Research paper

Assessing the Role of Latent Variables in Modelling and Predicting the Determinants of Obesity

A Structural Equation Modelling Approach

Peter Chimwanda *[1], **Edwin Rupi** [2]

[1] *Department of Mathematics, Chinhoyi University of Technology. Chinhoyi, Zimbabwe*
[2] *Department of Mathematics, Masvingo Teachers College Masvingo, Zimbabwe*

KEYWORDS	ABSTRACT
obesity determinants structural equation modelling latent variables predictive modelling construct	Obesity is a complex and multifactorial public health problem driven by the interaction of biological, behavioural, and environmental determinants. Accurately modelling these interdependencies remains methodologically challenging, particularly when conventional regression and machine-learning approaches prioritise prediction at the expense of interpretability. This study evaluates the contribution of latent variables to obesity modelling and prediction using a comparative structural equation modelling (SEM) framework. Using survey data from 2111 individuals aged 14 to 61 years, two SEM specifications were estimated: (i) a construct-based model incorporating three latent determinants namely, Demographic and Anthropometric Factors (DAF), Dietary and Eating Behaviour (DEB), and Lifestyle and Physical Activity (LPA), together with family history of overweight, and (ii) an observed-variable model in which all predictors were entered directly. Results indicate that both models capture meaningful relationships with obesity level (NObeyesdad), but they differ substantially in explanatory power and interpretability. The construct-based model explains 70.9% of the variance in obesity and identifies DAF, DEB, and family history as the dominant drivers, with lifestyle factors exerting weaker effects. In contrast, the observed-variable model explains 95.9% of the variation in obesity, largely driven by direct anthropometric indicators, especially weight, and multiple behavioural variables. However, this model is considerably more complex and mixes determinants with outcomes and correlates, raising concerns regarding overfitting and theoretical coherence. Overall, the findings demonstrate a clear trade-off between predictive accuracy and conceptual clarity. Latent variable modelling provides a parsimonious and theory-consistent representation of obesity determinants, while observed-variable SEM maximises statistical fit. These results suggest that for predictive purposes, the model without constructs is better.

1. Introduction

Obesity has emerged as one of the most significant global public health challenges of the twenty-first century, contributing substantially to premature mortality and the growing burden of non-communicable diseases. It is commonly defined as a state of chronic positive energy imbalance resulting in excessive accumulation of body fat to a degree that adversely affects health and reduces life expectancy (Asra & Farhath, 2019). In clinical and

epidemiological research, obesity is typically assessed using body mass index (BMI), with values of 30 kg/m² or higher indicating obesity in adults.

There is strong agreement in the literature that obesity does not arise from a single causal factor but represents a chronic, complex, and multifactorial condition shaped by the interaction of genetic, biological, environmental, behavioural, sociocultural, and economic determinants (Janez, 2024). Excess body weight develops gradually through sustained imbalance between energy intake and expenditure, reinforcing the view that obesity is a long-term process influenced by multiple interrelated risk factors (Asra & Farhath, 2019). Consequently, effective management requires lifelong strategies, with lifestyle interventions, particularly caloric reduction and increased physical activity, forming the foundation of prevention and treatment efforts (Janez, 2024).

Despite this consensus regarding complexity, accurate modelling and prediction of obesity remain methodologically challenging. Traditional regression-based approaches are constrained by assumptions of linearity, limited capacity to model interdependencies among predictors, and reliance on a relatively small number of observed variables. These limitations restrict their ability to capture the dynamic and interconnected pathways through which behavioural, socioeconomic, environmental, and biological factors jointly shape obesity risk.

Recent studies have therefore explored machine learning techniques to improve predictive accuracy. For example, Thamrin et al. (2021) evaluated Logistic Regression, Classification and Regression Trees, and Naïve Bayes models using national health survey data, reporting that Logistic Regression achieved the best performance, although agreement between predicted and observed obesity status remained moderate. Similarly, Görmez et al. (2025) reported an accuracy of approximately 72% using Logistic Regression. While such data-driven approaches identify important predictors, they offer limited insight into the underlying causal structure and mediating mechanisms linking determinants to obesity.

In response to these limitations, contemporary research increasingly advocates for modelling frameworks capable of representing both direct and indirect relationships among multiple determinants. Structural equation modelling (SEM) is particularly well suited to this task, as it allows for the simultaneous estimation of complex pathways, incorporation of latent constructs to represent unobservable theoretical concepts, and explicit modelling of measurement error.

However, despite the growing application of SEM in obesity research, limited attention has been given to evaluating the incremental value of latent variables in predictive modelling. Specifically, it remains unclear whether representing determinants as latent constructs improves model fit, interpretability, and predictive performance compared to SEM specifications based solely on observed variables.

Accordingly, this study aims to assess the role of latent variables in modelling and predicting obesity by comparing two SEM frameworks: (i) a latent-variable model in which key determinants are represented as theoretical constructs measured by multiple indicators, and (ii) an observed-variable model in which all predictors enter the model directly. By contrasting these two approaches using the same dataset, this study seeks to determine whether latent constructs provide substantive advantages in explaining obesity levels and improving classification performance. The findings are intended to contribute to methodological guidance for future obesity modelling studies and to support the development of more robust, interpretable, and policy-relevant predictive frameworks.

2. Literature Review: Application of Structural Equation Modelling in Obesity Research

Structural equation modelling (SEM) has become an important analytical approach in obesity research because it allows the simultaneous examination of complex relationships among Family and Genetic Background, Demographic and Anthropometric Factors, Dietary and Eating Behavior, and Lifestyle and Physical Activity. Unlike traditional regression methods, SEM enables the modelling of latent constructs and indirect pathways, providing deeper insight into the mechanisms through which obesity develops. Recent studies have applied SEM across diverse populations to clarify how contextual and individual-level determinants interact to influence obesity risk.

Santiago-Torres et al. (2016) investigated the influence of parental behaviours and the home environment on obesity-related metabolic risk among 187 Hispanic children aged 10 to 14 years in the United States. Using SEM, the authors modelled pathways linking parental dietary and physical activity behaviours to children's health outcomes through household food availability and access to physical activity resources. Higher parental consumption of sugar-sweetened beverages increased their availability at home, which in turn elevated children's intake and BMI. Parental physical activity predicted greater access to exercise equipment, leading to improved cardiovascular fitness. BMI and fitness subsequently predicted insulin resistance. These results highlight the mediating role of environmental factors and demonstrate the ability of SEM to uncover indirect causal pathways.

Mendez et al. (2023) examined socioeconomic status (SES), lifestyle behaviours, and obesity among 1,366 children in Argentina using SEM with latent variables. Obesity, SES, and healthy habits were each specified as latent constructs measured by multiple indicators, including BMI, waist-to-height ratio, and body fat percentage for obesity. Higher SES was associated with healthier behaviours, which in turn reduced obesity levels. The influence of SES on obesity was primarily indirect, mediated by lifestyle factors. This study illustrates the methodological advantage of latent variables in capturing complex constructs more reliably than single observed measures.

Mollaei et al. (2025) analysed data from 3,169 Iranian adults to examine how socioeconomic factors, personal habits, and chronic disease affect obesity through dietary quality and inflammation. SEM results showed that higher SES and healthier habits reduced obesity risk mainly through improved dietary quality and lower inflammatory potential of diets. The model explained approximately 69% of the variance in obesity status, demonstrating the strong explanatory capacity of SEM when behavioural and biological mediators are included.

Keenan et al. (2021) applied SEM to investigate psychosocial pathways linking food insecurity to obesity among adults in the United Kingdom. Food insecurity increased psychological distress, which in turn promoted maladaptive coping behaviours such as emotional eating and alcohol consumption, ultimately increasing BMI. The effect of food insecurity on obesity was largely indirect, reinforcing the importance of modelling mediators rather than relying solely on direct associations.

Collectively, these studies demonstrate that SEM is particularly well suited to obesity research due to its capacity to model complex, multilevel processes and indirect pathways. However, approaches differ substantially in their treatment of variables. Some studies rely primarily on observed indicators (e.g., BMI, food insecurity scores), whereas others explicitly construct latent variables to represent obesity, socioeconomic status, or behavioural patterns.

While latent variable models offer theoretical and statistical advantages such as reduced measurement error and improved construct validity, they also increase model complexity and estimation demands. Consequently, there remains limited empirical evidence comparing the performance of SEM models with and without latent constructs in the context of obesity prediction. This gap motivates the present study's comparative modelling framework.

3. Materials and Methods

This study employed an open-access dataset comprising 2111 individuals aged 14 to 61 years. The data were collected through an anonymous online survey and included 17 variables capturing demographic characteristics, anthropometric measures, dietary habits, lifestyle behaviours, and obesity status.

The variables included Gender, Age, Height, Weight, family history of overweight (yes/no), frequent consumption of high-calorie foods (FAVC), frequency of vegetable consumption (FCVC), number of main meals per day (NCP), consumption of food between meals (CAEC), smoking status (SMOKE), daily water intake (CH2O), calorie monitoring (SCC), physical activity frequency (FAF), time using technology (TUE), alcohol consumption frequency (CALC), mode of transportation (MTRANS), and the target variable NObeyesdad, representing obesity level Insufficient Weight, Normal Weight, Overweight Levels I and II, and Obesity Types I to III.

To evaluate the contribution of latent variables to obesity modelling and prediction, two structural equation models were specified and estimated using the same dataset and outcome variable, with the first model incorporating three latent constructs: Demographic and Anthropometric Factors (DAF), measured by age, gender, height, and weight to represent the individual's biological and physical profile commonly associated with obesity risk; Dietary and Eating Behaviour (DEB), measured by FAVC, FCVC, NCP, CAEC, CH2O, SCC, and CALC to capture dietary patterns, energy intake behaviours, and nutritional quality; and Lifestyle and Physical Activity (LPA), measured by FAF, TUE, MTRANS, and SMOKE to reflect physical activity, sedentary behaviour, transportation habits, and broader lifestyle choices influencing metabolic health.

In addition, Family and Genetic Background, measured by family history of overweight, was included as an observed exogenous variable.

The outcome variable was NObeyesdad, representing obesity level, which served as the endogenous variable for classification and prediction.

Structural paths were specified from DAF, DEB, LPA, and Family and Genetic Background to obesity level, allowing both direct and indirect relationships among determinants to be estimated.

The second model excluded latent variables and instead incorporated all predictors as observed variables directly linked to obesity level. Age, Gender, Height, Weight, dietary indicators and lifestyle indicators were specified individually as exogenous variables, and family history as an endogenous variable, all predicting NObeyesdad. This model retained the SEM framework but omitted the measurement component, thereby

serving as a benchmark for evaluating whether latent constructs improve explanatory power, model fit, and predictive performance.

Both models were evaluated using standard SEM goodness-of-fit indices and predictive performance measures. Comparative analysis focused on, Overall model fit, Explained variance in obesity level, Stability and interpretability of parameter estimates, and Classification performance for obesity categories.

This comparative framework enables an explicit assessment of whether latent variable modelling offers substantive methodological advantages over observed-variable specifications in predicting obesity outcomes.

4. Data Analysis, Results and Discussion

Table 1 Model fit statistics for the Model with constructs

Model	RMSEA	LO 90	HI 90	PCLOSE
Default model	.184	.181	.188	.000
Independence model	.207	.204	.210	.000

Table 2 Model fit statistics for the Model with no constructs

Model	RMSEA	LO 90	HI 90	PCLOSE
Default model	.131	.124	.138	.000
Independence model	.207	.204	.210	.000

Tables 1 and 2 show the RMSEA values of two models, one with and the other without constructs (latent variables). A model with good fit has an RMSEA that is between 0.05 and 0.08. None of these models has an RMSEA in that interval. However, for comparison purposes, it is the model with the smaller of the two RMSEAs which is better. The model with constructs has an RMSEA= 0.184 and for that without constructs RMSEA= 0.131, which makes it better.

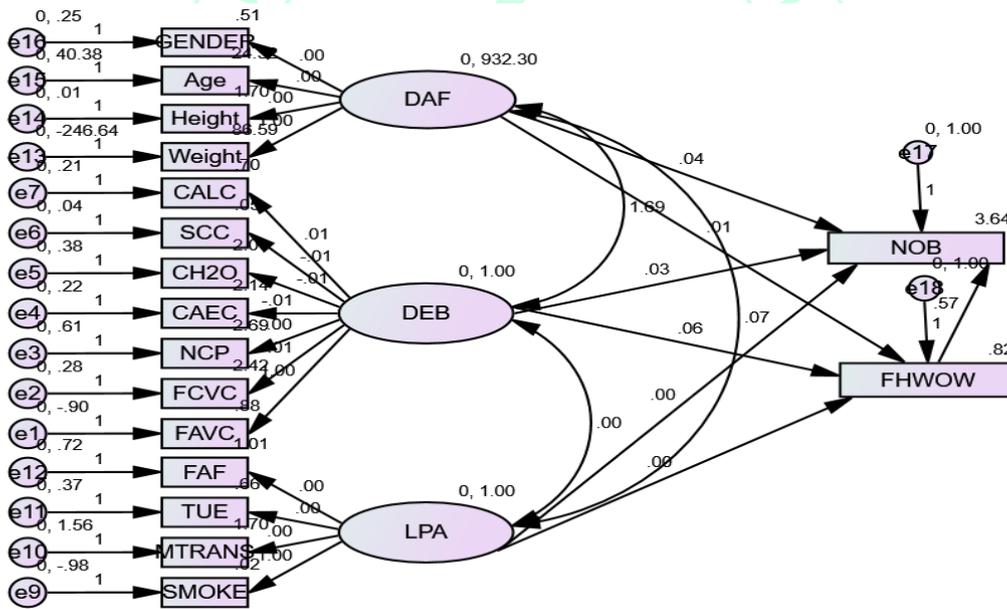

Fig. 1 Path diagram for the Model with Constructs

Fig. 1 shows the structure of the path diagram of the model with construct. There are only two endogenous variables in the data. The AI package, Chatgpt was used for grouping the indicators.
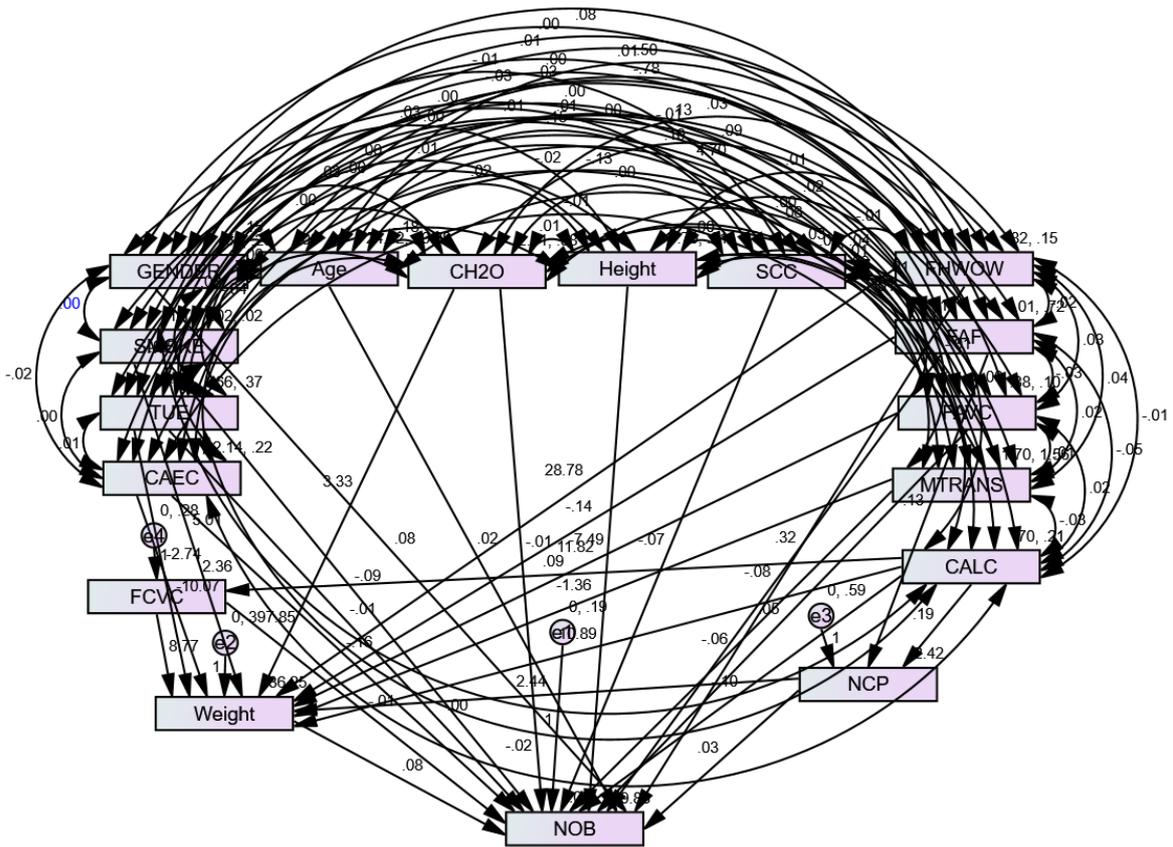
Fig. 2 Path diagram for the Model without Constructs

Fig. 2 is the path diagram for the model without constructs. NOB (level of obesity) is the main endogenous (dependent) variable in the model. All arrows pointing into NOB represent direct effects of predictors on obesity. Double arrowed curves are for capturing covariances and correlations among exogenous variables. This is for purposes of ensuring that all significant paths are accounted for.

Table 3 Regression coefficients for the Model with Constructs

			Estimate	S.E.	C.R.	P	Label
FHWOW	<---	DAF	.007	.001	12.642	***	par_18
FHWOW	<---	LPA	.004	.003	1.150	.250	par_21
FHWOW	<---	DEB	.058	.007	8.669	***	par_22
FAVC	<---	DEB	1.000				
FCVC	<---	DEB	-.010	.004	-2.936	.003	par_1
NCP	<---	DEB	-.004	.005	-.815	.415	par_2
CAEC	<---	DEB	-.010	.003	-3.286	.001	par_3
CH2O	<---	DEB	-.007	.004	-1.690	.091	par_4
SCC	<---	DEB	-.009	.001	-6.492	***	par_5
CALC	<---	DEB	.013	.003	4.260	***	par_6
SMOKE	<---	LPA	1.000				
MTRANS	<---	LPA	.004	.004	1.108	.268	par_7
TUE	<---	LPA	.002	.002	1.138	.255	par_8
FAF	<---	LPA	.000	.003	.084	.933	par_9
Weight	<---	DAF	1.000				
Height	<---	DAF	.002	.000	42.839	***	par_11
Age	<---	DAF	-.003	.003	-1.267	.205	par_12
GENDER	<---	DAF	.001	.000	3.940	***	par_13
NOB	<---	LPA	-.001	.002	-.391	.696	par_16
NOB	<---	DEB	.026	.005	5.695	***	par_17
NOB	<---	FHWOW	.574	.015	38.005	***	par_19
NOB	<---	DAF	.043	.001	58.123	***	par_20

The model with constructs (latent variables) is highly parsimonious at the structural level. NOB is explained by four predictors, DAF, DEB, LPA, and FHWOW. Measurement models show that DEB and DAF are well defined since they have several significant loadings. The construct, LPA, is weakly measured. Most of its indicators are non-significant.

Table 3 also gives the nature of the effects on NOB for the model with constructs. The strongest effects on NOB are provided by FHWOW, size 0.574 or 57.4% and p-value less than 0.001. Other significant effects on NOB for this model are DAF, 4.3% and DEB, 2.6%. LPA has no significant effects on NOB. This means that NOB is mainly driven by DAF, DEB and FHWOW as a key mediator. This provides a clean, theory-consistent explanation that obesity is driven by underlying demographic and anthropometric, and dietary constructs, partly through food-related behaviour.

Table 4 Regression coefficients for the Model without Constructs

			Estimate	S.E.	C.R.	P	Label
NCP	<---	FAF	.130	.020	6.594	***	par_27
NCP	<---	CALC	.185	.037	5.060	***	par_28
FCVC	<---	CALC	.090	.025	3.552	***	par_29
Weight	<---	CH2O	3.329	.740	4.499	***	par_17
Weight	<---	TUE	-2.743	.732	-3.747	***	par_18
Weight	<---	CALC	10.894	.985	11.062	***	par_19
Weight	<---	MTRANS	-1.355	.356	-3.809	***	par_20
Weight	<---	FAF	-.144	.533	-.271	.786	par_21
Weight	<---	CAEC	-10.072	.961	-10.479	***	par_22
Weight	<---	NCP	2.444	.566	4.315	***	par_23
Weight	<---	FCVC	8.774	.816	10.755	***	par_24
Weight	<---	SMOKE	5.006	3.060	1.636	.102	par_25
Weight	<---	FHWOW	28.783	1.184	24.314	***	par_26
Weight	<---	FAVC	11.821	1.420	8.324	***	par_30
NOB	<---	GENDER	.075	.025	3.016	.003	par_1
NOB	<---	Age	.021	.002	10.511	***	par_2
NOB	<---	Height	-7.489	.146	-51.127	***	par_3
NOB	<---	Weight	.077	.000	161.249	***	par_4
NOB	<---	FHWOW	.320	.031	10.391	***	par_5
NOB	<---	FAF	-.079	.012	-6.326	***	par_6
NOB	<---	FAVC	.050	.032	1.553	.120	par_7
NOB	<---	MTRANS	-.062	.010	-6.383	***	par_8
NOB	<---	CALC	-.103	.023	-4.543	***	par_9
NOB	<---	TUE	-.014	.017	-.860	.390	par_10
NOB	<---	SCC	-.072	.048	-1.504	.133	par_11
NOB	<---	CH2O	-.013	.016	-.766	.444	par_12
NOB	<---	CAEC	-.159	.022	-7.294	***	par_13
NOB	<---	NCP	.034	.012	2.757	.006	par_14
NOB	<---	FCVC	-.006	.018	-.350	.727	par_15
NOB	<---	SMOKE	-.086	.068	-1.264	.206	par_16

The model with observed variables only has a very dense structure with many direct paths among indicators going to NOB. It is less parsimonious and is more data-driven; interpretation is fragmented across many coefficients.

In this model, NOB is driven primarily by weight (effect size = 0.077) followed by height (with a large negative effect), age, gender, and several behavioural factors, that include FAF, MTRANS, CALC, CAEC, NCP, and FHWOW. A large number of predictors are statistically significant, which contributes to the very high R^2 value. While the model captures detailed direct associations, it mixes determinants, outcomes, and covariates of obesity, most notably weight, thereby artificially inflating its explanatory power.

Table 5 Squared Multiple Correlations (R^2)

Model with Constructs				Without Constructs	
FHWOW	0.052	SMOKE	0.490	FCVC	0.006
NOB	0.709	CALC	0.001	NCP	0.029
GENDER	0.002	SCC	0.002	Weight	0.415
Age	0.000	CH2O	0.000	NOB	0.959
Height	0.346	CAEC	0.000		
Weight	0.136	NCP	0.000		
FAF	0.000	FCVC	0.000		
TUE	0.000	FAVC	0.974		
MTRANS	0.000				

The decisive comparison comes from the squared multiple correlations (R^2) that is sown in Table 5. For the model with constructs, R^2 is 0.709 while for that without constructs R^2 is 0.959. The construct model explains about 71% of the variance in NOB. The non-construct model explains about 96%, which is extremely high and indicates superior explanatory power. Thus, in terms of explaining variation in NOB alone, the model without constructs is clearly better.

5. Conclusion

This study set out to evaluate whether incorporating latent constructs into structural equation models improves the modelling and prediction of obesity determinants. By comparing a construct-based SEM with an observed-variable SEM using the same dataset, the analysis highlights an important methodological trade-off. The construct-based model provides a clear and theoretically grounded representation of obesity as a function of underlying demographic and anthropometric characteristics, dietary behaviour, and family history of overweight. It explains a substantial proportion of variation in obesity level (70.9%), despite relying on only a small number of structural predictors. This parsimony enhances interpretability, reduces dimensionality, and supports generalisation across populations by focusing on stable conceptual mechanisms rather than numerous individual indicators.

Conversely, the observed-variable model achieves superior explanatory power (95.9%), driven primarily by direct anthropometric measures, especially weight and multiple behavioural variables. While this approach is advantageous for short-term prediction within the same dataset, it results in a dense and fragmented structure that blurs the distinction between causes, mediators, and consequences of obesity. Such models are more vulnerable to multicollinearity, overfitting, and reduced external validity.

Taken together, the findings suggest that latent variable SEM is preferable for theory development, policy analysis, and understanding the mechanisms underlying obesity, whereas observed-variable SEM may be better suited for applications where maximising predictive accuracy is the sole objective. Future research could extend this comparison using longitudinal data and out-of-sample validation to further assess the stability and generalisability of both modelling strategies.

References

1. Asra Fathima and Farhath Khanum (2019) OBESITY – An Overview, International Journal of Life Sciences Research Vol. 7, Issue 2.
2. Görmez Y, Yagin FH, Yagin B, Aygun Y, Boke H, Badicu G, De Sousa Fernandes MS, Alkhateeb A, Al-Rawi MBA and Aghaei M (2025) Prediction of obesity levels based on physical activity and eating habits with a machine learning model integrated with explainable artificial intelligence. *Front. Physiol.* 16:1549306. doi: 10.3389/fphys.2025.1549306
3. Janez, A.; Muzurovic, E.; Bogdanski, P.; Czupryniak, L.; Fabryova, L.; Fras, Z.; Guja, C.; Haluzik, M.; Kempler, P.; Lalic, N.; et al. Modern Management of Cardiometabolic Continuum: From Overweight/Obesity to Prediabetes/Type 2 Diabetes Mellitus. Recommendations from the Eastern and Southern Europe Diabetes and Obesity Expert Group, Diabetes Ther (2024) 15:1865–1892, https://doi.org/10.1007/s13300-024-01615-5.
4. Keenan Gregory S., Christiansen Paul and Hardman Charlotte A. (2021) Household Food Insecurity, Diet Quality, and Obesity: An Explanatory Model, Obesity published Wiley Periodicals LLC on behalf of The Obesity Society (TOS), VOLUME 29, NUMBER 1.
5. Mendez Ignacio, Fasano María Victoria and Orden Alicia B., Exploring factors associated with obesity in Argentinian children using structural equation modelling, Cad. Saúde Pública 2023; 39(7): e00087822.
6. Mollaei Somayeh, Roshani Daem, Salari-Moghddam Asma, Moradi Yousef and Moradpour Farhad (2025) Dietary and Lifestyle Predictors of Obesity: A Structural Equation Model Approach, Research Square.
7. Santiago-Torres M, Cui Y, Adams AK, Allen DB, Carrel AL, Guo JY, LaRowe TL, Schoeller DA. Structural equation modelling of the associations between the home environment and obesity-related cardiovascular fitness and

insulin resistance among Hispanic children. Appetite. 2016 Jun 1;101: 23-30. doi: 10.1016/j.appet.2016.02.003. Epub 2016 Feb 3. PMID: 26850309; PMCID: PMC5912911.

8. Thamrin SA, Arsyad DS, Kuswanto H, Lawi A and Nasir S (2021) Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018. Front. Nutr. 8:669155. doi: 10.3389/fnut.2021.669155.