BMC Public Health

RESEARCH

Open Access

Using interpretable machine learning methods to identify the relative importance of lifestyle factors for overweight and obesity in adults: pooled evidence from CHNS and NHANES

Zhiyuan Sun[1,2], Yunhao Yuan[3], Vahid Farrahi[4], Fabian Herold[5], Zhengwang Xia[6], Xuan Xiong[7], Zhiyuan Qiao[1], Yifan Shi[1], Yahui Yang[1], Kai Qi[1], Yufei Liu[8], Decheng Xu[1], Liye Zou[9*] and Aiguo Chen[1,2,10*]

Abstract

Background Overweight and obesity pose a huge burden on individuals and society. While the relationship between lifestyle factors and overweight and obesity is well-established, the relative contribution of specific lifestyle factors remains unclear. To address this gap in the literature, this study utilizes interpretable machine learning methods to identify the relative importance of specific lifestyle factors as predictors of overweight and obesity in adults.

Methods Data were obtained from 46,057 adults in the China Health and Nutrition Survey (2004–2011) and the National Health and Nutrition Examination Survey (2007–2014). Basic demographic information, self-reported lifestyle factors, including physical activity, macronutrient intake, tobacco and alcohol consumption, and body weight status were collected. Three machine learning models, namely decision tree, random forest, and gradient-boosting decision tree, were employed to predict body weight status from lifestyle factors. The SHapley Additive exPlanation (SHAP) method was used to interpret the prediction results of the best-performing model by determining the contributions of specific lifestyle factors to the development of overweight and obesity in adults.

Results The performance of the gradient-boosting decision tree model outperformed the decision tree and random forest models. Analysis based on the SHAP method indicates that sedentary behavior, alcohol consumption, and protein intake were important lifestyle factors predicting the development of overweight and obesity in adults. The amount of alcohol consumption and time spent sedentary were the strongest predictors of overweight and obesity, respectively. Specifically, sedentary behavior exceeding 28–35 h/week, alcohol consumption of more than 7 cups/week, and protein intake exceeding 80 g/day increased the risk of being predicted as overweight and obese.

*Correspondence:
Liye Zou
liyezou123@gmail.com
Aiguo Chen
agchen@nsi.edu.cn

Full list of author information is available at the end of the article

Conclusion Pooled evidence from two nationally representative studies suggests that recognizing demographic differences and emphasizing the relative importance of sedentary behavior, alcohol consumption, and protein intake are beneficial for managing body weight status in adults. The specific risk thresholds for lifestyle factors observed in this study can help inform and guide future research and public health actions.

Keywords Interpretable machine learning, Relative importance, Lifestyle, Overweight and obesity, Physical activity, Alcohol consumption

Background

Overweight and obesity are characterized by abnormal or excessive fat accumulation in the body, which increases the risk of adverse health outcomes [1]. Individuals with overweight and obesity are widely recognized as being at an increased risk for various non-communicable diseases such as type 2 diabetes, hypertension, and stroke, as well as various forms of cancer [2]. Therefore, overweight and obesity have become a fast-growing and serious public health issue in recent years [3]. Notably, the prevalence of overweight and obesity is rapidly increasing in both developing and developed countries [4, 5]. In particular, the World Health Organization (WHO) reported that in 2022, 1 in 8 people in the world was living with obesity. Among adults, 43% were overweight, and 16% were living with obesity [6]. Given the high prevalence and health consequences of overweight and obesity, these health conditions pose a huge burden on individuals and society [7].

Previous studies have found that a healthy lifestyle, including physical activity and a balanced diet, is crucial for weight management in adults [8]. Moreover, a healthy lifestyle is associated with higher odds of a metabolically healthy phenotype, which can reduce mortality risk and improve prognosis in overweight and obese adults [9, 10]. In contrast, an unhealthy lifestyle (e.g., lower levels of physical activity, and unbalanced diet) may increase the risk of overweight and obesity [11]. For example, in adults, lower levels of physical activity are associated with a greater increase in body mass index (BMI) [12]. In addition, big data analysis from nationwide health surveys has further confirmed the strong association between lifestyle factors and the development of overweight and obesity in adults [13, 14]. Accumulating evidence indicates that lifestyle factors, such as physical activity (or lack thereof), nutritional intake, tobacco use, and alcohol consumption, are closely related to overweight and obesity in adults. Specifically, spending more hours sedentary (i.e., being less physically active) is linked to a higher risk of overweight and obesity in adults, whereas higher levels of regular physical activity, both occupational and recreational, are associated with a lower risk [15, 16]. In addition, strong evidence suggests that nutrient intake in the daily diet influence the body weight status of adults [17]. In this context, an adequate intake of macronutrients, including carbohydrates, fats, and proteins, plays a critical role in preventing overweight and obesity in adults [18, 19]. The influence of tobacco and alcohol consumption on the development of overweight and obesity in adults has also been highlighted in the literature [20]. Compared with adults who have never smoked, current smokers have a lower prevalence of overweight and obesity [21]. However, higher alcohol consumption not only increases health risks [22, 23], but is also associated with a higher prevalence of overweight and obesity in adults [24]. In general, previous studies have identified lifestyle factors linked to overweight and obesity in adults, but traditional statistical methods often overlook the complexity and interactions of these factors.

Positive associations between a healthy lifestyle and lower development of overweight and obesity in adults have been confirmed in several studies [8, 16, 17]. However, most of these studies have focused on data from a single national health survey, and few studies have merged data from different databases, which limits the generalizability of their findings. More importantly, it is essential to recognize that different lifestyle factors may not be equally important for the development of overweight and obesity in adults [25]. For example, a higher amount of physical activity and less alcohol consumption are generally considered beneficial for preventing overweight and obesity, but their contributions are perhaps not of equal magnitude or, in the optimal case, may be synergistic. A study analyzing county-level obesity prevalence in the United States (US) also found that physical inactivity and diabetes prevalence are stronger predictors of obesity prevalence compared to poor mental health and being uninsured [26]. However, previous studies often use traditional statistical analysis methods that focus on linear relationships between specific lifestyle factors and the prevalence of overweight and obesity in adults [14, 24]. Thus, these previous studies are limited in their to unveil the unique and synergistic influence of specific lifestyle factors on the prevalence of overweight and obesity, including but not limited to the differences in the magnitude of the contributions. To address this limitation, interpretable machine learning methods offer an advantage as employing this analytical technique allows for revealing how individual lifestyle factors contribute to overweight and obesity.

Machine learning is an important research area in artificial intelligence [27], which aims to learn knowledge and

rules from complex data [28]. Compared with statistical analysis methods, machine learning focuses on achieving practical predictions and is generally better situated than the former in processing complex and big data sets [29]. Machine learning can predict future outcomes and trends and has been successfully applied in behavioral and health research [30]. For example, the decision tree (DT) algorithm has been used to predict physical activity behavior based on individual, demographic, psychological, behavioral, environmental, and physical factors [31]. In recent years, numerous studies using machine learning methods to analyze overweight and obesity-related outcomes have confirmed its effectiveness in identifying potentially high-risk individuals [32, 33].

Although more powerful than traditional statistical approaches, many machine learning techniques are often considered "black boxes" due to the difficulty in interpreting their predictions [34]. To better understand how models make predictions, scholars have proposed interpretable machine learning methods based on the SHapley Additive exPlanation (SHAP) [35]. SHAP is a unified interpretation method grounded in game theory [36] and provides global attributions that can explain the importance and impact of each feature affecting the model's output [37]. A recent study used machine learning algorithms, including logistic regression (LR), K-nearest neighbor, artificial neural network, DT, random forest (RF), gradient boosting machine, and CatBoost, to predict obesity risk in overweight adults. The results suggest that the last three tree-based algorithms achieved better accuracy, and the SHAP method identified waist and hip circumference as the strongest predictors of obesity risk [38]. Another study used LR, RF, XGBoost, and gradient-boosting decision tree (GBDT) algorithms to predict obesity risk. The SHAP analysis based on the RF model (best-performing) identified several important personal factors, such as self-awareness and body weight control experience [39]. Interpretable machine learning provides a new approach to understanding the important factors in the development of overweight and obesity in adults. However, to our knowledge, no study has applied the SHAP method to investigate whether different lifestyle factors can predict body weight status and to identify their relative importance for developing overweight and obesity in adults.

To address this gap in the literature, this study will pool data on lifestyle and body measurements from two different population-based and nationally representative databases and use interpretable machine learning methods to identify the relative importance of lifestyle factors in predicting overweight and obesity status in adults. Our findings will offer new insights for a more in-depth understanding of the association between specific lifestyle factors and overweight and obesity in adults that, in turn, may help to inform the development of more efficient prevention and intervention approaches.

Methods

Study sample

In this study, we selected China and the US as representative countries for developing and developed countries, respectively. In China, about 50% of adults are overweight or obese [40], whereas in the US, the obesity prevalence is much higher than the global average (16%), exceeding 40% [41]. Furthermore, these countries represent distinct economic and lifestyle environments, providing a diverse dataset for robust model evaluation. The China Health and Nutrition Survey (CHNS) and the National Health and Nutrition Examination Survey (NHANES) are two freely accessible, high-quality, and widely used nationwide surveys that encompass a broad range of lifestyle factors and health data.

CHNS is designed to examine the effects of health, nutrition, and family planning policies and programs on the Chinese population. It has been conducted since 1989, with the latest data collected in 2015. However, surveys before 2004 lacked data on sedentary behavior (i.e., time spent sedentary), and the 2015 survey lacked data on nutrient intake. Therefore, this study utilized data from the available surveys conducted in 2004, 2006, 2009, and 2011. The CHNS was approved by the University of North Carolina at Chapel Hill and the Chinese Center for Disease Control and Prevention, and all participants provided written informed consent. More details are available at https://www.cpc.unc.edu/projects/china.

NHANES is designed to assess the health status and behaviors of the US population through interviews and physical examinations. NHANES began with three discrete surveys and has since evolved into a continuous program, with data released every two years. However, surveys conducted before 2007–2008 did not include explicit questions about "sedentary behavior." Therefore, we selected four consecutive NHANES surveys, including data from 2007 to 2008, 2009–2010, 2011–2012, and 2013–2014. NHANES was approved by the National Center for Health Statistics Research Ethics Review Board, and all participants provided informed consent. More details are available at https://www.cdc.gov/nchs/nhanes/index.htm.

This study pooled data from CHNS and NHANES and excluded participants who (i) had any missing data, (ii) refused to answer or answered "don't know," and (iii) gave contradictory responses, such as answering "no" to the question "Do you participate in this activity?" but then reporting participation time for "How much time do you spend during a typical day?" Finally, data from 46,057 adults were included, with 30,666 from CHNS and 15,391 from NHANES. The flowchart of the sample selection is

shown in Fig. 1. Definitions of demographic characteristics, lifestyle factors, and body weight status are detailed in the following subsections.

Demographic information

Basic demographic information was collected, including country, race/ethnicity, sex, and age. Han adults were included, comprising over 85% of the total adults in CHNS. Hispanic, Non-Hispanic White, and Non-Hispanic Black adults from NHANES were also included. Sex was categorized as either male or female. Given that certain questionnaires in NHANES, such as the 2009–2010 Alcohol Use questionnaire, only disclosed data from participants aged 20 and above, we included only adults aged ≥ 20 in both CHNS and NHANES.

Lifestyles

Physical activity in adults was quantified using metabolic equivalent tasks (METs) from the updated compendium of physical activities [42]. METs provide a standardized method for comparing energy expenditure across activities, making it suitable for cross-national comparison. Time spent on different physical activities was multiplied by the corresponding METs score to calculate the amount of physical activity related to occupation, transportation, and recreation, which were then added to obtain the total physical activity level (MET h/week). To avoid over-representation of an activity in questions containing multiple specific activities, the mean METs score was weighted [43] by the reciprocal of the number of different physical activities. In addition, sedentary behavior (h/week) was quantified as time spent [44], such as watching television and using a computer.

Nutrient intake was quantified using daily dietary information obtained from interviews. To account for differences in the frequency of dietary data collection between CHNS and NHANES, the average nutrient intake was weighted by the reciprocal of the number of assessments [45, 46]. Dietary intakes included carbohydrates (g/day), fats (g/day), and proteins (g/day).

Tobacco consumption-related variables included smoking status and smoking amount. Smoking status was categorized as never smoker, former smoker, and current smoker [21]. In addition, smoking amount (cigarettes/day) was recorded, ranging from 1 to 95 for current smokers, and defined as 0 for never-smokers and former smokers.

Alcohol consumption-related variables included self-reported information on the frequency and amount of alcohol consumption. Alcohol consumption frequency was redefined as daily, 3–5 times/week, 1–3 times/week, 1–4 times/month, rarely, and none. The amount of alcohol consumption (cups/week) was calculated based on the approximate conversion of different types of alcohol.

A more detailed description of the assessment of lifestyle factors can be found in the supplementary material (Additional file S1).

Body weight status

BMI was used to evaluate body weight status in adults, calculated as weight (kg) / height (m^2). Body weight status in CHNS and NHANES was defined according to country-specific BMI standards [47, 48]. For the Chinese sample, 18.5 ≤ BMI < 24 was defined as normal body weight, 24 ≤ BMI < 28 as overweight, and BMI ≥ 28 as obese. For the American sample, 18.5 ≤ BMI < 25 was defined as normal body weight, 25 ≤ BMI < 30 as overweight, and

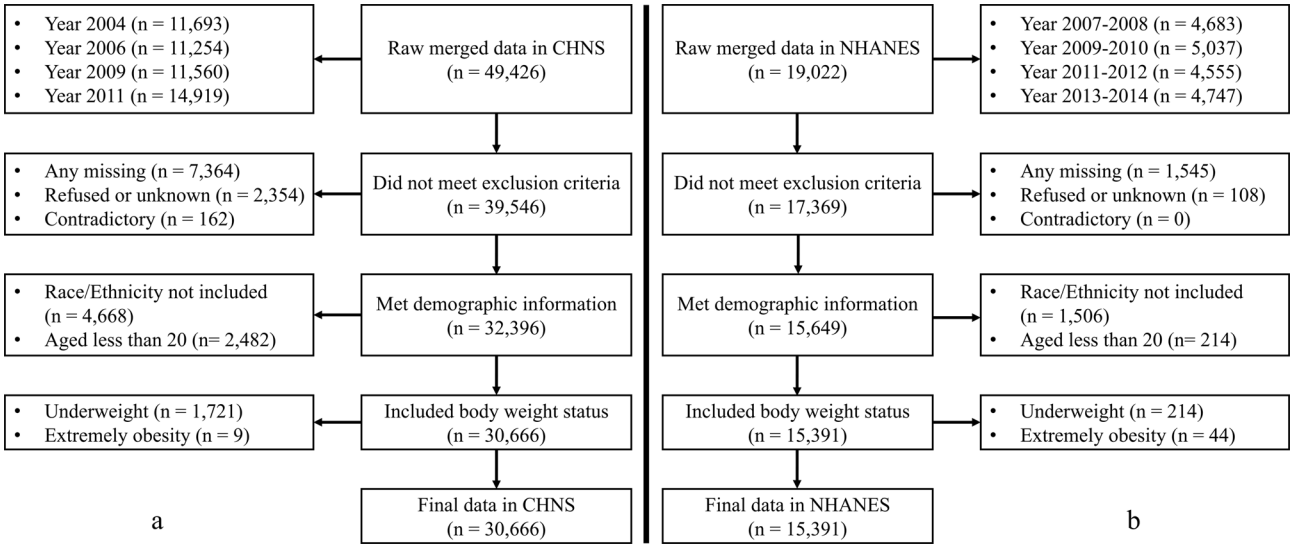

Fig. 1 Participants flowchart. *Note* Flowchart of the sample selection in CHNS (**a**) and NHANES (**b**). CHNS = China Health and Nutrition Survey; NHANES = National Health and Nutrition Examination Survey

BMI ≥ 30 as obese. Underweight with a BMI < 18.5 and extremely obese with a BMI > 60 were excluded.

Machine learning models

Machine learning models were established to predict body weight status from lifestyle characteristics, and the prediction results of the best-performing model were further explained to identify the relative importance of lifestyle factors for overweight and obesity in adults. Especially, to explore the possible different contributions of lifestyle factors to overweight and obesity, we established two types of binary classification models: one for predicting overweight and another for predicting obesity in adults. In addition, given the sample imbalance among normal body weight (n = 21,180), overweight (n = 15,327), and obesity (n = 9,550), we performed random sampling from the normal body weight group to reconstruct the dataset for overweight (normal and overweight, n = 30,654) and obesity (normal and obesity, n = 19,100) [49]. A flowchart visualizing the modeling is shown in Fig. 2.

Algorithms

Three tree-based algorithms, namely DT, RF, and GBDT, were selected to establish machine learning models. On the one hand, these algorithms have been successfully applied and have performed well in predicting overweight and obesity [50–52]. On the other hand, tree-based models offer better interpretability and are easier to understand in their decision-making process [53].

DT is one of the classic machine learning algorithms [54]. It recursively splits the dataset into smaller subsets and classifies them based on feature values. DT excels in classification tasks due to its simple structure and capacity to handle various features. RF is an ensemble learning method based on Bagging, which can include hundreds

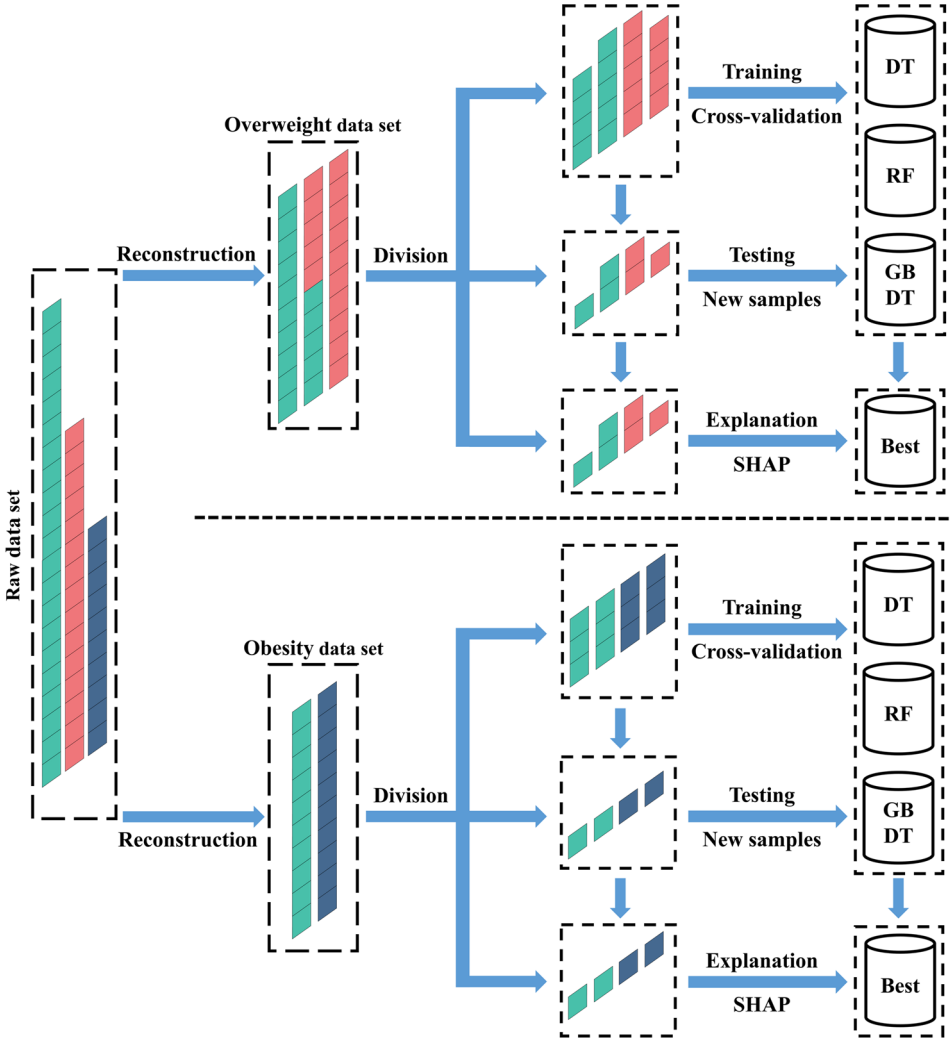

Fig. 2 Flowchart visualizing the modeling procedures. *Note* DT = Decision Tree; RF = Random Forest; GBDT = Gradient Boosting Decision Tree; SHAP = SHapley Additive exPlanation; Best = Best-performing

or thousands of decision trees [55]. It uses the Bootstrap method to randomly extract different sample subsets from the dataset multiple times to train multiple classifiers. Its final classification result is determined by a majority vote among the classifiers, with each classifier given equal weight in the voting process. GBDT is an ensemble learning method based on Boosting [56]. GBDT iteratively trains weak classifiers to minimize the loss function and combines them into a strong classifier. Its final classification result is determined by weighted voting, where classifiers with better prediction performance receive greater weight.

Training, testing, and explanation

The dataset was randomly divided into a training set, a test set, and an additionally defined explanation set in a 3:1:1 ratio, with each subset assigned different specific functions. For the overweight dataset, it was divided into an overweight training set (60%, $n=18,392$), an overweight test set (20%, $n=6,131$), and an overweight explanation set (20%, $n=6,131$). The obesity dataset was divided into an obesity training set (60%, $n=11,460$), an obesity test set (20%, $n=3,820$), and an obesity explanation set (20%, $n=3,820$).

Training

The training set was used to train the machine learning model. Given the randomness of the calculation, we combined the 10-fold cross-validation method in the training process [57]. It randomly splits the training set into 10 subsets of similar size and then performs 10 rounds of model training and evaluation. In each round, 9 subsets are selected for training and the remaining 1 subset is used for evaluation. Ten model performance evaluation indicators are obtained, and the average of the indicators is taken to evaluate the model performance. In addition, we were more concerned with the model's performance in predicting individuals as overweight adults and obese adults (positive class, label=1), rather than normal body weight adults (negative class, label=0). Therefore, precision was chosen as the key indicator to evaluate the model's performance in the training set. Precision refers to the proportion of samples predicted as positive by the model that are correctly predicted. The closer the precision is to 1, the more reliable the results predicted as positive by the model are.

Testing

The test set was used to assess the generalization ability of the machine learning model on new samples and select the best-performing model in predicting adults' body weight status from lifestyle factors. We calculated the precision and area under the ROC curve (AUC) score to evaluate the model's performance on the test set. In addition, for the best-performing model, we visualized its decision-making process. AUC is a comprehensive indicator that combines the true positive rate (TPR) and the false positive rate (FPR) of the model. TPR refers to the proportion of samples that are actually positive and are correctly predicted as positive by the model. FPR refers to the proportion of samples that are actually negative and are incorrectly predicted as positive by the model. The closer the AUC is to 1, the better the overall performance of the model.

Explanation

More importantly, an additional explanation set was defined specifically to retest the performance of the best-performing model and explain its predictions. The SHAP method was used to visualize the relative importance of various lifestyle factors in affecting the model's output, identifying which factors can influence the model's prediction of individuals as overweight and obese adults. SHAP calculates the Shapley value of each feature by traversing all possible subsets in the feature space to obtain the SHAP value of each feature. Analyzing these values allows us to understand the importance of each feature and the impact of different feature values on the model's output [53]. In addition, the Shapely Lorenz value was calculated, which uses Lorenz Zonoid decomposition and Partial Gini Contribution to measure the contribution of each feature, ensuring that the Shapely values are standardized [58].

Results

Participant characteristics

Compared to normal body weight adults, overweight and obese adults exhibited lower levels of occupation-related physical activity, total physical activity, carbohydrate intake, and smoking amount. Conversely, they had higher levels of transportation-related physical activity, recreation-related physical activity, time spent sedentary, fat intake, protein intake, and alcohol consumption (Table 1). In the overweight and obesity datasets, there were statistically significant differences in country, race/ethnicity, sex, and age factors among adults with different body weight statuses ($ps<0.05$). All lifestyle characteristics between overweight or obese adults and those with normal body weight were different at the 99% confidence level when demographic covariates were not considered ($ps<0.01$).

The relationship between body weight status and lifestyle factors was analyzed with consideration of demographic covariates. In the overweight dataset, body weight status showed negative correlations with occupation-related physical activity ($r=-0.022$), transportation-related physical activity ($r=-0.017$), total physical activity ($r=-0.023$), carbohydrate intake ($r=-0.017$), smoking

Table 1 Participant characteristics

Characteristics	Normal versus Overweight		Normal versus Obesity	
	Normal (*n* = 15,327)	Overweight (*n* = 15,327)	Normal (*n* = 9,550)	Obesity (*n* = 9,550)
Country[a]				
China	12,497 (81.5, 0)	10,036 (65.5, 0)	7,796 (81.6, 0)	3,340 (35, 0)
United States	2,830 (18.5, 1)	5,291 (34.5, 1)	1,754 (18.4, 1)	6,210 (65, 1)
Race/Ethnicity				
Han	12,497 (81.5, 0)	10,036 (65.5, 0)	7,796 (81.6, 0)	3,340 (35, 0)
Hispanic	614 (4, 1)	1,551 (10.1, 1)	396 (4.1, 1)	1,688 (17.7, 1)
Non-Hispanic White	1,653 (10.8, 2)	2,716 (17.7, 2)	993 (10.4, 2)	2,842 (29.8, 2)
Non-Hispanic Black	563 (3.7, 3)	1,024 (6.7, 3)	365 (3.8, 3)	1,680 (17.6, 3)
Sex				
Male	6,991 (45.6, 0)	7,734 (50.5, 0)	4,296 (45, 0)	4,126 (43.2, 0)
Female	8,336 (54.4, 1)	7,593 (49.5, 1)	5,254 (55, 1)	5,424 (56.8, 1)
Age (year)				
20–40	4,718 (30.8, 0)	3,398 (22.2, 0)	2,990 (31.3, 0)	2,402 (25.2, 0)
40–60	6,394 (41.7, 1)	7,163 (46.7, 1)	3,952 (41.4, 1)	3,964 (41.5, 1)
60-	4,215 (27.5, 2)	4,766 (31.1, 2)	2,608 (27.3, 2)	3,184 (33.3, 2)
Physical activity (MET h/week)[b]				
Occupation	89.9 (87.6, 92)	73.3 (71.5, 75.3)	90.1 (87.5, 92.8)	52.9 (50.8, 55)
Transportation	2.7 (2.5, 2.9)	3.3 (3.1, 3.6)	2.6 (2.4, 2.9)	4.6 (4.1, 5)
Recreation	6.8 (6.4, 7.1)	8.4 (8, 8.8)	6.9 (6.5, 7.3)	8.3 (7.9, 8.8)
Total	99.4 (97.2, 101.5)	85.1 (83.2, 87)	99.6 (96.9, 102.4)	65.8 (63.7, 68.1)
Sedentary (h/week)	23.5 (23.2, 23.8)	27.3 (26.9, 27.6)	23.5 (23, 23.9)	35.2 (34.7, 35.8)
Macronutrient intake (g/day)				
Carbohydrate	282 (280.2, 284)	272.3 (270.4,274.1)	283 (280.9, 285.2)	253.9 (251.9, 256)
Fat	73.9 (72.9, 75.1)	77.5 (75.8, 79.9)	73.2 (71.9, 74.8)	75.7 (75, 76.5)
Protein	67.8 (67.3, 68.3)	72.2 (71.7, 72.6)	67.9 (67.3, 68.4)	74.7 (74, 75.2)
Smoking status				
Never	10,128 (66.1, 0)	9,830 (64.1, 0)	6,352 (66.5, 0)	5,980 (62.6, 0)
Former	1,039 (6.8, 1)	1,962 (12.8, 1)	639 (6.7, 1)	1,812 (19, 1)
Current	4,160 (27.1, 2)	3,535 (23.1, 2)	2,559 (26.8, 2)	1,758 (18.4, 2)
Smoking amount (cigarettes/day)	4.4 (4.2, 4.5)	3.5 (3.4, 3.6)	4.3 (4.1, 4.5)	2.6 (2.4, 2.7)
Alcohol consumption frequency				
Daily	1,574 (10.3, 0)	1,567 (10.2, 0)	940 (9.8, 0)	597 (6.3, 0)
3–5 times/week	676 (4.4, 1)	878 (5.7, 1)	415 (4.3, 1)	475 (5, 1)
1–3 times/week	1,387 (9, 2)	1,759 (11.5, 2)	839 (8.8, 2)	1,092 (11.4, 2)
1–4 times/month	1,254 (8.2, 3)	1,541 (10.1, 3)	785 (8.2, 3)	1,322 (13.8, 3)
Rarely	737 (4.8, 4)	1,149 (7.5, 4)	470 (4.9, 4)	1,396 (14.6, 4)
None	9,699 (63.3, 5)	8,433 (55, 5)	6,101 (63.9, 5)	4,668 (48.9, 5)
Alcohol consumption amount (cups/week)	5.1 (5, 5.4)	7 (6.7, 7.2)	5.1 (4.9, 5.3)	9.3 (9, 9.6)

Note [a]Categorical variables are presented as sample counts (weighted %, label); [b] Continuous variables are presented as mean (95% confidence interval). MET = Metabolic Equivalent Task

status ($r=-0.05$), and smoking amount ($r=-0.052$), and showed positive correlations with time spent sedentary ($r=0.025$), fat intake ($r=0.017$), and protein intake ($r=0.041$). In the obesity dataset, body weight status exhibited negative correlations with occupation-related physical activity ($r=-0.031$), transportation-related physical activity ($r=-0.028$), recreation-related physical activity ($r=-0.035$), total physical activity ($r=-0.041$), carbohydrate intake ($r=-0.036$), smoking status ($r=-0.055$), and smoking amount ($r=-0.056$), and showed positive correlations with time spent sedentary ($r=0.051$), fat intake ($r=0.018$), protein intake ($r=0.026$), and alcohol consumption frequency ($r=0.033$). All reported correlation analyses were statistically significant at the 95% confidence level ($ps<0.05$) and were weak ($rs<0.1$).

Prediction performance of machine learning models

In the overweight dataset, the precision of the DT, RF, and GBDT models in the training set was 62.3%, 64.2%,

and 64.3%, respectively. Furthermore, we calculated the precision and AUC scores for these models in the overweight test set. The precision of the DT, RF, and GBDT models was 62.5%, 64.8%, and 65.2%, respectively, while the AUC was 57%, 58.2%, and 58.3%, respectively. The GBDT model was considered the best-performing model in classifying normal body weight and overweight adults from lifestyle factors. We visualized the decision-making process of a decision tree within the GBDT model. As shown in Fig. 3a, this tree contained 6,131 samples, and the root node used age ≤ 0.5 as the splitting point. Among them, 1,696 samples (age = 0, 20–40 years) that met this splitting point were assigned to the left child node and split again with race/ethnicity ≤ 0.5 as the new splitting point. The log odds of the samples in this child node being predicted as positive was −0.1, resulting in a transformed probability of 0.48 via the sigmoid function, with the majority class being negative. On the other hand, 4,435 samples (age = 1 or 2, over 40 years) that did not meet the root node splitting point were assigned to the right child node, and continued to be split downward according to occupation-related physical activity ≤ 232.3 until they reached the leaf nodes.

In the obesity dataset, the precision of the DT, RF, and GBDT models in the training set was 77%, 78.6%, and 78.8%, respectively. We also calculated the precision and AUC scores for these models in the obesity test set. The precision of the DT, RF, and GBDT models was 77.4%, 78.1%, and 78.5%, respectively, while the AUC was 72.6%, 73.1%, and 73.6%, respectively. The GBDT model indicated strong performance, particularly in distinguishing obese from normal body weight adults, and the decision-making process of one of its trees is shown in Fig. 3b. This tree contained 3,056 samples, with the root node using carbohydrate intake ≤ 148.5 as the splitting point. The left child node contained 332 samples that met this criterion and continued to be split according to age ≤ 0.5. The log odds of the samples in this child node was 0.1, resulting in a transformed probability of 0.52, with the majority class being positive. The right child node contained 2,724 samples that did not meet the root node criterion and used sedentary activity ≤ 63.6 as the new splitting point. The tree splits were terminated upon reaching the maximum depth, leaving the nodes as leaf nodes.

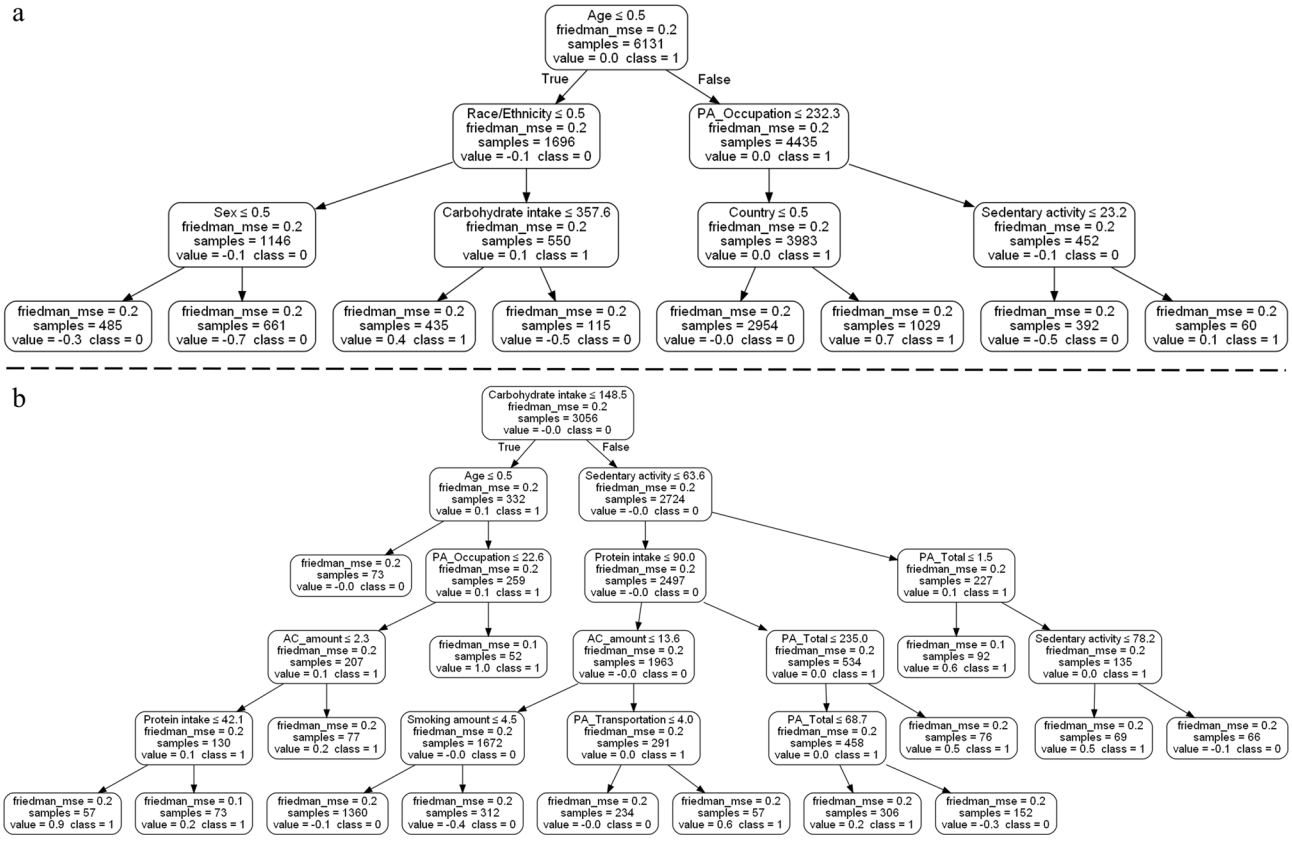

Fig. 3 Decision-making process of machine learning models. *Note* a: Decision-making process in the overweight test set. b: Decision-making process in the obesity test set. The value represents the log odds of predicting the samples in the node as positive, and the class indicates the majority class (1 = positive; 0 = negative) of the samples in the node. Country, race/ethnicity, sex, and age are categorical variables. For example, age ≤ 0.5 results in two branches: one for 0 (20–40 years), and the other for 1 (40–60 years) and 2 (over 60 years). PA = Physical Activity; AC = Alcohol Consumption

The relative importance of lifestyle factors for overweight and obesity in adults

In the overweight explanation set, the precision and AUC of the GBDT (best-performing) model were 66.9% and 59.3%, respectively. Figure 4a describes the 10 features that have the greatest influence on the model's output, ranked as country, age, race/ethnicity, alcohol consumption amount, smoking amount, sex, protein intake, sedentary behavior, occupation-related physical activity, and alcohol consumption frequency. Alcohol consumption amount was considered the most critical lifestyle factor of overweight in adults, and as its feature value increases, the model will tend to predict individuals as overweight adults. Lower smoking amount, higher protein intake, more time spent sedentary, lower occupation-related physical activity, and higher alcohol consumption frequency were also important factors for being predicted as overweight adults (Fig. 4b). Specifically, consuming more than 7 cups of alcohol/week, smoking less than 2 cigarettes/day, consuming more than 80 g protein/day, spending more than 28 h/week in sedentary behaviors, engaging in less than 7.5 MET-h/week of occupation-related physical activity, and any frequency of alcohol

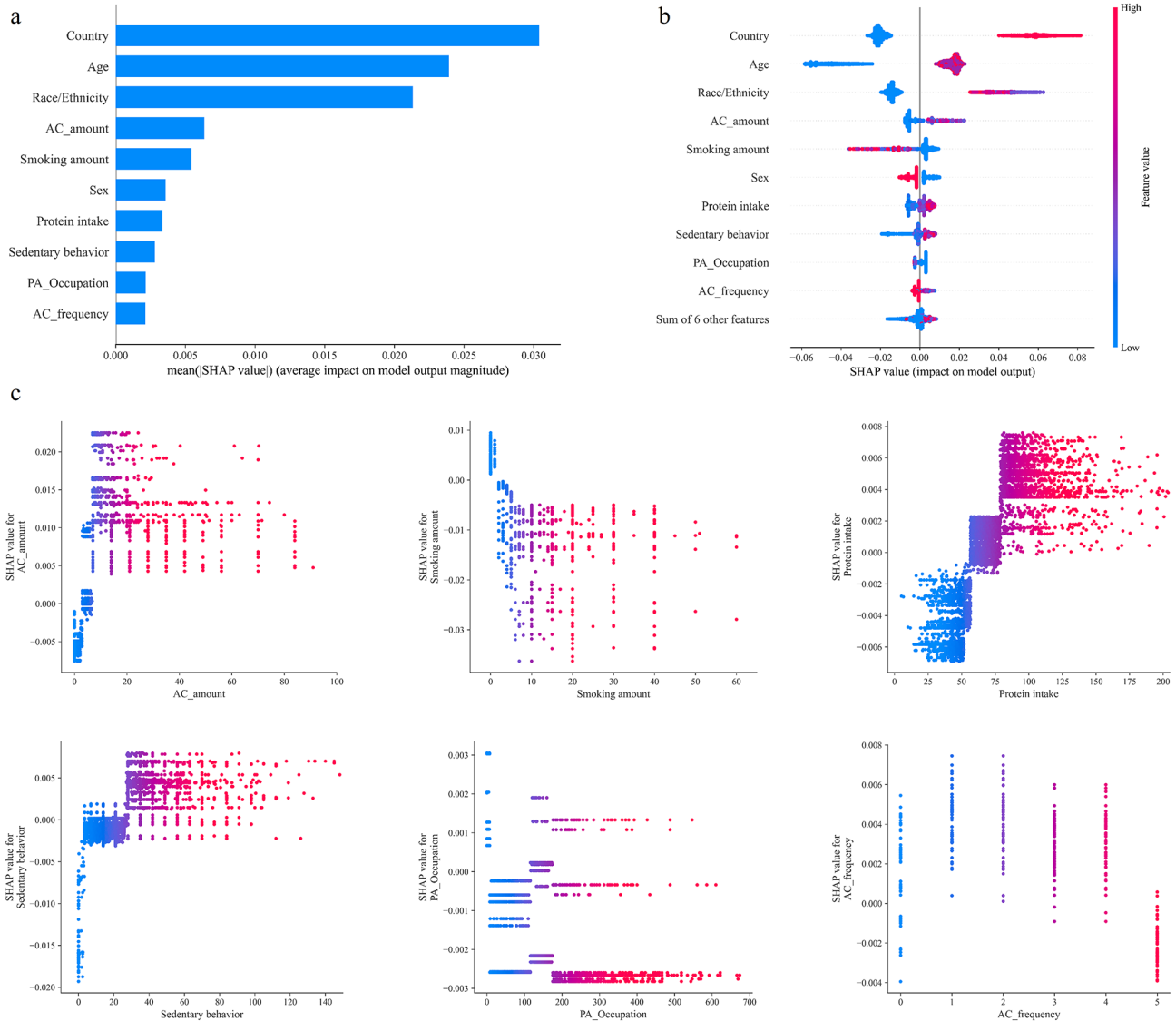

Fig. 4 The relative importance of lifestyle factors for overweight in our sample of adults. *Note* a: Feature importance; b: SHAP results; c: Impacts of specific features on the model's output. From top left to bottom right: alcohol consumption amount, smoking amount, protein intake, sedentary behavior, occupation-related physical activity, and alcohol consumption frequency. The horizontal axis represents the actual value of the specific feature, while the vertical axis represents the SHAP value corresponding to the feature (i.e., the impact of the feature on the model's output; a positive value indicates a positive influence, while a negative value indicates a negative influence). In b and c, each point represents a sample, and the color indicates the feature value (red and blue corresponding to high and low values, respectively). PA = Physical Activity; AC = Alcohol Consumption

consumption can increase the risk of being predicted as overweight adults (Fig. 4c).

In the obesity explanation set, the precision and AUC scores of the GBDT model were 77.8% and 73%, respectively. As shown in Fig. 5a, the factors that have the greatest impact on the model's output were ranked as country, race/ethnicity, age, sedentary behavior, smoking amount, alcohol consumption amount, carbohydrate intake, recreational-related physical activity, protein intake, and total physical activity. Time spent sedentary was considered the most critical lifestyle factor of obesity in adults, and as its feature value increases, the model tends to predict individuals as obese adults. Lower smoking amount, higher alcohol consumption amount, lower carbohydrate intake, lower recreational-related physical activity, and higher protein intake were also important factors for being predicted as obese adults (Fig. 5b). Specifically, spending more than 35 h/week in sedentary behaviors, smoking less than 2 cigarettes/day, consuming more than 7 cups of alcohol/week, having a carbohydrate intake of less than 225 g/day, engaging in less than 3 MET-h/week of recreational-related physical activity, and consuming more than 80 g protein/day have been

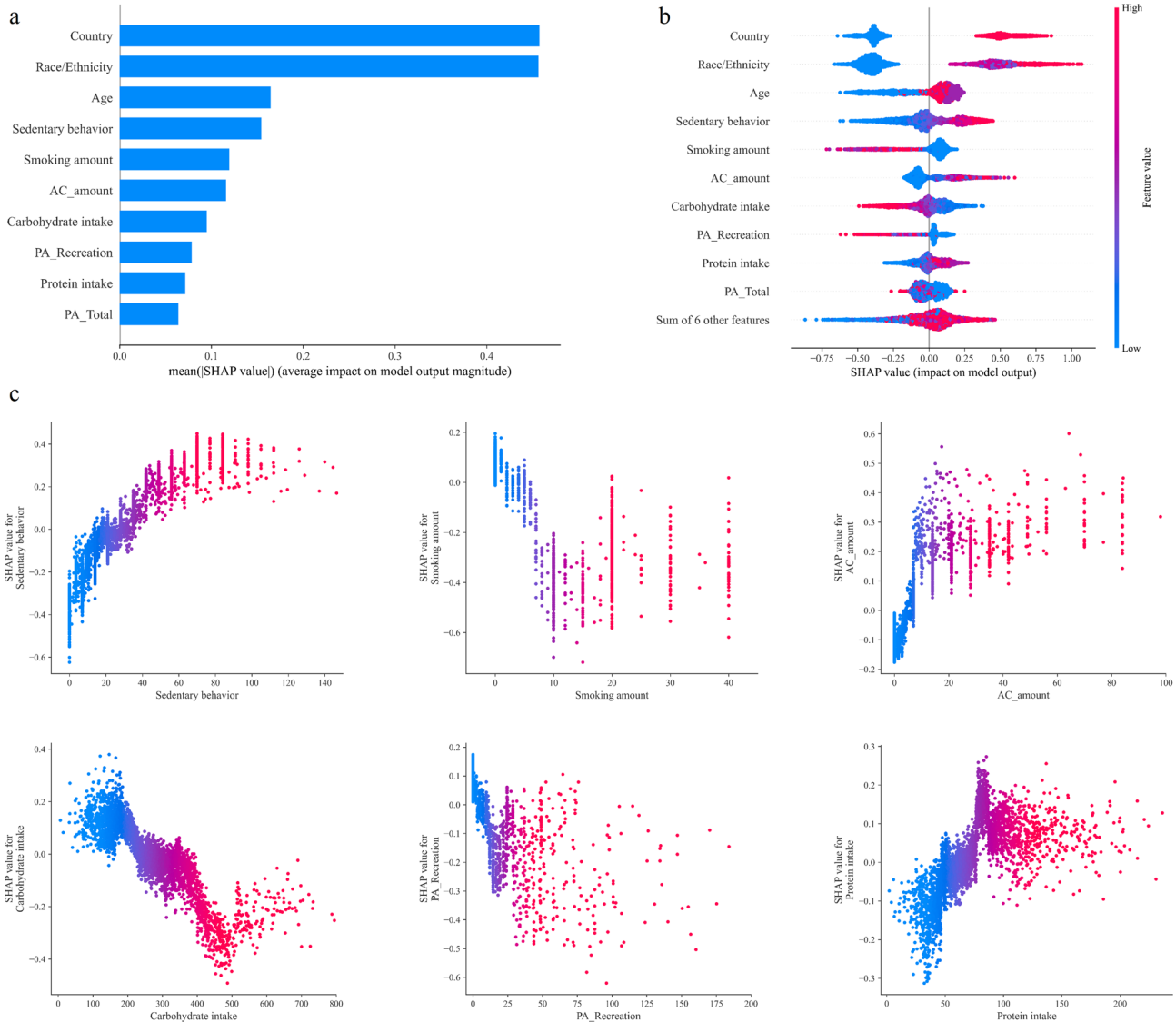

Fig. 5 The relative importance of lifestyle factors for obesity in adults. *Note* a: Feature importance; b: SHAP results; c: Impacts of specific features on the model's output. From top left to bottom right: sedentary behavior, smoking amount, alcohol consumption amount, carbohydrate intake, recreational-related physical activity, and protein intake. The horizontal axis represents the actual value of the specific feature, while the vertical axis represents the SHAP value corresponding to the feature (i.e., the impact of the feature on the model's output; a positive value indicates a positive influence, while a negative value indicates a negative influence). In b and c, each point represents a sample, and the color indicates the feature value (red and blue corresponding to high and low values, respectively). PA = Physical Activity; AC = Alcohol Consumption

observed to increase the risk of being predicted as obese adults (Fig. 5c).

In addition, the Shapely Lorenz values of the GBDT model in the overweight and obesity datasets were calculated, which are described in detail in the supplementary material, as shown in Additional file S2.

Discussion

Given that different lifestyle factors may not be equally important for the development of overweight and obesity in adults, this study used interpretable machine learning methods to identify the relative importance of specific lifestyle factors for being predicted as overweight and obese adults. Pooled data from two large-scale population-based studies, encompassing over 40,000 adults, indicated that time spent sedentary, amount of alcohol consumption and smoking, and protein intake were important lifestyle factors that are associated with overweight and obesity in adults. Notably, higher levels of alcohol consumption and sedentary behavior were the strongest predictors of being classified as overweight and obese, respectively. These findings align with previous studies that emphasize sedentary behavior and alcohol consumption as critical targets for obesity prevention and intervention [16, 24].

Regarding demographic factors, our study provides evidence that country, race/ethnicity, and age are important determinants of overweight and obesity in adults. Firstly, living in a high-income country typically fosters a specific — also referred to as "Western" lifestyle (e.g., the ability to purchase high-calorie foods and high time spent in sedentary behaviors) that contributes to a higher prevalence of overweight and obesity [59]. In low- and middle-income countries, affected by processes such as globalization and urbanization, traditional eating habits are increasingly being replaced by Western high-calorie diets [60]. Secondly, recent studies suggest that genetic differences among racial/ethnic groups may influence metabolic rate, energy expenditure, and fat storage, which in turn is reflected in the prevalence rates of overweight and obesity [61]. Moreover, racial/ethnic groups often differ in terms of socioeconomic status and public resource allocation. For instance, lower socioeconomic status is often associated with a higher prevalence of obesity, as it may limit access to healthy foods and opportunities for physical activity [62]. Thirdly, the prevalence of overweight and obesity in adults increases with age [63], though it may decrease in older adults due to health issues (e.g., malnutrition) [64]. Specifically, the prevalence of obesity is highest among middle-aged adults, which may be related to work pressure, which may result in a lack of physical activity, and irregular eating habits [65]. Taken together, the findings of this study support the notion that a strong association exists between demographic factors and overweight and obesity status. Future research should consider other demographic factors (e.g., socioeconomic status [62]) to advance our understanding of the relationship between lifestyle factors and overweight and obesity in adults.

More importantly, our findings indicate that time spent sedentary, amount of alcohol consumption and smoking, and protein intake are important lifestyle factors associated with overweight and obesity in adults. To begin with, this observation aligns with findings from a systematic review that reported a relationship between higher levels of sedentary behavior and weight gain in adults [66]. Prolonged periods of sedentary behavior reduce daily energy expenditure. When energy expenditure is lower than energy intake, unconsumed energy is stored as fat, contributing to the development of overweight and obesity [44]. Sedentary behavior may also decrease the activity of lipoprotein lipase, which can negatively affect fat metabolism and utilization, thereby exacerbating fat accumulation and weight gain [67]. Besides, there is substantial evidence that alcohol consumption is associated with overweight and obesity [68], which is consistent with our findings. Alcohol is inherently high in energy, and its consumption increases total energy intake, particularly when paired with high-energy foods, potentially leading to an energy surplus and subsequent weight gain [69]. Alcohol consumption also influences neurotransmitters involved in appetite regulation, which can lead to increased consumption of energy-dense foods [70]. In contrast to alcohol consumption, our study found that tobacco consumption is negatively associated with overweight and obesity. Evidence suggests that smoking may reduce body weight by elevating the resting metabolic rate while diminishing the expected increase in food intake associated with this metabolic increase [71]. However, since smoking is a significant risk factor for adverse health outcomes [72], using smoking as a weight management strategy is not recommended for adults. Finally, we observed that excessive protein intake is associated with overweight and obesity in adults. It is important to note that the evidence regarding the relationship between protein intake and overweight and obesity in adults is mixed. Previous studies have indicated that a high-protein diet increases energy expenditure and satiety, which can aid in weight loss [73, 74]. However, other studies observed that protein intake is positively correlated with overweight and obesity [75, 76]. For example, a cross-sectional study that examined the relationship between macronutrient intake and adiposity in adults found a positive correlation between protein intake and BMI, body fat percentage, sagittal abdominal diameter, and waist circumference in men [75]. Another survey of the Chinese adults population also emphasized that higher protein and fat intake are associated with an increased risk of being overweight

and obese [76]. These mixed findings may partly result from the effects of different protein sources (animal versus plant protein) on overweight and obesity [77]. Thus, further studies are needed to clarify the underlying biological mechanisms that may explain this phenomenon.

Additionally, three machine learning models, namely DT, RF, and GBDT, were established to predict the body weight status in adults. The precision and AUC of the GBDT (best-performing) model were 65.2% and 58.3% in the overweight test set, and 78.5% and 73.6% in the obesity test set. Lin et al. developed nine machine learning models to identify relevant risk factors for overweight [78]. The results suggest that the CatBoost model achieved a precision of 81% in the test set, outperforming other models, such as RF, SVM, and LR. This model also outperformed the GBDT model used in our study. The potential reason may be that the association between different variables is an important factor determining the performance of the machine learning models. On the one hand, compared with the waist and hip circumference factors included in the CatBoost model, lifestyle factors may be relatively minor contributors to predicting overweight in adults. On the other hand, Cheng et al. used multiple machine-learning methods to predict adult obesity from physical activity levels [79]. The random subspace and classification via regression models achieved the highest AUC of 64.3%. The performance of these two models in predicting obesity was lower than that of the best-performing model in our study. This is not surprising, as our study included a more comprehensive set of lifestyle factors, with physical activity or its absence (i.e., time spent sedentary) considered as only single factors in our analysis. This also indicates that including more relevant factors in the models can improve their performance. In addition, future research can refer to the SAFE framework [80] to comprehensively measure the application of artificial intelligence from the perspectives of sustainability, accuracy, fairness, and explainability.

Finally, our study provides thresholds for how exposure to specific lifestyle factors is associated with overweight and obesity in adults. Specifically, spending more than 28–35 h/week in sedentary behaviors, consuming more than 7 cups of alcohol/week, and consuming more than 80 g of protein/day were observed to increase the risk of being predicted as overweight and obese adults. In addition, insufficient occupational and recreational physical activity, higher carbohydrate intake, and any level of alcohol consumption also increased the risk of being predicted as overweight or obese. In this context, the WHO has emphasized the importance of a healthy lifestyle for body weight management, which includes limiting screen time and engaging in regular physical activity [6]. The weight management strategies recommended by the Dietary Guidelines for Chinese Adults include engaging in moderate-intensity aerobic exercise for 2.5–5 h/week, limiting sedentary activity to less than 2–4 h/day, consuming no more than 70 g of livestock and poultry meat/day, and strictly limiting alcohol consumption [81]. The obesity strategies proposed by the Centers for Disease Control and Prevention include following the Dietary Guidelines for Americans, engaging in moderate-intensity physical activity for at least 2.5 h/week, sleeping for 7–9 h/day, and managing stress effectively [82]. Compared with these guidelines, our findings are more liberal. In particular, our findings suggest (i) that time spent in sedentary behavior should be limited to 28–35 h/week, which is higher than the guidelines, and (ii) that alcohol consumption should be limited to no more than 7 cups/week, rather than being completely restricted. These findings suggest that also less restrictive lifestyle guideline criteria are perhaps effective in preventing overweight and obesity, although future studies in this direction are required to draw more robust conclusions. In addition, the risk thresholds for various types of physical activity (e.g., recreational) and specific macronutrient intakes (e.g., protein) were quantified in our study. These unique contributions of our work may offer valuable information for updating and refining body weight management strategies in adults.

The following limitations need to be considered when interpreting the findings of our study. Firstly, predicting body weight status from lifestyle factors can be challenging because of the myriad of factors influencing overweight and obesity that were not assessed by the surveys or included in our models (e.g., sleep duration is associated with body weight status [83], but was not assessed in CHNS). In addition, given that the definition of the confounding variables (e.g., socioeconomic status) differ between the two databases, it was difficult to specifically control for the effects of these factors. Secondly, using more sophisticated models (e.g., complex neural network models) may help improve prediction performance but at the expense of interpretability. Thirdly, this study only included adults from China and the US. Future research should include data from other populations to assess the generalizability of our findings to different cultural contexts.

Conclusion

Pooled evidence from two nationally representative studies, encompassing a total of 46,057 adults, suggests that recognizing demographic differences and emphasizing the relative importance of sedentary behavior, alcohol consumption, and protein intake are beneficial for managing body weight status in adults. Our findings provide actionable insights for developing targeted interventions that focus on reducing sedentary behavior and alcohol consumption to manage overweight and obesity rates in

diverse populations. In addition, the specific risk thresholds for lifestyle factors observed in this study can help inform and guide future research and public health actions. Future studies should explore integrating other lifestyle factors, and validate the findings in populations outside China and the US to confirm the generalizability of our observations.

Abbreviations

CHNS	China Health and Nutrition Survey
NHANES	National Health and Nutrition Examination Survey
BMI	Body Mass Index
DT	Decision Tree
RF	Random Forest
GBDT	Gradient Boosting Decision Tree
SHAP	SHapley Additive explanation
METs	Metabolic Equivalent Tasks
AUC	Area Under the ROC Curve

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12889-024-20510-z.

Supplementary Material 1

Supplementary Material 2

Acknowledgements
Thanks to all authors for their contributions in writing and editing of the manuscript.

Author contributions
ZYS participated in the design of the study, collection and analysis of data, and drafting and editing of the manuscript; YHY, VF, FH, ZWX, XX, YFS, YHY, KQ, and YFL contributed to the drafting and editing of the manuscript; ZYQ and DCX participated in data collection and analysis; AGC and LYZ conceived of the study, and participated in its design and coordination and helped to draft and edit the manuscript. All authors contributed to the manuscript writing and editing.

Funding
This research was supported by grants from the National Social Science Foundation of China (23ATY008) and the Fok Ying Tong Education Foundation (141113).

Data availability
The public datasets used in this study are freely available at the following links:https://www.cpc.unc.edu/projects/china and https://www.cdc.gov/nchs/nhanes/index.htm.

Declarations

Ethics approval and consent to participate
The participant data for this study are open source and available in two public datasets. CHNS was approved by the University of North Carolina at Chapel Hill and the Chinese Center for Disease Control and Prevention. NHANES was approved by the National Center for Health Statistics Research Ethics Review Board.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]College of Physical Education, Yangzhou University, Yangzhou 225127, China
[2]School of Sport and Brain Health, Nanjing Sport Institute, Nanjing 210014, China
[3]School of Information Engineering, Yangzhou University, Yangzhou 225127, China
[4]Institute for Sport and Sport Science, TU Dortmund University, 44227 Dortmund, Germany
[5]Research Group Degenerative and Chronic Diseases, Movement, Faculty of Health Sciences Brandenburg, University of Potsdam, 14476 Potsdam, Germany
[6]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[7]Department of Physical Education, Nanjing University, Nanjing 210033, China
[8]Department of Sport, Gdansk University of Physical Education and Sport, Gdansk 80–336, Poland
[9]Body-Brain-Mind Laboratory, School of Psychology, Shenzhen University, Shenzhen 518060, China
[10]Nanjing Sport Institute, Nanjing 210014, China

Received: 15 August 2024 / Accepted: 24 October 2024
Published online: 01 November 2024

References

1. Apovian CM, Obesity: definition, comorbidities, causes, and Burden. Am J Managed Care. 2016;22:s176–85.
2. World Health Organization. Obesity and its root causes. https://www.who.int/india/Campaigns/world-obesity-day. Accessed 05 Mar 2024.
3. Ng M, Fleming T, Robinson M, Thomson B, Graetz N, Margono C, et al. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet. 2014;384:766–81.
4. Jaacks LM, Vandevijvere S, Pan A, McGowan CJ, Wallace C, Imamura F, et al. The obesity transition: stages of the global epidemic. The lancet. Diabetes Endocrinol. 2019;7:231–40.
5. Teufel F, Seiglie JA, Geldsetzer P, Theilmann M, Marcus ME, Ebert C, et al. Body-mass index and diabetes risk in 57 low-income and middle-income countries: a cross-sectional study of nationally representative, individual-level data in 685 616 adults. Lancet. 2021;398:238–48.
6. World Health Organization. Obesity and overweight. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight. Accessed 03 May 2024.
7. Flegal KM, Kit BK, Orpana H, Graubard BI. Association of all-cause mortality with overweight and obesity using standard body mass index categories: a systematic review and meta-analysis. JAMA. 2013;309:71–82.
8. O'Neil CE, Deshmukh-Taskar P, Mendoza JA, Nicklas TA, Liu Y, Relyea G, et al. Dietary, Lifestyle, and Health Correlates of Overweight and Obesity in Adults 19 to 39 Years of Age. Am J Lifestyle Med. 2011;6:347–58.
9. Naja F, Itani L, Nasrallah MP, Chami H, Tamim H, Nasreddine L. A healthy lifestyle pattern is associated with a metabolically healthy phenotype in overweight and obese adults: a cross-sectional study. Eur J Nutr. 2019;59:2145–58.
10. Ju Q, Wu X, Li B, Peng H, Lippke S, Gan Y. Regulation of craving training to support healthy food choices under stress: A randomized control trial employing the hierarchical drift-diffusion model. Appl Psychol Health Well Being. 2024;16:1159–77.
11. Lee KX, Quek KF, Ramadas A. Dietary and Lifestyle Risk Factors of Obesity Among Young Adults: A Scoping Review of Observational Studies. Curr Nutr Rep. 2023;12:733–43.
12. Ekström S, Andersson N, Kull I, Georgelis A, Ljungman PLS, Melén E, et al. Changes in lifestyle, adiposity, and cardiometabolic markers among young adults in Sweden during the COVID-19 pandemic. BMC Public Health. 2023;23:1026.
13. He K, Du S, Xun P, Sharma S, Wang H, Zhai F, et al. Consumption of monosodium glutamate in relation to incidence of overweight in Chinese adults: China Health and Nutrition Survey (CHNS). Am J Clin Nutr. 2011;93:1328–36.
14. Hunt KJ, St. Peter JV, Malek AM, Vrana-Diaz C, Marriott BP, Greenberg D. Daily Eating Frequency in US Adults: Associations with Low-Calorie Sweeteners, Body Mass Index, and Nutrient Intake (NHANES 2007–2016). Nutrients. 2020;12:2566.

15. Su C, Jia XF, Wang ZH, Wang HJ, Ouyang YF, Zhang B. Longitudinal association of leisure time physical activity and sedentary behaviors with body weight among Chinese adults from China Health and Nutrition Survey 2004–2011. Eur J Clin Nutr. 2017;71:383–8.
16. Zhang Y, Yang J, Hou W, Arcan C. Obesity Trends and Associations with Types of Physical Activity and Sedentary Behavior in US Adults: National Health and Nutrition Examination Survey, 2007-2016. Obesity. 2020;29:240–50.
17. Chang T, Ravi N, Plegue MA, Sonneville KR, Davis MM. Inadequate Hydration, BMI, and Obesity Among US Adults: NHANES 2009–2012. Ann Fam Med. 2016;14:320–4.
18. Yuan X, Wei Y, Jiang H, Wang H, Wang Z, Dong M, et al. Longitudinal Relationship between the Percentage of Energy Intake from Macronutrients and Overweight/Obesity among Chinese Adults from 1991 to 2018. Nutrients. 2024;16:666.
19. Zhang X, Zhang J, Du W, Su C, Ouyang Y, Huang F, et al. Multi-Trajectories of Macronutrient Intake and Their Associations with Obesity among Chinese Adults from 1991 to 2018: A Prospective Study. Nutrients. 2021;14:13.
20. Yuan F, Wu M, Li W, Zhang H. The effect of self-perceived stress, the history of smoking and drinking on weight status in Chinese adults - evidence from the 2015 China Health and Nutrition Survey. Medicine. 2020;99:e21159.
21. Ellison-Barnes A, Yeh H-C, Pollack CE, Daumit GL, Chander G, Galiatsatos P, et al. Weighing cessation: Rising adiposity of current smokers in NHANES. Prev Med. 2023;175:107713.
22. Sun T, Lv J, Zhao X, Li W, Zhang Z, Nie L. In vivo liver function reserve assessments in alcoholic liver disease by scalable photoacoustic imaging. Photoacoustics. 2023;34:100569.
23. Liu Y, Yin H, Liu X, Zhang L, Wu D, Shi Y, et al. Alcohol use disorder and time perception: The mediating role of attention and working memory. Addict Biol. 2024;29:e13367.
24. White GE, Mair C, Richardson GA, Courcoulas AP, King WC. Alcohol Use Among U.S. Adults by Weight Status and Weight Loss Attempt: NHANES, 2011–2016. Am J Prev Med. 2019;57:220–30.
25. Fu Y, Gou W, Hu W, Mao Y, Tian Y, Liang X, et al. Integration of an interpretable machine learning algorithm to identify early life risk factors of childhood obesity among preterm infants: a prospective birth cohort. BMC Med. 2020;18:184.
26. Allen B. An interpretable machine learning model of cross-sectional U.S. county-level obesity prevalence using explainable artificial intelligence. PLoS ONE. 2023;18:e0292341.
27. An R, Shen J, Wang J, Yang Y. A scoping review of methodologies for applying artificial intelligence to physical activity interventions. J Sport Health Sci. 2023;00:1–14.
28. Sun Z, Yuan Y, Dong X, Liu Z, Cai K, Cheng W, et al. Supervised Machine Learning: A New Method to Predict the Outcomes following Exercise Intervention in Children with Autism Spectrum Disorder. Int J Clin Health Psychol. 2023;23:100409.
29. Sun Z, Yuan Y, Xiong X, Meng S, Shi Y, Chen A. Predicting academic achievement from the collaborative influences of executive function, physical fitness, and demographic factors among primary school students in China: ensemble learning methods. BMC Public Health. 2024;24.
30. Farrahi V, Rostami M. Machine learning in physical activity, sedentary, and sleep behavior research. J Activity Sedentary Sleep Behav. 2024;3.
31. Farrahi V, Niemela M, Karmeniemi M, Puhakka S, Kangas M, Korpelainen R, et al. Correlates of physical activity behavior in adults: a data mining approach. Int J Behav Nutr Phys Act. 2020;17:94.
32. Ferreras A, Sumalla-Cano S, Martínez-Licort R, Elío I, Tutusaus K, Prola T, et al. Systematic Review of Machine Learning applied to the Prediction of Obesity and Overweight. J Med Syst. 2023;47:8.
33. Colmenarejo G. Machine Learning Models to Predict Childhood and Adolescent Obesity: A Review. Nutrients. 2020;12:2466.
34. McCoy LG, Brenna CTA, Chen SS, Vold K, Das S. Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based. J Clin Epidemiol. 2022;142:252–7.
35. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf. Accessed 18 Feb 2024.
36. Štrumbelj E, Kononenko I. An Efficient Explanation of Individual Classifications using Game Theory. J Mach Learn Res. 2010;11:1–18.
37. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci. 2019;116:22071–80.
38. Lin W, Shi S, Huang H, Wen J, Chen G. Predicting risk of obesity in overweight adults using interpretable machine learning algorithms. Front Endocrinol. 2023;14:1292167.
39. Jeong S, Yun SB, Park SY, Mun S. Understanding cross-data dynamics of individual and social/environmental factors through a public health lens: explainable machine learning approaches. Front Public Health. 2023;11:1257861.
40. Wang Y, Pan L, He H, Li Z, Cui S, Yang A, et al. Prevalence, associated factors, and gene polymorphisms of obesity in Tibetan adults in Qinghai, China. BMC Public Health. 2024;24:305.
41. Dettoni R, Bahamondes C, Yevenes C, Cespedes C, Espinosa J. The effect of obesity on chronic diseases in USA: a flexible copula approach. Sci Rep. 2023;13:1831.
42. Ainsworth BE, Haskell WL, Whitt MC, Irwin ML, Swartz AM, Strath SJ, et al. Compendium of physical activities: an update of activity codes and MET intensities. Med Sci Sports Exerc. 2000;32:S498–504.
43. Du H, Bennett D, Li L, Whitlock G, Guo Y, Collins R, et al. Physical activity and sedentary leisure time and their associations with BMI, waist circumference, and percentage body fat in 0.5 million adults: the China Kadoorie Biobank study. Am J Clin Nutr. 2013;97:487–96.
44. Tremblay MS, Aubert S, Barnes JD, Saunders TJ, Carson V, Latimer-Cheung AE, et al. Sedentary Behavior Research Network (SBRN) - Terminology Consensus Project process and outcome. Int J Behav Nutr Phys Act. 2017;14:75.
45. Huang L, Wang L, Jiang H, Wang H, Wang Z, Zhang B, et al. Trends in Dietary Carbohydrates, Protein, and Fat Intake and Diet Quality Among Chinese Adults, 1991–2015: Results from the China Health and Nutrition Survey. Public Health Nutr. 2022;26:1–31.
46. Cheng Z, Fu F, Lian Y, Zhan Z, Zhang W. Low-carbohydrate-diet score, dietary macronutrient intake, and depression among adults in the United States. J Affect Disord. 2024;352:125–32.
47. Zhang MJ, Zhang MZ, Yuan S, Yang HG, Lu GL, Chen R, et al. A nutrient-wide association study for the risk of cardiovascular disease in the China Health and Nutrition Survey (CHNS) and the National Health and Nutrition Examination Survey (NHANES). Food Funct. 2023;14:8597–603.
48. Risk Factor Collaboration N. C. D. Worldwide trends in underweight and obesity from 1990 to 2022: a pooled analysis of 3663 population-representative studies with 222 million children, adolescents, and adults. Lancet. 2024;403:1027–50.
49. He H, Garcia EA. Learning from Imbalanced Data. IEEE Trans Knowl Data Eng. 2009;21:1263–84.
50. Rodriguez-Pardo C, Segura A, Zamorano-Leon JJ, Martinez-Santos C, Martinez D, Collado-Yurrita L, et al. Decision tree learning to predict overweight/obesity based on body mass index and gene polymporphisms. Gene. 2019;699:88–93.
51. Rehkopf DH, Laraia BA, Segal M, Braithwaite D, Epel E. The relative importance of predictors of body mass index change, overweight and obesity in adolescent girls. Int J Pediatr obesity: IJPO: official J Int Association Study Obes. 2011;6:e233–42.
52. Hammond R, Athanasiadou R, Curado S, Aphinyanaphongs Y, Abrams C, Messito MJ, et al. Predicting childhood obesity using electronic health records and publicly available data. PLoS ONE. 2019;14:e0215571.
53. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. Nat Mach Intell. 2020;2:56–67.
54. Quinlan JR. Induction of Decision Trees. Mach Learn. 1986;1:81–106.
55. Breiman L. Random Forests. Mach Learn. 2001;45:5–32.
56. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001;29:1189–232.
57. Dugan TM, Mukhopadhyay S, Carroll A, Downs S. Machine Learning Techniques for Prediction of Early Childhood Obesity. Appl Clin Inf. 2015;6:506–20.
58. Giudici P, Raffinetti E. Shapley-Lorenz eXplainable Artificial Intelligence. Expert Syst Appl. 2021;167:114104.
59. Swinburn BA, Sacks G, Hall KD, McPherson K, Finegood DT, Moodie ML, et al. The global obesity pandemic: shaped by global drivers and local environments. Lancet. 2011;378:804–14.
60. Bell AC, Ge K, Popkin BM. The Road to Obesity or the Path to Prevention: Motorized Transportation and Obesity in China. Obes Res. 2012;10:277–83.
61. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518:197–206.
62. Wang Y, Beydoun MA. The obesity epidemic in the United States–gender, age, socioeconomic, racial/ethnic, and geographic characteristics: a systematic review and meta-regression analysis. Epidemiol Rev. 2007;29:6–28.

63. Shi XD, He SM, Tao YC, Wang CY, Jiang YF, Feng XW, et al. Prevalence of obesity and associated risk factors in Northeastern China. Diabetes Res Clin Pract. 2011;91:389–94.
64. Peeters A, Barendregt JJ, Willekens F, Mackenbach JP, Al Mamun A, Bonneux L. Obesity in adulthood and its consequences for life expectancy: a life-table analysis. Ann Intern Med. 2003;138:24–32.
65. Zhang Y, Jiao Y, Lu K, He G. Influencing factors of overweight and obesity among Chinese adults: a meta-analysis. Chin J Public Health. 2015;31:232–5.
66. Thorp AA, Owen N, Neuhaus M, Dunstan DW. Sedentary behaviors and subsequent health outcomes in adults a systematic review of longitudinal studies, 1996–2011. Am J Prev Med. 2011;41:207–15.
67. Hamilton MT, Hamilton DG, Zderic TW. Role of low energy expenditure and sitting in obesity, metabolic syndrome, type 2 diabetes, and cardiovascular disease. Diabetes. 2007;56:2655–67.
68. Popa A, Fratila O, Rus M, Aron R, Vesa C, Pantis C, et al. Risk factors for adiposity in the urban population and influence on the prevalence of overweight and obesity. Experimental Therapeutic Med. 2020;20:129–33.
69. Traversy G, Chaput JP. Alcohol Consumption and Obesity: An Update. Curr Obes Rep. 2015;4:122–30.
70. Yeomans MR, Caton S, Hetherington MM. Alcohol and food intake. Curr Opin Clin Nutr Metab Care. 2003;6:639–44.
71. Audrain-McGovern J, Benowitz NL. Cigarette smoking, nicotine, and body weight. Clin Pharmacol Ther. 2011;90:164–8.
72. World Health Organization. Tobacco. https://www.who.int/news-room/fact-sheets/detail/tobacco. Accessed 02 Jun 2024.
73. Noakes M, Keogh JB, Foster PR, Clifton PM. Effect of an energy-restricted, high-protein, low-fat diet relative to a conventional high-carbohydrate, low-fat diet on weight loss, body composition, nutritional status, and markers of cardiovascular health in obese women. Am J Clin Nutr. 2005;81:1298–306.
74. Hansen TT, Astrup A, Sjodin A. Are Dietary Proteins the Key to Successful Body Weight Management? A Systematic Review and Meta-Analysis of Studies Assessing Body Weight Outcomes after Interventions with Increased Dietary Protein. Nutrients. 2021;13.
75. Brandhagen M, Forslund HB, Lissner L, Winkvist A, Lindroos AK, Carlsson LM, et al. Alcohol and macronutrient intake patterns are related to general and central adiposity. Eur J Clin Nutr. 2012;66:305–13.
76. Tian Y, Jiang G, Chang G, Yang Y, Li Z, Gao J, et al. Analysis of key factors in diet of residents with overweight/obesity control in Tianjin. Occupation Health. 2009;25:2662–5.
77. Lin Y, Bolca S, Vandevijvere S, De Vriese S, Mouratidou T, De Neve M, et al. Plant and animal protein intake and its association with overweight and obesity among the Belgian population. Br J Nutr. 2011;105:1106–16.
78. Lin W, Shi S, Lan H, Wang N, Huang H, Wen J, et al. Identification of influence factors in overweight population through an interpretable risk model based on machine learning: a large retrospective cohort. Endocrine. 2024;83:604–14.
79. Cheng X, Lin SY, Liu J, Liu S, Zhang J, Nie P et al. Does Physical Activity Predict Obesity-A Machine Learning and Statistical Method-Based Analysis. Int J Environ Res Public Health. 2021;18.
80. Babaei G, Giudici P, Raffinetti E. A Rank Graduation Box for SAFE AI. Expert Syst Appl. 2025;259:125239.
81. National Health Commission of the People's Republic of China. Dietary Guidelines for Chinese Adult. http://www.nhc.gov.cn/. Accessed 04 Jun 2024.
82. Centers for Disease Control and Prevention. Obesity Strategies: What Can Be Done. https://www.cdc.gov/obesity/php/about/obesity-strategies-what-can-be-done.html. Accessed 04 Jun 2024.
83. Benfica M, Silva K, Oliveira M, Ribeiro T, Messa R, Tamura E, et al. Sleep duration on overweight and obesity: an overview of systematic reviews. Sleep Med. 2024;115:S63–4.

Publisher's note