Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 44 No. 9
September 2023

Analyzing the Significant Factors Influencing Obesity of Adolescent Girls through Statistical Methods

Maria Vinitha A[1]*, T. Pramananda Perumal[2]
[1]Presidency College, Chennai-600 005, Tamil Nadu, India.
[2]Presidency College, Chennai-600 005, Tamil Nadu, India.

Abstract
Objectives: Obesity is an important health issue. It is strongly correlated with numerous risk factors, causing a significant impact on increasing chronic diseases and mortality rates in the current global scale. These risk factors of obesity encompass several challenges. The main objective of our study is to collect primary data of adolescent girls in eight categories to identify a set of risk factors of obesity. **Methods:** In this study, the primary data can be collected with a specific focus on anthropometric measures, dietary habits, food frequency, physical activity pattern, non-communicable diseases, health issues, psychosocial behavior and health survey which helps to investigate the relationship between various factors in predicting health outcomes. Statistical methods are used to assess the whole dataset and to investigate the multi-collinearity among the collected dataset by applying Chi-square test, Sampling Adequacy test, Exploratory Factor Analysis and Confirmatory Factor Analysis. **Findings:** The relationship between the attributes /variable and BMI are identified by using Chi-square test. From the dataset, insignificant and irrelevant attributes are removed. Then, the eight factors are transformed into thirteen factors using Exploratory Factor Analysis and these 13 factors are validated using Confirmatory Factor Analysis. These 13 factors are considered as evidences for the prevalence of overweight and obese. **Novelty:** The constant focus on the obesity factors of the adolescent age group will further reduce or prevent the obesity occurrences and to improve the healthy lifestyle for the future society.

Keywords: *Adolescent girls, Overweight, Obesity, Risk factors, Exploratory Factor Analysis, Confirmatory Factor Analysis.*

1. Introduction

Obesity is an important health problem that is strongly linked to many chronic illnesses, with harmful impacts and long-term consequences not only for patients but also for their families. Problems with nutrition or malnutrition are a double burden in Southeast Asia because the number of cases of malnutrition and malnourishment is still very high and the number of cases of obesity has also increased significantly over time [1]. Obesity prevalence in Indonesia has increased over the years, according to data, gathered. Obesity prevalence among adult men has reached 13.9% in 2007, 7.8% in 2010, and 19.7% in 2013, whereas that among adult women has reached 14.8% in 2007, 15.5% in 2010 and 32.9% in 2013. According to the same survey, the prevalence of obesity in men and women has dropped marginally to 14.5% and 29.3% respectively by 2018 [1]. Furthermore, compared to developed regions, emerging regions are expected to have a significantly larger proportional increase in the number of overweight and obese people between 2005 and 2030. Population growth, age, urbanization and lifestyle changes, such as increasing total calorie intake and reduced physical activity, together contribute to an epidemic of overweight and obesity in emerging countries [2]. Stress factors related to food choices have demonstrated a new trend being explored. Precision health, or the use of genetic information, could reveal how eating choices affect adolescent obesity[3]. Increased habituation to less active lifestyle patterns among adolescents too causes the prevalence of overweight and obesity. The intervention of the internet and modern urbanization that influence diverse lifestyle variations, overweight and obesity among adolescents. Increased time of playing sports and outdoor activities instead of spending time in indoor activities like watching television, playing videogames and operating computers or mobiles

reduces the prevalence of overweight and obese. Various investigations too determine that lack of physical activity has statistical significance with overweight and obesity [4, 5].

Factor analysis is especially useful for reducing a large number of linked variables to a manageable amount before employing them in other analyses such as multiple regression or multivariate analysis of variance. Kaiser's Criterion and the Scree test are used to determine the number of components to be extracted. To reduce the number of variables with large loadings on each factor, the Varimax orthogonal factor rotation approach is used. Convergent validity is proved when the extracted average variance is larger than or equal to 0.5. The results show that factor analysis not only detects irrelevant items but also allows for the extraction of valuable factors from the data set [6]. Confirmatory Factor Analysis (CFA) and Exploratory Factor Analysis (EFA) are two methodologies frequently used in scale development and scale adaption research. If the relationship between the items is unknown, EFA should be utilized; however, if the relationship has been tested and the factors and associated items are known, CFA should be used [7].

The investigation has been performed to find and evaluate the risk factors that cause the incidence of overweight and obesity among adolescent girls. Proper data on adolescent dietary patterns, habits and nutritional conditions if derived, could assist in planning the strategies that are needed to strengthen adolescent well-being and health. The purpose of this study is to identify the key factors and its relevant variables from numerous risk indicators influencing obesity among adolescent girls based on the hidden factors in Anthropometric profile, Dietary pattern, Food frequency, Physical activity, Family history of Non-Communicable Diseases (NCD), Self-health issues, Psychosocial behavior and Health survey.

This paper is organized as follows. In the section 2, we describe the previous works, related to our study. In section 3, the details of the system (data) to be studied and tools for study are explained. In section 4, we present our results and discussion on a) Chi-Square test between attributes and BMI b) Sampling Adequacy test c) Exploratory Factor Analysis d) Confirmatory Factor Analysis. In section 5, we highlight the principal outcome of the study.

2. Related works

There are several papers which deals with the risk and prediction of obesity. Their works have been studied and attempted to comprehend the method that they demonstrated. Obesity and overweight continue to be major human health issues in India however too in other nations for the past two decades. This problem is less found in early childhood and aged women however is rising among adolescents swiftly and records higher among women who have got children, particularly among girls or women with poor education. In developed nations, it happens largely in at-risk groups of economically weaker sections and the reverse happens in developing countries as the changes in home cooking and underweight are more related. Obesity has a tendency to increase with larger income groups. In developed nations, nearly twice the women group will experience obesity in comparison with men in the lower socioeconomic sections [8]. Researches have been performed by various groups on obesity and overweight associated risk analysis to determine causal factors.

Kabbaoui et al. [9] have investigated the proportional differences between normal weight and overweight/obese groups using the Chi-Squared test. $P < 0.05$ is considered statistically significant. In order to evaluate the relationship between the factors of interest and overweight, including obesity, they have applied the Logistic Regression. The relationship between the factors considered and overweight or obesity have been established by applying uni-variate analysis. In order to evaluate the degree of relationship, they have set the analysis with a basic Odds Ratio (OR) value and with confidence intervals (CIs) of 95%. They have performed a multivariate analysis by using Logistic Regression in order to find associations between various factors. Factors relations by including $\alpha < 20\%$ to uni-variate analysis have applied in the initial model of multivariate logistic model with extra-factors, stated to be more related to overweight and obesity in earlier research. The significance level in multivariate analyses has been set as $P < 0.05$.

They have conducted a survey of 1561 participants with boys 782 and girls 779 by collecting details of anthropometric measurements like age, height and weight, socioeconomic factors like average family income, parent education, duration of viewing computers and television, duration of playing sports inside and outside of school and transporting mode like walking, a motorcycle to school and dietary behavior like frequency of consuming breakfast, lunch and dinner, frequency of consuming fruits, vegetables, soft drinks and snacks like cakes, biscuits, dairy products, sweets and chocolates.

Chandini et al.[10] have studied the risk factors connected to overweight and obesity in females of reproductive age from 15 to 49 years in city regions of Rajbiraj Municipality, Saptari. They have calculated the sample size by applying a single proportion method, presuming the jointed prevalence rate of overweight and obesity to be 22%, CI of 95%, margin error of 7% and non-response rate of 10%.They have applied Chi-square to determine the factor association. They have considered the BMI, waist circumference and waist-to-hip ratio as dependent variables in their study. They have considered socio-economic and demographic factors like age, marital status, education, occupation, family income and family size, duration, frequency and intensity of physical activities while at work, travel and leisure in a normal week period according to the International Physical Activity Questionnaire (IPAQ), frequency of diet consumption like carbohydrates, fat, calorie, etc. according to WHO standards, health-related characteristics like menstrual disorders, use of contraceptive uses, and lastly the behavioral aspects like watching television during eating, sleep, consuming food outside, smoking and alcohol consumption.

Viswambharan et al.[11] have performed a study with 1011 adolescents aged between 15 and 18 of which 526 from government schools, 389 from aided schools and 96 from private schools in Kozhikode Corporation, Kerala, South India. The study group have comprised 475 girls and 536 boys. They have used a questionnaire which contains questions, associated with socio-demographic factors, diet, physical activity and other known risk factors of obesity in adolescents. They have used socio-demographic factors like age, gender, religion, studying standards, birthplace, the status of immunization, family type, parents' education and occupation, family income, dietary behaviors like food consumption habits, consumption frequency, dietary foods that cause overweight and obesity, the nature, duration and frequency of physical activities, duration of viewing television and computer, and sleeping duration. They have asserted the study with OR and with a CI of 95% confidence interval to assess the risk variables by applying uni-variate analysis.

Hazzaa et al.[12] have performed a cross-sectional study among school adolescent children that has been carried out in three major cities Al-Khobar, Jeddah and Riyadh in Saudi Arabia. They have chosen participants of 2906 secondary school children with 1400 boys and 1506 girls aged between 14 and 19 years through a random multistage layered cluster sampling method. They have used the Arab Teens Lifestyle Study (ATLS) research instrument to estimate lifestyle details comprised of 47 items. The initial 5 items are age, height, weight, waist circumference and education. The items between 6and 34 included the physical activities questions, the items between 35 and 37 have measured the sedentary lifestyles and the items between 38 and 47 have concentrated on dietary patterns. They have used the anthropometric measurements data comprising weight, height, waist circumference, waist/height ratio, behaviors comprised television and computer viewing, and playing video games, physical activities inquired employing a standard questionnaire, and dietary habits of consumption frequency per week. They have applied logistic regression to investigate the relationship between lifestyle variables and obesity.

Rocha et al [13] have focused to investigate the association between factors like environment, socioeconomic, and nutrition and childhood overweight and obesity. They have performed a population-based study of children aged between 2 and 6 years in Ceará, Brazil. The nutritional condition of children has been evaluated by z scores of BMI for grouping into overweight and obesity. They have used the model of ordinal logistic regression to determine the association

between the variables and, overweight and obesity. Their survey has involved1038 males and 1021 females among 2059 children with an average age of 46 ± 17 months children, having an overweight prevalence of 12% and obesity prevalence of 8% with 95% CI respectively. They have found, applying multivariate analysis, the likelihood of child obesity intensified for increased family income. Further, families with a lesser number of children have more than 30% lesser number of overweight children.

Pratap Singh et al. [14] have examined the incidence of overweight and obesity in adolescents aged 10 to 14, i.e., classes 5 to 9, from schools in various areas of Chandigarh. They have selected nine coeducational schools in order to contain participants from diverse socioeconomic levels. They have included 1030 students, of which 502 and 528 students are from the government and private schools respectively. On the whole, the prevalence of overweight and obesity, assessed by utilizing age-related BMI limits are indicated as 9.9 % and 14.0 % correspondingly. The incidence of overweight with a BMI of 23 is seemed to be 10.3% and 9.4 % correspondingly in boys and girls and in the case of obesity with a BMI of 27 has been set as 13.3% and 14.7% corresponding in boys and girls. The statistical significance for associations has been determined by applying a uni-variate analysis between the obesity risk with sex, socio-economic level and physical activity.

Anuradha et al. [15] have assessed the incidence of overweight and obesity and their relationship with environmental and social factors in adolescents from schools in Tirupati town of Andhra Pradesh, India. They have collected data by administering the interview method to children aged between 12 and16 years. They have sampled the data that comprised 2258 subjects with boys of 1097 and girls of 1161. They have ascertained the overweight and obesity through BMI on the basis of recommendations by the Centre for Disease Control and Prevention 2000. They have also collected data on environmental and social factors by applying experimented and approved questionnaires.

Duante et al.[16] have studied to focus on the research gap of the prevalence of overweight and obesity in adults that has exhibited a gradual increase. They have included 9076 adults over 20years and collected socio-economic, socio-demographic anthropometric, clinical and health and dietary data. In controlling the consequences of the factors, the variables are substantially related to overweight and obesity among adults comprised energy consumption, residence type, age, social status, wealth, education and smoking habit. The odds of being overweight and obese have been set to 29% high for the individuals who have consumed recommended energy in comparison with those who did not, and 28% high for the individuals living in city regions than village regions. The odds of overweight and obesity have increased as socio-economic conditions have increased and the population have matured. Adults with life partners are highly possible to be overweight and obese on comparing single. As compared to the adults with no education, the odds of being overweight and obese are larger among adults who have gone to high school and vocational graduates and twofold high among college graduates. This analysis further has contributed solid proof of the variables related to overweight and obesity, which has been focused on multiple section method by designing the most efficient programs.

Niranjan et al. [17] have investigated the prevalence of overweight and obesity among 984 affluent students who had an age between 10 and 16 years from three schools in Raichur city. They have measured height and weight, and collected the data like diet types, junk food consumption, sleeping hours, exercise routine, viewing television and information about the detrimental effects due to obesity through questionnaires. They have determined the prevalence of obesity in girls is 6% in comparison with boys of 3.6% and it has been observed in the ages of 13 and 15 years. Further, in their analysis, on studying the factors, junk food intake ($P = 0.001$), exercise routine ($P = 0.004$), and sleeping duration over 8 hours ($P = 0.007$) have a substantial impact on obesity. Adolescents who are not performing the routine exercise, have huge junk food consumption, and slept over 8 hours is possible to be vulnerable to obesity. A mixed diet have not shown any significant ($P = 0.233$) association with obesity.

Song et al. [18] have performed their study on 14618 adult participants which include 7799 males and 6819 females ages over 35 years from the Cardiovascular Risk Survey, carried out in 2010. They have obtained data from interviews and physical examination methods. They have used these samples to assess the prevalence of overweight with a BMI between 24 and 28, and obesity with BMI over 28 in Xinjiang Province. The influencing variables are studied by applying Chi-Square statistical techniques. They have found that the overall prevalence of overweight is 36.5% with 40% in males and 33.4% in females and the prevalence of obesity is 26.5% with 27.2% in males and 25.8% in females. They have finally concluded that females have a higher prevalence of overweight and obesity than males ($p<0.05$).

Liu et al. [19] have focused to assess the prevalence and controlling factors of general and abdominal obesity and overweight in rural areas of China. They have selected 39034 participants, aged between 18 and 79 years from the Henan Rural Cohort Study for their cross sectional study. They have showed the prevalence of overweight, general and abdominal obesities due to normalized ages are 34.97%, 16.82%, and 43.71% respectively in the adults of the common Chinese population. The prevalence of overweight, general and abdominal obesities due to gender differences is 36.04%, 18.98% and 35.37% in the case of men and 34.55%, 15.42%, and 49.13% in the case of females correspondingly. The subpopulation study has demonstrated overweight, general and abdominal obesity rates occurred with significant differences, however they are generally larger in whole subpopulations. Moreover, their analysis has determined a U-shaped statistical significance in relations between the prevalence of overweight, general and abdominal obesity and age classes. Further, the prevalence of subjects with both abnormal BMI and waist circumference has approximately 40%. The other controlling factors like age, marital status and cohabit, per capita monthly income and unhealthy living style are independent of overweight, general and abdominal obesities. They have conclusively found that overweight and obesity are intense in a rural areas in China.

Lee and Ham [20] have studied the factors affecting underweight and obesity participants comprising 4,895 children from 59 different primary schools of entire Korea. They have grouped the children based on their weight < 5% BMI for age as underweight, 5%-85% BMI for age as normal weight, and ≥ 85% BMI for age as overweight or obese. They have adopted the questionnaire comprised of demographic features, health conditions, dietary habits and physical exercise of children and environmental attributes of schools. Their results have demonstrated that demographic factors like age and gender, health conditions like atopic dermatitis and ill health, and the diet types like diet behavior and unbalanced diet are more related to the underweight ($p < .05$), whereas the demographic elements like age and gender, health conditions i.e., poor health, diet behavior like fast food intake consumption and diet patterns, and school environment features like rural area are more related to overweight and obesity ($p < .05$).

More than 3 folds percentage of adolescents have been influenced by obesity since 1970 in the United States [21]. A joint prevalence of overweight and obese are found to be 29.8% as stated by a countrywide analysis performed in 2015[22]. An analysis carried out in Egypt too has determined a larger section of the population, affected by overweight and obese, i.e. by 20% and 10.7% respectively [23]. The differences in the prevalence of obesity and overweight might be because of the variation in the reference norms and moreover because of the geographical areas subjected to analysis. Nevertheless, an identical study to the present approach with a lesser population of overweight and obesity is observed in the studies that are performed in Pondicherry [24] and Rajasthan [25].

Khadilkar et al. have determined that the prevalence of overweight had been higher among girl students of adolescents where upon it reduced later, whereas the prevalence of obesity have been higher in those aged above17 years [22]. The age-wise prevalence as stated in diverse studies has indicated the highest prevalence in late adolescents. It is agreement with the results obtained in the present study. Thus overweight and obesity depicted had a rise in late adolescence

and beyond which might be because of variations in puberty and too because of feeding junk food and living life with a lack of activity. As India a nation with various traditions and customs, rather than a nation-wide standard, setting norms that are common for various country parts may contribute to a more precise identification of food intake. Once the socio-demographic risk variables are evaluated, the present investigation has depicted a maximum prevalence in the upper middle class 88 (4.4%) [25, 17].

Singh et al. [26] have performed the study on the prevalence of obesity among Asian Indians of vegetarian dietary patterns. In this transition, whole plant foods (fruit, vegetables, nuts, seeds, unrefined whole grains) are replaced by refined carbohydrates, fast foods/snack foods/processed foods, and fried foods. There is also evidence of a transition to cooking oils with more at herogenic effects. Also this provides the evidence of a nutrition transition increases the rate of NCDs despite the continued high prevalence of vegetarianism. Kotian et al. [4] have determined that the overweight risk is seemed to be two folds high among adolescents belonging to higher socio-economic status. In other words, the adolescents belonging to higher socio-economic status are positively correlated with the incidence of overweight and obesity.

3. Methodology

A systematic approach is followed to conduct this study. This section provides details about the study sites and participants, sample size estimation, data collection and tools for analysis.

3.1 Study sites and participants

The dataset for this study has been randomly collected from the sites viz. seven different Schools and five different Colleges, located in the Chennai city, Tamil Nadu, India, during the period between January and March in the year 2019. The selection of Schools and Colleges has been based on the diverse backgrounds of students and the availability of playgrounds, canteens and academic activities of the Institutions.

The participants for the data collection are adolescent girl students, aged from 16 to 19 years who are studying in the Standards XI and XII in Schools and UG Degree I and II years in Colleges. Consent has been obtained from both the participants and parents before data collection.

3.2 Sample size estimation

This study has planned with a sample size of more than 1537 (but in our study, we have taken the sample size as 2000) in order to have a confidence level of 95% and a margin of error within ±2% by setting population proportion 0.2. The sample size has been estimated by applying the formula in Eq. (1)

$$N = \frac{z^2 \hat{p}(1-\hat{p})}{\in^2} \quad \text{.....(1)}$$

where, n is the sample size, z is the z-score associated with a level of confidence, p is the sample proportion, and $\in$ is the margin of error[27].

3.3 Data Collection

Using a standard questionnaire, data such as Demographic information, Anthropometric measures, Dietary pattern, Dietary behavior, Food Frequency, Physical activity pattern, Family history of Non-Communicable Diseases (NCD), Health issues, Psychosocial behavior, and Health survey details have been recorded. By applying the WHO (World Health Organization) standards, Body mass index (BMI) is calculated using the formula, given by Eq. (2).

$$\mathrm{BMI} = \mathrm{W\,(Kg)} / \mathrm{H^2(m^2)} \quad \text{..... (2)}$$

Where, W is weight (in Kilograms) and H is height (in meters) of an adolescent girl student.

The WHO classifies BMI values into different categories based on the following ranges [28]:

BMI range	Classification
< 18.50	Underweight
18.50 – 24.9	Normal
25.00 – 29.00	Overweight

>30.00	Obesity

3.4 Statistical tools for analysis

The main aim of this study is to find the relationship between various attributes and BMI measure. In order to associate these attributes it is required to explain the statistical methods such as Exploratory Factor Analysis (EFA), Confirmatory Factor Analysis (CFA) and Chi-Square test for independence of attributes. Before applying the EFA, it is mandatory to check the data screening of Kaiser Meyer Olkin test (KMO) and Bartlett's test of Sphericity.

3.4.1 Sampling Adequacy test

Once the data has been ingested, we must determine whether factor analysis is feasible. There are two approaches to determining factorability or sampling adequacy.

a) Kaiser-Meyer-Olkin Test

The KMO test determines whether data is suitable for factor analysis. It determines the adequacy of each observed variable as well as the overall model. KMO calculates the proportion of variance in all observed variables. A lower proportion is better suited for factor analysis. KMO values range from 0 to 1. KMO values less than 0.6 are considered inadequate.

b) Bartlett's Test

This test is used to determine whether a correlation matrix is appropriate for Factor Analysis. A factor analysis should not be used if the test is found to be statistically insignificant. The p-value for this Bartlett's test is 0. It evaluates the claim of all variables are uncorrelated and as a result, it is observed as inappropriate for factor analysis because the observed correlation matrix is an identity matrix.

3.4.2 Exploratory Factor Analysis

The purpose of EFA is to find the hidden factors that may be influencing the observable variables and to comprehend how these variables are related to one another.

a) Factor Extraction: Factor extraction methods are procedures used in factor analysis to detect and extract hidden features from the data that has been collected. These strategies helps in reducing data dimensionality and identifying the underlying structure or patterns among variables. There are various factor extraction methods such as Principal Component Analysis (PCA), Principal Axis Factoring (PAF), Maximum Likelihood Estimation (MLE) and few others.

b) Principal Component Analysis (PCA): PCA is a well-known unsupervised learning technique for reducing data dimensionality. It improves interpretability while also minimizing information loss. It aids in identifying the most significant features in a dataset and simplifies data plotting in 2D and 3D. PCA assists in determining a sequence of linear variable combinations.

c) Factor loading: Factor loadings are numerical coefficients that show how much of an observed variable's variance is explained by the related latent factor. The percentage of variance is to compute the squares of factor loadings.

d) Factor Rotation: A rotation approach (eg., Varimax, Obliman, Promax) is frequently used after factor extraction to make the factors easier to comprehend. Rotation helps to match the variables more closely with the identified factors.

3.4.3 Confirmatory Factor Analysis

The CFA is used to confirm or validate the number of factors used to classify each object. The CFA tools namely Average Variance Explained (AVE), Construct Validity such as Composite reliability (CR) and Discriminant validity (DV) are discussed in the following:

a) Average Variance Explained: In construct validation, the average variance extracted (AVE) is a convergent validity indicator. The convergent validity refers to the degree to which multiple measures of the same construct are related to one another.

b) Composite Reliability: A measure of internal consistency reliability called composite reliability measures how consistently a factor is measured across all of its indicators. In more detail, CR consider both the shared and unique variances of the construct's indicators and assesses the ratio of a construct's true score variance to its total score variance.

c) Discriminant Validity: It shows the measures of dimensions that should not be substantially associated to each other are not found to be significantly connected to each other. The total number of discriminant validity coefficients should

be considerably lower than that of convergent validity coefficients.

3.4.4 Chi-Square test

A Chi-Square test is a statistical test to decide whether there is a correlation between the categorical variables. Also it compares observed and expected outcomes. The purpose of this test is to determine whether a difference between actual and predicted data is due to chance or to a relationship between the variables under consideration. As a result, the Chi-Square test is an excellent choice for assisting us in comprehending and interpreting the relationship between our two categorical variables.

The Chi-Square test statistic is expressed in the Eq. (3). [29]

$$\chi^2 = \sum \left(\frac{O_i - E_i}{E_i}\right)^2, i = 1,2,3,\dots,n \quad \dots\dots\dots\dots\dots (3)$$

where, O_i – is the observed frequency

E_i – is the expected frequency

The Chi-Square test degrees of freedom (df) is calculated using the formula shown in Eq. (4)

df = (r-1) (c-1) …….. (4)

where, r – is the No. of rows, c – is the No. of columns

The p-value represents the probability. The following are the distinct values of p that reflect different hypothesis interpretations:

Hypothesis rejected – P<= 0.05

Hypothesis accepted –P>=0.05

3.5 Ethical agreement

A verbally informed agreement is made to enroll the participants by the head of the department of respective schools and colleges followed by oral consensus from the parents of every participant before surveying.

4. Results and Discussion

The above mentioned categories in the previous section 3.3 are used to test the relationships with Body Mass Index (BMI). Further, a Factor Analysis have been performed by applying Python based sklearn decomposition. The decomposed factors are determined from the Eigen values using a scree plot. The analysis and discussion of the Chi-Square test for independence of attributes and factor analysis will be covered in the next Sections.

4.1 Chi-Square test

The relationship between BMI and all 50 attributes are explained by Chi-Square test for Independence of Attributes which helps to find out association between the Attributes and BMI. A P-value less than 0.05 have been statistically significant and a P-value less than 0.01 have set as highly significant is shown in Table 2. Chi-Squareχ^2, degree of freedom (df) and P-value have been evaluated and these are included in Table 1 in 3 extreme columns.

Table 1.Chi-Square test between attributes and BMI

Attributes	Range	χ^2	df	p value
Age (in completed years)	16 to 19	33.64	9	0.0001**
Income (Rs.)	<10000 to ≥200000	69.96	12	0.000**
Family Type	Nuclear, Jointed, Extended	15.19	6	0.0188*
Diet Type	Vegetarian, Vegan, Lacto Veg, Non Veg	10.55	9	0.0007**
Diet Regularity	Regular, Irregular	1.26	6	0.9738
Meal Timing	Regular, Sometime Irregular, Mostly Irregular	14.01	6	0.0295*
Meal Skipping	Yes, No	0.12	3	0.989
Meal Skip	Breakfast, Lunch, Dinner	23.02	12	0.0275*
Meal Skip Frequency	Daily, Weekly Once, Weekly Thrice, Rarely	17.04	9	0.048*
Compensate meals by soft drinks	No, Yes	7.13	3	0.0676*

Appetite	Poor, Fair, Good	10.44	6	0.1072
Physical Activity	Very Poor, Poor, Moderate, Good, Very Good, Intense	20.61	15	0.1495
Presence of self NCD	No NCD, Obesity, BP, Diabetes, CVD	113.43	3	0.000**
Obesity in family members	No, Yes	38.48	3	0.000**
CVD in family members	No, Yes	6.87	3	0.076*
BP in family members	No, Yes	11.33	3	0.010*
DM in family members	No, Yes	17.15	3	0.000**
OP in family members	No, Yes	0.44	3	0.93
Cancer in family members	No, Yes	7.055	3	0.07
Menstrual cycle disorder	No, Yes			
Wheat Cereals	Regular, Fairly, Rarely	13.21	6	0.039
Rice Cereals	Regular, Fairly, Rarely	6.44	6	0.375
Pulses	Regular, Fairly, Rarely	3.33	6	0.765
Yellow Vegetables	Regular, Fairly, Rarely	7.48	6	0.278
Green Leafy vegetables	Regular, Fairly, Rarely	4.95	6	0.55
Fruits	Regular, Fairly, Rarely	2.59	6	0.858
Milk Products	Regular, Fairly, Rarely	10.29	6	0.113
Fish	Regular, Fairly, Rarely	25.68	6	0.000**
Mutton and Chicken	Regular, Fairly, Rarely	6.2	6	0.401*
Oily Food	Regular, Fairly, Rarely	6.83	6	0.336*
Cholesterol	Regular, Fairly, Rarely	30.74	6	0.000**
Animal Fat	Regular, Fairly, Rarely	13.11	6	0.041*
Sweets and Chocolates	Regular, Fairly, Rarely	41.2	6	0.000**
Instant Foods	Regular, Fairly, Rarely	14.07	6	0.029*
Coffee and Tea	Regular, Fairly, Rarely	3.24	6	0.777
Menstural cycle	Regular, Irregular	9.58	3	0.022*
Dysmenorrhea	Yes, No	22.9	3	0.000**
Health Issues-Appetite problems	Yes, No	3.68	3	0.297
Indigestion	Yes, No	3.23	3	0.357
Mouth Alcers	Yes, No	0.24	3	0.971
Anemia	Yes, No	4.59	3	0.204
Food Allergy	Yes, No	2.77	3	0.428
Hypothyroidism	Yes, No	3.78	3	0.286
Acidity	Yes, No	8.91	3	0.03
TV viewing weekstart	Regular, Fairly, Rarely	17.51	9	0.041*
TV viewing weekend	Regular, Fairly, Rarely	7.35	9	0.601
Advertisement Food choice	Yes, No, Neutral	8.71	6	0.19
TV Commercials	Yes, No, Neutral	11.74	6	0.067
Parents Influence	Never, Sometimes, Often, Always	3.05	6	0.802

Friends Influence	Never, Sometimes, Often, Always	2.4	6	0.879

Table 2. Significant attributes and BMI

Attributes	Range	χ^2	df	p value
Age (in completed years)	16 to 19	33.64	9	0.0001**
Income (Rs.)	<10000 to ≥200000	69.96	12	0.000**
Family Type	Nuclear, Jointed, Extended	15.19	6	0.0188*
Diet Type	Vegetarian, Vegan, Lacto Veg, Non Veg	10.55	9	0.0007**
Meal Timing	Regular, Sometime Irregular, Mostly Irregular	14.01	6	0.0295*
Meal Skip	Breakfast, Lunch, Dinner	23.02	12	0.0275*
Meal Skip Frequency	Daily, Weekly Once, Weekly Thrice, Rarely	17.04	9	0.048*
Compensate meals by soft drinks	No, Yes	7.13	3	0.0676*
Presence of self NCD	No NCD, Obesity, BP, Diabetes, CVD	113.43	3	0.000**
Obesity in family members	No, Yes	38.48	3	0.000**
CVD in family members	No, Yes	6.87	3	0.076*
BP in family members	No, Yes	11.33	3	0.010*
DM in family members	No, Yes	17.15	3	0.000**
Fish	Regular, Fairly, Rarely	25.68	6	0.000**
Mutton and Chicken	Regular, Fairly, Rarely	6.2	6	0.401*
Oily Food	Regular, Fairly, Rarely	6.83	6	0.336*
Cholesterol	Regular, Fairly, Rarely	30.74	6	0.000**
Animal Fat	Regular, Fairly, Rarely	13.11	6	0.041*
Sweets and Chocolates	Regular, Fairly, Rarely	41.2	6	0.000**
Instant Foods	Regular, Fairly, Rarely	14.07	6	0.029*
Menstural cycle	Regular, Irregular	9.58	3	0.022*
Dysmenorrhea	Yes, No	22.9	3	0.000**
TV viewing weekstart	Regular, Fairly, Rarely	17.51	9	0.041*

* p<0.05, significant at 5 percent, ** p<0.01, Highly significant at 1 percent

From the above Table 1, in Demographic profile the p value of age and income are less than 0.01 which is highly significant at 1% level whereas family type have significant association at the 5% level. This suggests that BMI is associated with age, income, and family type. In Dietary habits, the p-value for diet type is significant at the 1% level (p<0.01), indicating a strong association with BMI. The p-values for Meal Timing, Meal Skip, Meal Skip Frequency and Compensate meals by soft drinks are all significant at the 5% level (p<0.05), suggesting an association with BMI. Diet Regularity, Meal Skipping, and Appetite on the other hand, have p-values greater than 0.05 and are not associated with BMI.

In Family history, affliction of NCD, family obese and family DM all have significant at 1 % level (p<0.01), indicating an association with BMI. Similarly, family CVD and family BP have the p values are less than 0.05 which are significant at 5% level, whereas family OP, family cancer,

menstrual cycle disorder are insignificant with the prevalence of overweight and obese.

The attribute Physical Activity p value 0.1495 is more than 0.05 which is not significant at 5 % level. It says that Physical activity does not associate with BMI. Among the Food Frequency category, fish, cholesterol, sweet and chocolate have highly significant associations with BMI at the 1% level, while mutton and chicken, oily food, animal fat, and instant foods have significant associations at the 5% level. Wheat cereals, rice cereals, pulses, yellow vegetables, green leaves, fruits, milk products, and coffee and tea do not have significant associations with BMI at the 5% level.

Regarding Health Issues, Menstrual and Dysmenorrhea are significant at 5% and 1% levels respectively, while the other attributes such as Appetite problems, Indigestion, Mouth Ulcers, Anemia, Food Allergy, Hypothyroidism and Acidity are not significant at the 5% level. This suggests that how these attributes are associated to BMI. In terms of Entertainment attributes, only TV viewing week start is significantly associated with BMI ($p<0.05$). However, the other attributes (TV viewing weekend, Ad Food choice, and TV Commercials) are not associated with BMI ($p>0.05$).The attributes Parents and Friends have p values greater than 0.05 and are therefore not significant at the 5% level. This suggests that these attributes are not related to BMI.

We conclude that from Table 2, some attributes are significantly related to BMI, while others are not. This enables to know how BMI is associated with other attributes. This insight is minimal sufficient to explain the impact of BMI. So the researchers thought of reducing the size of the attributes using factor analysis.

4.2 Sampling Adequacy test

The Kaiser-Meyer-Olkin measure of sampling adequacy and the Bartlett's test of sphericity are used to determine if the given data matrix is appropriate for factor analysis or not. The estimated KMO and Bartlett coefficient values, together with probability values, is calculated using Python programming and are shown in the following Table 3.

Table 3. KMO and Bartlett test of Sphericity measures

Factors	KMO test	Bartlett test	P value	Decision
Anthropometric Measures	0.780	23902.562	0.000***	Accepted
Dietary Habits	0.611	393.177	0.000***	Accepted
Food Frequency	0.638	1583.266	0.000***	Accepted
Physical Activity Pattern	0.696	3082.179	0.000***	Accepted
Non-Communicable Diseases	0.569	657.604	0.000***	Accepted
Health Issues	0.500	846.731	0.000***	Accepted
Psychosocial Behavior	0.569	2011.608	0.000***	Accepted
Health Survey	0.832	3986.048	0.000***	Accepted

*** Significant at 1% level ($p<0.01$)

In our study, the KMO test value for all the eight factors is found to be more than that of the threshold value 0.5 which satisfied the requirements according to the authors Kaiser [30] and Field [31]. The sampling adequacy is satisfied from KMO test to proceed further to apply Factor analysis. The Bartlett's test of sphericity probability values of all eight factors are less than 0.01 which is significant at 1 % level. Both KMO and Bartlett test confirm that the given data matrix is suitable for Factor Analysis.

4.3 Factor Loading

The Exploratory Factor Analysis is to identify the sub-factors within defined eight factors. This process identified irrelevant and insignificant attributes are dropped. The significant attributes are taken for further consideration. Finally these eight factors are redeveloped into thirteen factors and validated using Confirmatory Factor Analysis. These thirteen factors are considered as evidences to the prevalence of overweight and obese.

The factor analysis is applied for the data matrix and its factor loading and percentages of variance

are explained, estimated and presented in the following Table 4.

Table 4.Factor Loading and Percentage of variance explained

Factors		**Item description**	**Std. Factor loading**	**% of variance explained**
Anthropometric Measures (ANT)		Percent Body Fat	0.9680	86.68 %
		Body Mass Index	0.9680	
		Weight	0.9498	
		Waist circumference	0.8313	
Dietary Habits (DH)		Meal-timing	0.7405	51.15 %
		Meal-Freq	0.7263	
		Meal-Skip	0.6774	
Food Frequency	Junk Food (JF)	Instant food	0.7615	16.28 %
		Sweet &Chocolates	0.7477	
		Oily-food	0.6086	
	Healthy Food (HF)	Fruits	0.7432	14.96 %
		Milk-Products	0.6622	
		Green Leaf Vegetable	0.6431	
	Non-Vegetarian (NV)	Fish	0.8400	14.03 %
		Mutton &Chicken	0.7893	
	Cereal (CER)	Pulses	0.7756	12.48 %
		Cereals	0.7393	
Physical Activity Pattern (PAP)		PAP- Duration	0.9163	78.12 %
		PAP-Frequency	0.9104	
		PAP–Intensity	0.8216	
Non-communicable diseases	NCD1	Diabetes Mellitus Family	0.8108	34.27 %
		Blood Pressure Family	0.8104	
	NCD2	NCDSELF	0.8553	33.65 %
		Obesity Family	0.7732	
Health Issues (HI)		Appetite Problems	0.8910	79.39 %
		Acidity	0.8910	
Psychosocial Behavior	Physical Influence (PI)	Friends Influence	0.8735	39.24 %
		Parents Influence	0.7878	
	Entertainment Influence (EI)	TV Viewing Weekend	0.6655	25.96 %
		Ad Food Choice	0.8332	
		TV Viewing Weekday	0.7422	
Health Survey (HS)		Energy/Fatigue	0.8180	53.03 %
		Emotional Wellbeing	0.7928	
		General Health	0.7533	
		Social Functioning	0.6921	
		Body Pain	0.6904	
		Role Limitations due To Physical health	0.6022	

In the above table, the loadings of factor analysis in which factor values are above 0.45 have been decided in forming the 13 group factors. Another confirmation of the number of factors is done through Scree plot.

4.4 Principal Component Analysis

The procedure of analysis and finding principal (statistically significant) components /factors is done using a scree plot. A Scree plot is a line plot between Eigen values of factors and factor numbers in the dataset. The Scree plot is used to determine the number of factors to retain. Here the scree plot shows the dataset's Eigen values on the y-axis and the factor numbers on the x-axis. This plot is generated while performing sklearn decomposition factor analysis in Python. When the factors in the dataset are having Eigen value greater than 1, they are considered to be the major factors. In our study, the scree plot of Eigen values greater than 1 reveals that dimensionality of entire 50 factors in the dataset has reduced into 13 grouped factors as shown in Figure 1.

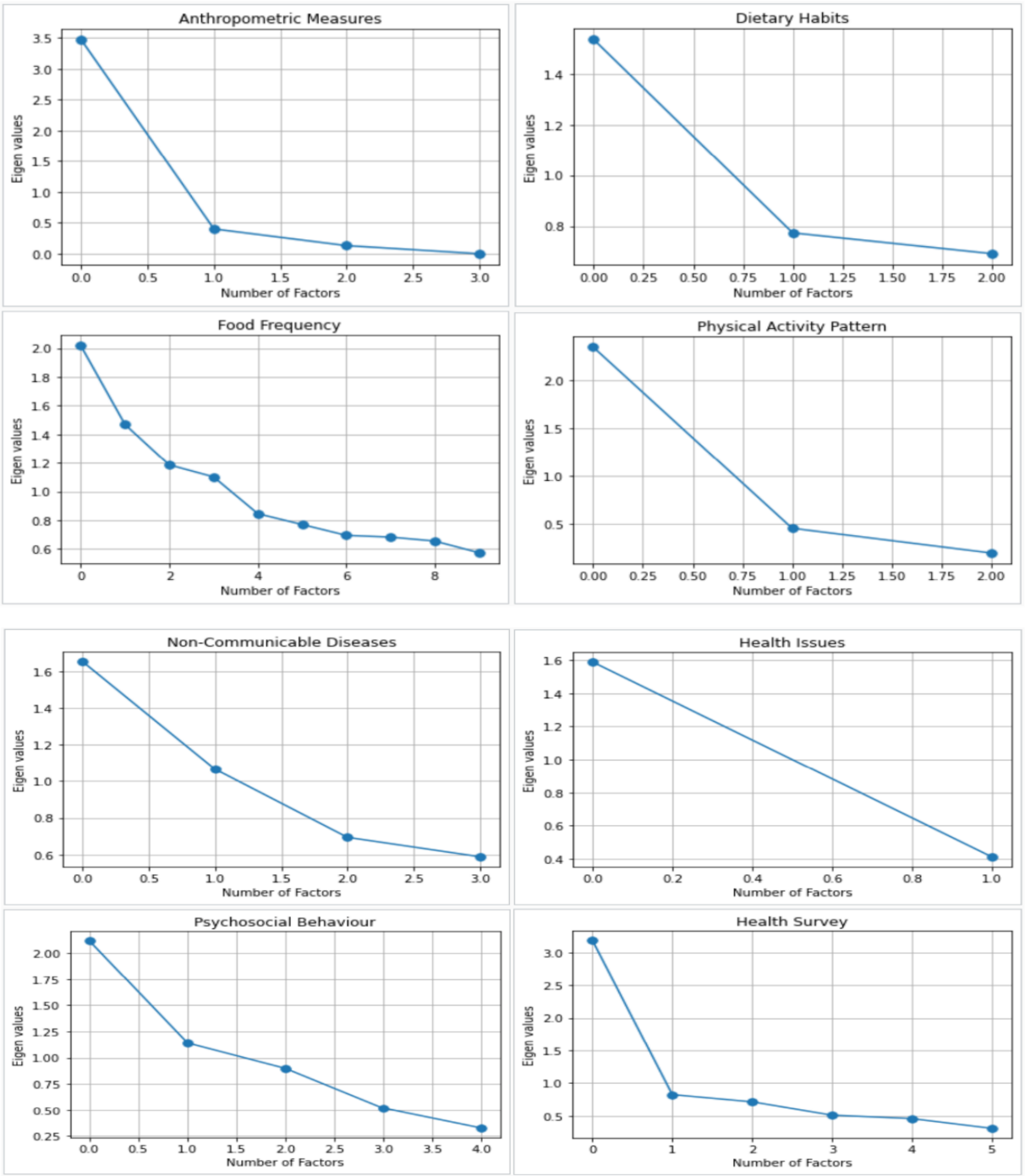

Figure 1. Scree plots of Eigen value while performing sklearn decomposition factor analysis

in Python

4.5 Confirmatory Factor Analysis

The extracted factors from Table 3 should be validated using Average Variance Explained, Construct Validity such as Composite reliability and Discriminant validity which are discussed in the following Table 5.

Table 5. Average Variance Explained and Construct Reliability

S.No	Factors	No. of items	AVE	CR	Decision
1	Anthropometric (ANT)	4	0.8668	0.9628	Accepted
2	Dietary Habits (DH)	3	0.5116	0.7583	Accepted
3	Junk Food (JF)	3	0.5031	0.7505	Accepted
4	Healthy Food (HF)	3	0.4681	0.7245	Accepted
5	Non-Vegetarian (NV)	2	0.6643	0.7981	Accepted
6	Cereals (CER)	2	0.5741	0.7293	Accepted
7	Physical Activity Pattern (PAP)	3	0.7812	0.9144	Accepted
8	NCD1	2	0.6571	0.7931	Accepted
9	NCD2	2	0.6647	0.7982	Accepted
10	Health Issues (HI)	2	0.7939	0.8851	Accepted
11	Physical Influence (PI)	3	0.6088	0.8219	Accepted
12	Entertainment Influence (EI)	3	0.6225	0.7668	Accepted
13	Health Survey (HS)	6	0.5310	0.8657	Accepted

The AVE values of ANT, DH, JF, HF, NV, CER, PAP, NCD1, NCD2, HI, PI, EI and HI are 0.8668, 0.5116, 0.5031, 0.4681, 0.6643, 0.5741, 0.7812, 0.6571, 0.6647, 0.7939, 0.6088, 0.6225 and 0.5310 respectively. These observed AVE values are higher than 0.5 and Construct Reliability values are higher than 0.7 [32]. Both AVE and CR met the requirement for model performance in the current study. Table 4 shows the Construct Reliability (CR) and Average Variance Extracted (AVE) values. The third validity Discriminant validity is discussed in the following Table 6.

Table 6 Discriminant validity – AVE and Correlation Square

Factors	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**
ANT (F1)	**0.8668**												
DH (F2)	0.00113	**0.5116**											
JF (F3)	0.00674	0.01023	**0.5031**										
HF (F4)	0.00047	0.01356	0.00000	**0.4681**									
NV (F5)	0.00009	0.00051	0.00000	0.00000	**0.6643**								
CER (F6)	0.00436	0.00041	0.00000	0.00000	0.00000	**0.5741**							
PAP (F7)	0.00144	0.00018	0.00059	0.00013	0.00062	0.00107	**0.7812**						
NCD1 (F8)	0.00816	0.00495	0.00016	0.00114	0.00061	0.00101	0.00041	**0.6571**					
NCD2 (F9)	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.6647**				

	905	598	116	176	017	354	009	000	**47**				
HI (F10)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.79**			
	058	001	166	004	189	045	002	001	002	**39**			
PI (F11)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	**0.60**		
	040	005	004	006	084	233	076	133	045	202	**88**		
EI (F12)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.62**	
	085	013	008	034	007	007	000	002	039	101	000	**25**	
HS (F13)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.53**
	008	064	011	000	001	001	097	000	011	000	015	000	**10**

Diagonal values are AVE of each factor and off-diagonal values are correlation square between factors.

From the above Table 5, the pair of factored squared correlation values are given in the off-diagonal and Average Variance Explained are given in diagonal. This guideline of discriminant validity according to Hair et al. [32], all the correlation squared correlation values are less than AVE. Therefore it is concluded that each factor is unique with its variables and distinct with each other. From the three statistical tools namely AVE, CR and DV, it is inferred that all the factors and their variables are validated.

5 Conclusions

The findings of this study show the empirical evidence of identifying the relevant factors from the dataset, influencing obesity among adolescent girls. It is possible to conclude that factor analysis is a promising method for identifying major factors. The thirteen major factors are extracted by Exploratory Factor Analysis and validated by Confirmatory Factor Analysis and considered as evidences to the prevalence of overweight and obese.

Further enhancing the rigor of this work, the dataset may be expanded by including an extensive sample size. This expansion will then encompass a wider range of individuals, including those classified as extreme obese. Moreover, obesity can further be categorized into multiple obese levels based on other relevant factors.

References

[1] Thamrin et al., "Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018", Frontiers in Nutrition, Vol. 8:669155:1-15, 2021.

[2] T Kelly et al., “Global burden of obesity in 2005 and projections to 2030”, International Journal of Obesity, Vol.32, pp.1431–1437, 2008.

[3] E.T. Campbell et al., “Adolescent obesity in the past decade: A systematic review of genetics and determinants of food choice”, Journal of the American Association of Nurse Practitioners, Vol.31 (6):344-351, 2019.

[4] Kumar SG, Kotian S, Kotian MS, “Prevalence and determinants of overweight and obesity among adolescent school children of South Karnataka, India”, Indian J Community Med, 35:176, 2014.

[5] Mithra et al., “Lifestyle factors and obesity among adolescents in rural South India” Asian Journal of Pharmaceuticaland Clinical Research, Vol.8(6):81-83, 2015.

[6] Noora Shrestha, “Factor Analysis as a Tool for Survey Analysis”, American Journal of Applied Mathematics and Statistics”, Vol.9 (1):4-11, 2021.

[7] Fatih ORCAN, “Exploratory and Confirmatory Factor Analysis: Which One to Use First?”, Journal of Measurement and Evaluation in Education and Psychology, Vol.9(4):424-421, 2018.

[8] Michael Grabner, “BMI trends, socioeconomic status, and the choice of dataset”, The European Journal of Obesity, Vol.5, pp.112–126, 2012.

[9] Mohemad El Kabbaoui et al, “Prevalence of and risk factors for overweight and obesity among adolescents in Morocco”, EMHJ, Vol. 24(6):512-521, 2018.

[10] Shah et al., "Risk factors associated with overweight and obesity among reproductive aged (15-19) years females in urban areas of Rajbiraj Municipality Saptari", Advances in Obesity, Weight Management & Control, Vol.10(3):84–88, 2020.

[11] Viswambharan JK et al., "Prevalence and determinants of obesity among adolescent school children of North Kerala", International Journal of Community Medicine and Public Health, Vol.7 (8):3142-3148, 2020.

[12] Al-Hazzaa et al., "Lifestyle factors associated with overweight and obesity among Saudi adolescents", BMC Public Health, Vol.12 (354):1-11, 2012.

[13] Rocha et al., "Environmental, Socioeconomic, Maternal, and Breastfeeding Factors Associated with Childhood Overweight and Obesity in Ceará, Brazil: A Population-Based Study", International Journal of Environmental Research and Public Health, Vol.17, 1557:1-11, 2020.

[14] Dhirandra Pratap Singh et al., "Prevalence and associated factors of overweight/obesity among school going children in Chandigarh, India", John Wiley and Sons, Vol. 46 (5):571-575, 2020.

[15] Anuradha et al., "Effect of social and environmental determinants on overweight and obesity prevalence among adolescent school children", Indian Journal of Endocrinology and Metabolism, Vol. 19(2):283-287, 2015.

[16] Duante et al., "Factors Associated with Overweight and Obesity among Adults 20.0 Years and Over: Results from the 2013 National Nutrition Survey, Philippines", Philippine Journal of Science, Vol.148 (1): 7-20, 2019.

[17] Niranjan N et al., "Prevalence of overweight and obesity and factors affecting it in affluent adolescents of Raichur city, International Journal of Contemporary Pediatrics, Vol.2(1):21-24, 2015.

[18] Song N. et al., "Prevalence of overweight and obesity and associated risk factors among adult residents of northwest China: a cross-sectional study", BMJ Open, Vol.9. pp.1-9, 2019.

[19] X. Liu et al., "Prevalence and influencing factors of overweight and obesity in a Chinese rural population: the Henan Rural Cohort Study", Scientific Reports, Vol.8, pp.1-11, 2018.

[20] G.Lee, O.K.Ham, "Factors Affecting Underweight and Obesity among Elementary School Children in South Korea", Asian Nursing Research, Vol.9(4):298-304, 2015.

[21] Cheryl D.Fryar, Margaret D.Carroll and Cynthia L. Ogden,"Prevalence of overweight and obesity among children and adolescents aged 2-19 years: United States, 1963–1965 Through 2011–2012", National Centre for Health Statistics, pp.1-6, 2016.

[22] V.V Khadilkar et al, "Overweight and obesity prevalence and body mass index trends in Indian children", International Journal of Pediatric Obesity, Vol.6:e216-e224, 2011.

[23] Mohamed A Talat and Eman El Shahat, "Prevalence of overweight and obesity among preparatory school adolescents in Urban Sharkia Governorate, Egypt", Egypt, EgyptianPediatric Association Gazette, 20-5, 2016.

[24] Prasad Rv, Bazroy J and Singh Z, "Prevalence of overweight and obesity among adolescent students in Pondicherry, South India", International Journal of Nutrition, J Nutr Pharmacol Neurol Dis., Vol.6:72, 2016.

[25] Rohilla R, Rajput M, Rohilla J, Malik M, Garg D, Verma M, "Prevalence and correlates of overweight/obesity among adolescents in an Urban City of North India", J Fam Med Prim Care, 3:404, 2014.

[26] Singh et al., "Global epidemiology of obesity, vegetarian dietary patterns, and non-communicable disease in Asian Indians", American Journal of Clinical Nutrition, Vol.100, pp.1-6, 2014.

[27] OualidS. How to calculate sample size? 2022. https://www.linkedin.com/pulse/how-calculate-sample-size-oualid-soula/

[28] WHO, "Obesity and Overweight: Fact sheet

311", WHO Media Centre, 2016. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

[29] C.R. Kothari, "Research Methodology Methods and Techniques", Second Edition, New Age International Publishers, 2004.

[30] Kaiser, H. F., "An index of factorial simplicity", Psychometrika, Vol. 39(1):31-36, 1974.

[31] Field, A., "Discovering statistics using SPSS". Sage Publications, 2013.

[32] Hair Jr, Joe, et al., "A primer on partial least squares structural equation modeling (PLS-SEM)", Sage publications, 2021.