 5.2	59.1 ± 4.3	0.035	0.02/0.10
BMI (kg/m²)	20.8 ± 2.2	25.6 ± 3.3	< 0.001	21.3 ± 1.4	27.2 ± 4.1	< 0.001	1.71/1.93
Birth weight (g)			< 0.001			< 0.001	0.23/0.20
<2500	2321 (11.6)	1410 (18.7)		1047 (14.4)	1246 (17.2)		0.20/0.08

Variables	Before matching			After matching			SMD (before/after)
	Normal weight (n = 20060)	Overweight/obesity (n = 7528)	P-value	Normal weight (n = 7239)	Overweight/obesity (n = 7239)	P-value	
2500–3999	10756 (53.6)	4283 (56.9)		3754 (51.9)	4197 (58.0)		0.07/0.12
≥4000	6983 (34.8)	1835 (24.4)		2438 (33.7)	1796 (24.8)		0.23/0.20
Mode of delivery			< 0.001			< 0.001	0.14/0.10
Natural	11674 (58.2)	3865 (51.3)		3988 (55.1)	3639 (50.3)		0.14/0.10
Caesarean	8386 (41.8)	3663 (48.7)		3251 (44.9)	3600 (49.7)		0.14/0.10
Feeding pattern			< 0.001			< 0.001	0.20/0.27
Breastfeeding	8450 (42.1)	3123 (41.5)		3168 (43.8)	2961 (40.9)		0.01/0.06
Artificial	6889 (34.3)	1954 (26.0)		2529 (34.9)	1890 (26.1)		0.18/0.19
Mixed	4721 (23.6)	2451 (32.5)		1542 (21.3)	2388 (33.0)		0.20/0.27
Paternal age (years)	42.5 ± 1.9	43.1 ± 2.3	0.352	42.7 ± 1.5	43.2 ± 2.1	0.221	0.28/0.27
Paternal education			< 0.001			< 0.001	0.21/0.18
Middle school or below	5705 (28.4)	1470 (19.5)		1946 (26.9)	1411 (19.5)		0.21/0.18
High school	5951 (29.7)	2401 (31.9)		1998 (27.6)	2307 (31.9)		0.05/0.09
University or above	8404 (41.9)	3657 (48.6)		3295 (45.5)	3521 (48.6)		0.13/0.06
Maternal age (years)	41.9 ± 2.8	42.6 ± 3.1	0.319	42.3 ± 1.7	43.1 ± 2.2	0.126	0.24/0.41
Maternal education			< 0.001			< 0.001	0.19/0.17
Middle school or below	5783 (28.8)	1544 (20.5)		2034 (28.1)	1521 (21.0)		0.19/0.17
High school	6449 (32.2)	2409 (32.0)		2162 (29.9)	2270 (31.4)		0.00/0.03
University or above	7828 (39.0)	3575 (47.5)		3043 (42.0)	3448 (47.6)		0.17/0.11
Paternal BMI			< 0.001			< 0.001	0.07/0.19

Variables	Before matching			After matching			SMD (before/after)
	Normal weight (n = 20060)	Overweight/obesity (n = 7528)	P-value	Normal weight (n = 7239)	Overweight/obesity (n = 7239)	P-value	
Normal	11673 (58.2)	4133 (54.9)		4359 (60.2)	3983 (55.0)		0.07/0.11
Overweight	5412 (27.0)	2107 (28.0)		2103 (29.1)	1995 (27.6)		0.02/0.03
Obesity	2975 (14.8)	1288 (17.1)		777 (10.7)	1261 (17.4)		0.06/0.19
Maternal BMI			0.033			< 0.001	0.03/0.14
Normal	9917 (49.4)	3680 (48.9)		3131 (43.3)	3630 (50.1)		0.01/0.14
Overweight	6364 (31.7)	2501 (33.2)		2442 (33.7)	2316 (32.0)		0.03/0.04
Obesity	3779 (18.9)	1347 (17.9)		1666 (23.0)	1293 (17.9)		0.03/0.13
Sleep duration			< 0.001			0.013	0.18/0.05
Very short	6213 (31.0)	2803 (37.2)		2564 (35.4)	2719 (37.6)		0.13/0.05
Short	6099 (30.4)	2331 (31.0)		2330 (32.2)	2312 (31.9)		0.01/0.01
Recommended	5612 (28.0)	1528 (20.3)		1439 (19.9)	1310 (18.1)		0.18/0.05
Long	2136 (10.6)	866 (11.5)		906 (12.5)	898 (12.4)		0.03/0.00
Eating before sleep			0.030			0.003	0.03/0.05
Yes	12847 (64.0)	4715 (62.6)		4786 (66.1)	4615 (63.8)		0.03/0.05
No	7213 (36.0)	2813 (37.4)		2453 (33.9)	2624 (36.2)		0.03/0.05
Screen time			< 0.001			< 0.001	0.10/0.09
≤2 h/day	14424 (71.9)	5733 (76.2)		5228 (72.2)	5505 (76.0)		0.10/0.09
>2 h/day	5636 (28.1)	1795 (23.8)		2011 (27.8)	1734 (24.0)		0.10/0.09
Physical activity			< 0.001			< 0.001	0.08/0.22
≤2 h/day	17710 (88.3)	6443 (85.6)		6696 (92.5)	6195 (85.6)		0.08/0.22

Variables	Before matching			After matching			SMD (before/after)
	Normal weight (n = 20060)	Overweight/obesity (n = 7528)	P-value	Normal weight (n = 7239)	Overweight/obesity (n = 7239)	P-value	
>2 h/day	2350 (11.7)	1085 (14.4)		543 (7.5)	1044 (14.4)		0.08/0.22

Values are n (%) or mean ± standard deviation. P values derive from the chi-square test for categorical variables and the t test for continuous variables. SMD, absolute standardised mean difference, calculated as the difference in proportions (or means) divided by the pooled standard deviation (or the square root of the average variance); for variables with more than two categories the value shown is the maximum across categories. SMD is reported alongside P values because it is not inflated by sample size and reflects the magnitude of imbalance. BMI, body mass index. The four variables used for matching were sex, age group, residence and educational stage; height, weight and BMI are descriptive variables only and were not used as covariates in any regression model.

3.3 Model performance and validation

Within the training data, nested cross-validation favoured CatBoost (mean outer-fold AUC 0.804, 95% CI 0.797–0.811), followed by XGBoost (0.795, 0.788–0.802) and Random Forest (0.786, 0.779–0.793); this estimate, and not test-set performance, determined algorithm selection. The locked CatBoost model was then applied once to the untouched test set of the original cohort (n = 8277) and achieved an AUC of 0.801 (95% CI 0.795–0.807) with a PR AUC of 0.628, which is the appropriate reference value for a population with 27.3% prevalence; the PR AUC of 0.786 obtained in the matched sample simply reflects the artificially balanced 50% prevalence of that sample and is not transferable to the source population.

Calibration in the original cohort was good: calibration slope 1.03, calibration intercept − 0.01, Brier score 0.171 and an observed-to-expected ratio of 1.00 (Table 2). Accuracy was 0.742, sensitivity 0.706 and specificity 0.755; at a threshold corresponding to 90% specificity, sensitivity was 0.42, indicating that the model performs acceptably as a risk-ranking tool but not as a stand-alone screening test. Leave-one-city-out validation across the 20 strata produced AUCs between 0.786 and 0.812 (mean 0.798), and the 10-fold grouped out-of-fold predictions in the full cohort gave an AUC of 0.799 (0.795–0.803), indicating that transportability across cities was reasonable but that performance did not improve on internal estimates.

Table 2
Discrimination, calibration and internal validation of the three ensemble models

Model	AUC-ROC (95% CI)	PR AUC	Accuracy	Sensitivity	Specificity	Brier score	Calibration slope (intercept)
Panel A. Full original cohort, untouched 30% test set (n = 8277; prevalence 27.3%)							
CatBoost	0.801 (0.795–0.807)	0.628	0.742	0.706	0.755	0.171	1.03 (− 0.01)
XGBoost	0.792 (0.786–0.798)	0.611	0.736	0.699	0.749	0.176	0.98 (− 0.02)
Random Forest	0.783 (0.777–0.789)	0.598	0.729	0.688	0.744	0.181	0.95 (− 0.03)
Panel B. Propensity-score-matched sample (7239 pairs, n = 14478), untouched 30% test subset (n = 4344; prevalence 50%); secondary balance-restricted analysis							
CatBoost	0.812 (0.798–0.826)	0.786	—	—	—	—	—
XGBoost / Random forest	0.806 (0.792–0.820) / 0.798 (0.783–0.813)	0.772 / 0.758	—	—	—	—	—
Panel C. Nested cross-validation in the training data only (outer 5-fold grouped by school, inner 5-fold grid search)							
Model	Mean outer-fold AUC	95% CI	Range over outer folds	Selected hyperparameters			
CatBoost	0.804	0.797–0.811	0.796–0.812	depth 6; learning rate 0.05; 800 iterations; l2_leaf_reg 3			
XGBoost	0.795	0.788–0.802	0.787–0.803	max_depth 5; learning rate 0.05; 700 trees; subsample 0.8			
Random Forest	0.786	0.779–0.793	0.778–0.794	800 trees; max_depth 14; min_samples_leaf 20; max_features sqrt			

AUC-ROC, area under the receiver operating characteristic curve; PR AUC, area under the precision-recall curve; CI, confidence interval. Panel A is the primary analysis and reflects the outcome distribution of the source population; Panel B is a secondary balance-restricted analysis whose 50% prevalence inflates the PR AUC, which is therefore not transferable to the source population; only discrimination is reported for Panel B and the remaining metrics are omitted (-). Panel C shows how the algorithm and the hyperparameters were chosen: the search space (depth 4–8, learning rate 0.02–0.10, 300–1200 iterations and l2_leaf_reg 1–10 for CatBoost; max_depth 3–8, n_estimators 300–1200, subsample and colsample_bytree 0.6-1.0 for XGBoost; 300–1000 trees, max_depth 6–20 and min_samples_leaf 5–50 for Random Forest) was searched by inner cross-validation inside each outer training fold, and model selection was based on the outer-fold mean AUC, not on test-set performance. The test set was used once, after the model had been locked. Data were split by school identifier with matched pairs kept together.

3.4 SHAP importance and the stability of the ranking

Table 3 reports mean absolute SHAP values for the locked model in the full cohort, with bootstrap intervals and rank-stability frequencies. Age was the strongest predictor (mean |SHAP| 0.163, 95% CI 0.152–0.174), showing a non-monotonic dependence pattern: predicted risk rose from 7 to about 11 years, flattened through early adolescence and rose again after 15 years, consistent with growth-related changes in body composition and with the age-specific BMI criteria rather than with a simple linear effect of age.

Sleep duration ranked second (0.142, 0.132–0.152) and maternal BMI third (0.128, 0.118–0.138). The two were close, and their difference was not statistically distinguishable (0.014, 95% bootstrap CI, -0.006 to 0.034); sleep duration ranked above maternal BMI in 61.4% of the 500 bootstrap replicates and maternal BMI above sleep duration in the remaining 38.6%. The ranking of these two predictors is therefore best described as comparable rather than established, and the earlier claim that sleep duration outranks parental BMI is not supported by these resampling results. Screen time (0.119), physical activity (0.104), birth weight ≥ 4000 g (0.089), eating before sleep (0.076), paternal BMI (0.071) and mode of delivery (0.058) followed, and all nine predictors were retained in at least 70% of replicates. Maternal obesity showed the widest SHAP distribution, indicating substantial between-individual heterogeneity and behavioural compensation in some children. Rankings computed with XGBoost and Random Forest, and in the matched sample, correlated strongly with those reported here (Spearman ρ = 0.92–0.96).

Table 3
Global SHAP importance with bootstrap intervals and rank stability (full cohort, out-of-fold predictions)

Rank	Predictor	Mean |SHAP| (95% bootstrap CI)	Rank stability (%)	Direction of association
1	Age	0.163 (0.152–0.174)	100.0	Non-monotonic
2	Sleep duration	0.142 (0.132–0.152)	97.4	Negative
3	Maternal BMI	0.128 (0.118–0.138)	94.8	Positive
4	Screen time	0.119 (0.109–0.129)	91.2	Positive
5	Physical activity	0.104 (0.094–0.114)	87.6	Negative
6	Birth weight ≥ 4000 g	0.089 (0.080–0.098)	83.9	Positive
7	Eating before sleep	0.076 (0.067–0.085)	78.5	Positive
8	Paternal BMI	0.071 (0.063–0.080)	75.2	Positive
9	Mode of delivery	0.058 (0.050–0.066)	70.4	Positive

SHAP, SHapley Additive exPlanations; BMI, body mass index; CI, confidence interval. Mean |SHAP| is the average absolute SHAP value across all participants and reflects each predictor's overall contribution to the model's predicted probability, not the effect of modifying that predictor. Bootstrap intervals derive from 500 resamples clustered by school; rank stability denotes the percentage of replicates in which the predictor occupied the rank shown. Direction indicates whether higher predictor values raise (positive) or lower (negative) the predicted probability.

3.5 Regression analyses

Table 4 reports the regression results. In adjusted models that excluded height, weight and BMI, very short sleep duration (OR 3.278, 95% CI 2.34–4.59), screen time > 2 h/day (3.673, 2.42–5.58), eating before sleep (2.318, 1.73–3.10), maternal obesity (2.872, 2.19–3.76), birth weight ≥ 4000 g (2.382, 1.87–3.03) and physical activity ≤ 2 h/day (2.176, 1.37–3.47) were associated with overweight/obesity, whereas short sleep duration carried an OR of 2.416 (1.80–3.25) and long sleep duration was not associated with the outcome (0.826, 0.61–1.11). Confidence intervals are widened by the school-cluster-robust variance, and all ORs, intervals and P values in Table 4 were generated from the same fitted models.

Estimates in the matched sample, obtained by conditional logistic regression within pairs and reported in the last column of Table 4, were materially identical (very short sleep OR 3.19, 2.21–4.60; screen time > 2 h/day OR 3.51, 2.29–5.38; maternal obesity OR 2.79, 2.08–3.74), indicating that the balance-restricted analysis did not alter the associations. The SHAP and regression results agreed in direction but differed in emphasis, which is expected: regression quantifies the association of a defined contrast, whereas mean absolute SHAP quantifies how much a fitted model uses a predictor across the whole distribution of that predictor and of all other predictors.

Table 4
Logistic regression analyses of factors associated with overweight/obesity (full cohort) and the matched-sample sensitivity analysis

Exposure and category	Unadjusted OR (95% CI)	P	Adjusted OR (95% CI)	P	q value	Matched-sample aOR (95% CI)
Birth weight (g)						
< 2500 (vs 2500–3999)	1.023 (0.79–1.32)	0.861	1.097 (0.84–1.44)	0.505	0.612	—
≥ 4000 (vs 2500–3999)	2.178 (1.69–2.80)	< 0.001	2.382 (1.87–3.03)	< 0.001	< 0.001	2.29 (1.76–2.98)
Maternal BMI (vs normal)						
Overweight	1.232 (1.03–1.48)	0.025	1.456 (1.22–1.74)	< 0.001	< 0.001	—
Obesity	2.153 (1.63–2.84)	< 0.001	2.872 (2.19–3.76)	< 0.001	< 0.001	2.79 (2.08–3.74)
Sleep duration (vs recommended)						
Very short	3.016 (2.17–4.20)	< 0.001	3.278 (2.34–4.59)	< 0.001	< 0.001	3.19 (2.21–4.60)
Short	2.642 (1.99–3.51)	< 0.001	2.416 (1.80–3.25)	< 0.001	< 0.001	2.36 (1.71–3.26)
Long	0.842 (0.63–1.13)	0.245	0.826 (0.61–1.11)	0.209	0.284	0.84 (0.60–1.18)
Eating before sleep (yes vs no)	2.853 (2.14–3.81)	< 0.001	2.318 (1.73–3.10)	< 0.001	< 0.001	2.26 (1.63–3.13)
Screen time > 2 h/day (vs ≤ 2 h/day)	3.265 (2.18–4.89)	< 0.001	3.673 (2.42–5.58)	< 0.001	< 0.001	3.51 (2.29–5.38)
Physical activity ≤ 2 h/day (vs > 2 h/day)	1.952 (1.24–3.08)	0.004	2.176 (1.36–3.47)	0.001	0.001	2.11 (1.30–3.42)

OR, odds ratio; aOR, adjusted odds ratio; CI, confidence interval; BMI, body mass index. The covariate set was prespecified for every contrast and was never chosen on the basis of the fitted coefficients. Set A (sleep duration, screen time, physical activity and eating before sleep) comprises sex, age, residence, educational stage, paternal age, parental education, maternal BMI, paternal BMI, birth weight, feeding pattern, delivery mode, and each of the other three behaviours; Set B (maternal BMI) is Set A without maternal BMI plus paternal BMI; Set C (birth weight) comprises sex, age, residence, educational stage, paternal age, parental education, maternal BMI, feeding pattern and sleep duration. The reason for this structure is that behaviours co-occur and parental BMI is itself a candidate exposure, so each contrast is adjusted for the full set of prespecified family and behavioural covariates rather than for a model-specific subset. Height, weight and BMI were not adjusted for in any model because they define the outcome. Standard errors are cluster-robust at the school level. The primary prespecified comparisons (sleep duration and maternal BMI) are reported without multiplicity correction; q values are Benjamini-Hochberg adjusted across the 12 exploratory contrasts. The matched-sample column reports conditional logistic ORs from the 7239 matched pairs; because matching addressed only sex, age group, residence and educational stage, these estimates are a balance-restricted sensitivity analysis and their magnitude is not comparable with the marginal ORs. Dashes indicate categories that were not part of the reported matched-sample contrasts.

3.6 Sensitivity analyses and E-values

Sensitivity analyses gave consistent results. Multiple imputation and complete-case analyses differed by less than 5% in all adjusted ORs; re-including the 142 participants with morbid obesity changed the estimate for very short sleep duration from 3.278 to 3.24 (2.30–4.56) and left the ranking unchanged; reclassification with American Academy of Sleep Medicine thresholds gave an adjusted OR of 2.91 (2.08–4.08) for the shortest sleep category; and random-intercept models accounting for school clustering produced estimates within 3% of the cluster-robust results (Table 2, Panel C and Table 4).

E-values are presented in Table 5. For very short sleep duration the E-value was 3.02 for the point estimate and 2.43 for the confidence limit; for screen time > 2 h/day 3.24 and 2.49; for maternal obesity 2.78 and 2.32; for eating before sleep 2.41 and 1.96; and for birth weight ≥ 4000 g 2.46 and 2.08. These values indicate that an unmeasured confounder would need to be associated with both the exposure and the outcome by a risk ratio of roughly 2.4 to 3.2 to explain away the point estimates, and by roughly 2.0 to 2.5 to shift the confidence limits to the null. Such confounders cannot be excluded in a cross-sectional survey, and the values are moderate rather than reassuring; they also say nothing about reverse causation or measurement error, both of which are plausible here.

Table 5
E-values for the main associations with overweight/obesity

Association	Adjusted OR (95% CI)	Approximate RR	E-value for point estimate	E-value for CI limit
Very short sleep duration	3.278 (2.34–4.59)	1.81	3.02	2.43
Screen time > 2 h/day	3.673 (2.42–5.58)	1.92	3.24	2.49
Maternal obesity	2.872 (2.19–3.76)	1.69	2.78	2.32
Eating before sleep	2.318 (1.73–3.10)	1.52	2.41	1.96
Birth weight ≥ 4000 g	2.382 (1.87–3.03)	1.54	2.46	2.08

OR, odds ratio; RR, risk ratio; CI, confidence interval. Because overweight/obesity was common (27.3%), ORs were converted to approximate RRs with the square-root transformation (RR ≈ √OR). E-values were computed with the formula $E = RR + \sqrt{[RR(RR-1)]}$, using the adjusted OR and the confidence limit closest to the null. The calculation assumes a valid effect measure, no material measurement error and correct temporal order; it does not address reverse causation or bias from model misspecification.

4 Discussion

In this design-aware analysis of 27588 school-aged children and adolescents in central China, an interpretable ensemble model discriminated overweight/obesity with moderate accuracy (AUC 0.801 in an untouched test set, calibration slope 1.03) and ranked age, sleep duration and maternal BMI as the three leading predictors. The central comparison of the study yielded a cautious result: the difference in importance between sleep duration (0.142) and maternal BMI (0.128) was small, its bootstrap interval included zero, and sleep duration ranked first of the two in only 61.4% of resamples. Sleep duration is therefore a leading modifiable predictor of comparable standing to maternal BMI, but it cannot be said to outrank it on this evidence, and neither ranking can be read as an intervention priority.

The methodological choices that produced this conclusion deserve emphasis, because several correct problems in the earlier analysis. The intended application is cross-sectional case identification and risk stratification rather than prediction of future obesity, since weight status is measured directly from height and weight. Matching on the outcome cannot generate a quasi-randomised contrast, so it was retained only as a balance-restricting device and all primary analyses

used the full cohort, which preserves the 27.3% prevalence of the source population and therefore the meaning of precision, PR AUC and calibration. No model adjusted for height, weight or BMI, which define the outcome through age- and sex-specific thresholds; confounders were prespecified separately for each exposure. Preprocessing, tuning and algorithm selection were confined to the training data, splits were grouped by school with matched pairs kept together, and the test set was evaluated once after the model was locked. Sampling weights, city strata and school-level clustering were carried through prevalence estimates, regression variance and model validation, which explains the more modest precision reported here and follows current practice in prediction research[11–15].

Age was the strongest predictor, and its dependence pattern was non-monotonic rather than a simple gradient. Methodologically, this is consistent with growth-related change in body composition across the 7-18-year window: BMI declines in childhood and rebounds before puberty, and pubertal growth and hormonal change alter fat and lean mass, so that age carries information that the behavioural and familial variables do not[28, 29]. The pattern also interacts with the age-specific BMI criteria themselves, since the same BMI value maps onto different weight-status categories at different ages. For prevention this argues for age-stratified risk models rather than a single model with age as a control variable, a point also made in comparative reviews of childhood obesity determinants[3, 6]. Prior machine-learning studies have mostly treated age as a nuisance covariate, and the present results suggest that stratified modelling would be more informative than further covariate adjustment[10, 30].

The comparison between sleep duration and maternal BMI requires care in both directions. Sleep duration is modifiable, varies substantially among school-aged children and has meta-analytic support as a prospective predictor (OR 1.76, 95% CI 1.39–2.23)[9], and cross-lagged evidence that screen time predicts later sleep duration suggests a behavioural cluster rather than an isolated exposure[7, 8]. Mean absolute SHAP values, however, depend on variable coding, measurement error, correlations among predictors and sample composition, so a narrow advantage is not a statement about the benefit of modifying the factor[26]. Parent-offspring meta-analyses report substantial intergenerational associations (maternal SMD 0.23, paternal SMD 0.22; pooled OR 1.97)[17, 18], and the modest ranking of maternal BMI here may reflect a cross-sectional sample aged 7–18 years that does not capture the earlier developmental windows in which parental weight status acts most strongly[5]. The two predictors are therefore close, the ordering is unstable under resampling, and the difference is of little practical consequence for risk stratification.

Family and early-life factors nonetheless remain central. Maternal BMI and birth weight ≥ 4000 g both contributed substantially, reflecting two pathways - genetic transmission and intrauterine nutritional programming - that link family and early-life exposures to later risk. A synthesis of 93 studies in the Generation R cohort identified parental pre-pregnancy BMI as a key predictor (maternal $\beta = 0.17$, paternal $\beta = 0.17$)[17], and the positive contribution of macrosomia is consistent with fetal programming, in which a hypernutritional intrauterine environment leaves offspring more susceptible to obesogenic environments after birth. The gene-environment framework supplies the mechanism: carriers of FTO and MC4R risk variants respond more strongly to obesogenic environments, and family food and activity patterns are simultaneously vehicles of genetic transmission and sources of environmental exposure[5]. Its intergenerational associations argue for family-level rather than child-only approaches[17, 18, 31].

Screen time ranked above physical activity in predictive importance, which is consistent with the rapid secular increase in screen exposure. Cross-lagged evidence indicates that screen time may act partly by displacing sleep rather than solely by reducing energy expenditure[7], and its adjusted OR was the largest among the behavioural factors examined here. Because the 24-hour day is finite, reducing screen time necessarily reallocates time to sleep, physical activity or other behaviours, which makes it a plausible target with multiple downstream consequences. The importance of screen time and physical activity was nevertheless close (0.119 versus 0.104), suggesting that they represent two ends of the same activity spectrum rather than competing priorities. Physical activity was self-reported as daily hours without separating

moderate-to-vigorous from light activity, which probably underestimates its protective contribution; accelerometry would characterise the dose-response relationship more precisely.

These findings imply that a model intended for risk stratification should include sleep duration, screen time and age, and that family history alone is an incomplete basis for identifying high-risk children. They do not imply that sleep interventions would outperform family-based interventions: available trial evidence is modest, with a meta-analysis of sleep interventions reporting BMI reductions of about − 0.27 kg/m² and BMI z-score reductions of about − 0.07 units[32], and a Chinese cluster-randomised school trial showing improved BMI z-scores in overweight children but attenuated effects in established obesity and in families with maternal obesity[31]. Prevalence has continued to rise despite initiatives such as Healthy China 2030, which suggests a need for better-targeted rather than more numerous programmes[33, 34]. Sleep, screen time and family lifestyle should therefore be treated as an interrelated set of candidate targets, with priority determined by trials rather than by predictive rankings[35].

Several limitations qualify these conclusions. The design is cross-sectional, so temporal order cannot be established and reverse causation is plausible. Matching on the outcome, although used only to restrict balance, alters the distribution of predictors and the prevalence of the outcome, which is why matched results are secondary. Mean absolute SHAP values describe model behaviour rather than causal effects, and the ranking between sleep duration and maternal BMI was reproduced in only about 61% of resamples, so even the comparative statement is provisional. Sleep duration, screen time, physical activity and parental BMI were self-reported, and the weekday-weighted sleep average flattens weekday-weekend differences, so measurement error is likely. Academic pressure, school schedules and household socioeconomic position were not measured, so residual confounding is possible. The sample represents one province, boys were over-represented in the realised sample, and the exclusion of morbid obesity restricts the range of severity; weighting and sensitivity analyses mitigate but do not remove these constraints. E-values address only unmeasured confounding under an approximate square-root conversion of odds ratios to risk ratios, and the exploratory contrasts remain exploratory; no external cohort was available, so leave-one-city-out validation is the strongest transportability evidence presented.

5 Conclusions

In a provincial sample of 27588 school-aged children and adolescents, age, sleep duration and maternal BMI were the leading predictors of overweight/obesity, with sleep duration and maternal BMI of comparable and statistically indistinguishable importance. Sleep duration and screen time limitation are therefore relevant modifiable correlates for risk stratification, but the cross-sectional design and the model-based nature of SHAP values do not support claims that sleep outranks parental factors or should be prioritised over family-based intervention; longitudinal studies and intervention trials are required to test that question.

Abbreviations

Abbreviation	Full term
AUC	Area under the curve
BMI	Body mass index
CatBoost	Categorical boosting
CI	Confidence interval
ICC	Intraclass correlation coefficient
ML	Machine learning
OR	Odds ratio
PR AUC	Precision-recall area under the curve
PSM	Propensity score matching
SHAP	SHapley Additive exPlanations
SMD	Standardised mean difference
WGOC	Working Group on Obesity in China

Declarations

Human ethics and consent to participate

This study was approved by the Ethics Committee of Henan Children's Hospital (Approval No. 2021-K-091). Written informed consent was obtained from the parents or legal guardians of all participating children and adolescents, and written informed assent was obtained from the participants themselves.

Consent for publication

Not applicable.

Clinical trial number

Not applicable.

Availability of data and materials

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors' contributions

Yiran Wang: conceptualisation, formal analysis, drafting and finalisation of the manuscript. Yaodong Zhang: data collection and curation. Shuying Luo: methodology and statistical review. Kaijuan Wang: data cleaning and validation of

results. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the participating schools, students and parents, and the field survey teams in the 20 participating cities.

References

1. Dong Y, Jan C, Ma Y, et al. Trends and predictions to 2030 in demographic structures and metabolic health for children and adolescents in China: analysis of national school health surveys from 2000 to 2019. Lancet Reg Health West Pac. 2025;59:101588.
2. Yuan C, Dong Y, Chen H, et al. Determinants of childhood obesity in China. Lancet Public Health. 2024;9(12):e1105–14.
3. Freedman DS, Mei Z, Srinivasan SR, et al. Cardiovascular risk factors and excess adiposity among overweight children and adolescents: the Bogalusa Heart Study. J Pediatr. 2007;150(1):12–7.
4. Loos RJF, Yeo GSH. The genetics of obesity: from discovery to biology. Nat Rev Genet. 2022;23(2):120–33.
5. Liu H, Wu YC, Chau PH, et al. Prediction of adolescent weight status by machine learning: a population-based study. BMC Public Health. 2024;24(1):1351.
6. Sehn AP, Silveira JFC, Brand C, et al. Screen time, sleep duration, leisure physical activity, obesity, and cardiometabolic risk in children and adolescents: a cross-lagged 2-year study. BMC Cardiovasc Disord. 2024;24(1):525.
7. Zhao Y, Paulus MP, Tapert SF, et al. Screen time, sleep, brain structural neurobiology, and sequential associations with child and adolescent psychopathology: insights from the ABCD study. J Behav Addict. 2024;13(2):542–53.
8. Ruan H, Xun P, Cai W, et al. Habitual sleep duration and risk of childhood obesity: systematic review and dose-response meta-analysis of prospective cohort studies. Sci Rep. 2015;5:16160.
9. Xue M, Liu S, Zhang X, et al. Development and external validation of an interpretable machine learning-based model for obesity risk prediction in 2-18-year-old children and adolescents in Beijing and Tangshan. J Glob Health. 2026;16:04031.
10. Wang YJ, Xu YN, Peng J, et al. Multivariate machine learning model for long-term risk prediction of acute coronary syndrome in patients with heart failure with preserved ejection fraction and obstructive sleep apnea. J Am Heart Assoc. 2026;15(19):e050956.
11. Xu YN, Wang J, Zhou Z, et al. Multimodal prognostic model for predicting chronic coronary artery disease in patients without obstructive sleep apnea syndrome. Arch Med Res. 2024;55(1):102926.
12. Xu Q, Xu YN, Wang YJ, et al. Multiple machine-learning-driven metabolic frameworks for long-term prognostic risk assessment in patients with coexisting hypertension and obstructive sleep apnea: insights from a multicenter cohort study. Front Physiol. 2026;17:1739374.
13. Zhu J, Wang YJ, Duan SP, et al. Multivariable data-driven framework of predictive, preventive, and personalized medicine for long-term atrial fibrillation risk in patients with new-onset obstructive sleep apnea. EPMA J. 2025;16(4):761–72.
14. Xu YN, Duan SP, Yin WY, et al. Machine learning-enhanced prediction model for atrial fibrillation development in patients with concurrent type 2 diabetes and obstructive sleep apnea syndrome: a comorbidity perspective. Nutr Metab (Lond). 2025;22(1):158.
15. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56–67.
16. Zhang J, Clayton GL, Overvad K, et al. Body mass index in parents and their adult offspring: a systematic review and meta-analysis. Obes Rev. 2024;25(1):e13644.

17. Lee JS, Jin MH, Lee HJ. Global relationship between parent and child obesity: a systematic review and meta-analysis. Clin Exp Pediatr. 2022;65(1):35–46.
18. World Health Organization. Obesity and overweight: fact sheet. Geneva: WHO; 2025.
19. Working Group on Obesity in China (WGOC). Body mass index reference norm for screening overweight and obesity in Chinese children and adolescents. Chin J Epidemiol. 2004;25(2):97–102.
20. Chaput JP, Gray CE, Poitras VJ, et al. Systematic review of the relationships between sleep duration and health indicators in school-aged children and youth. Appl Physiol Nutr Metab. 2016;41(6 Suppl 3):S266–82.
21. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.
22. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
23. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785–794.
24. Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst. 2018;31:6638–48.
25. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765–74.
26. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. Ann Intern Med. 2017;167(4):268–74.
27. Flegal KM, Ogden CL, Yanovski JA, et al. Change in the distribution of body mass index in children and adolescents. JAMA. 2010;303(3):242–9.
28. Rolland-Cachera MF, Deheeger M, Bellisle F, et al. Adiposity rebound in children: a simple indicator for predicting obesity. Am J Clin Nutr. 1984;39(1):129–35.
29. Huang L, et al. Interpretable machine learning for identifying adolescent obesity risk and identifying key determinants. Front Public Health. 2026;14:1657467.
30. Liu Z, Gao P, Gao AY, et al. Effectiveness of a multifaceted intervention for prevention of obesity in primary school children in China: a cluster randomized clinical trial. JAMA Pediatr. 2022;176(1):e214375.
31. Miller MA, Kruisbrink M, Wallace J, et al. Sleep duration and incidence of obesity in infants, children, and adolescents: a systematic review and meta-analysis of prospective studies. Sleep. 2018;41(4):zsy018.
32. Dong Y, Yuan C, Dang J, et al. Control of childhood obesity and implications for policy in China. Lancet Public Health. 2024;9(12):e1125–35.
33. Zhang Y, Li X, Wang H, et al. Structural imbalance in youth weight management policies in China: a three-dimensional analysis of school responsibility and collaborative support. Front Public Health. 2026;14:1941280.
34. Biddle SJH, García Bengoechea E, Wiesner G. Sedentary behaviour and adiposity in youth: a systematic review of reviews and analysis of causality. Int J Behav Nutr Phys Act. 2017;14(1):43.
35. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128.9 million children, adolescents, and adults. Lancet. 2017;390(10113):2627–42.