BMC Medical Research
Methodology

RESEARCH

Open Access

Development of a diagnostic predictive model for determining child stunting in Malawi: a comparative analysis of variable selection approaches

Jonathan Mkungudza[1], Halima S. Twabi[1*] and Samuel O. M. Manda[2]

Abstract

Background Childhood stunting is a major indicator of child malnutrition and a focus area of Global Nutrition Targets for 2025 and Sustainable Development Goals. Risk factors for childhood stunting are well studied and well known and could be used in a risk prediction model for assessing whether a child is stunted or not. However, the selection of child stunting predictor variables is a critical step in the development and performance of any such prediction model. This paper compares the performance of child stunting diagnostic predictive models based on predictor variables selected using a set of variable selection methods.

Methods Firstly, we conducted a subjective review of the literature to identify determinants of child stunting in Sub-Saharan Africa. Secondly, a multivariate logistic regression model of child stunting was fitted using the identified predictors on stunting data among children aged 0–59 months in the Malawi Demographic Health Survey (MDHS 2015–16) data. Thirdly, several reduced multivariable logistic regression models were fitted depending on the predictor variables selected using seven variable selection algorithms, namely backward, forward, stepwise, random forest, Least Absolute Shrinkage and Selection Operator (LASSO), and judgmental. Lastly, for each reduced model, a diagnostic predictive model for the childhood stunting risk score, defined as the child propensity score based on derived coefficients, was calculated for each child. The prediction risk models were assessed using discrimination measures, including area under-receiver operator curve (AUROC), sensitivity and specificity.

Results The review identified 68 predictor variables of child stunting, of which 27 were available in the MDHS 2016–16 data. The common risk factors selected by all the variable selection models include household wealth index, age of the child, household size, type of birth (singleton/multiple births), and birth weight. The best cut-off point on the child stunting risk prediction model was 0.37 based on risk factors determined by the judgmental variable selection method. The model's accuracy was estimated with an AUROC value of 64% (95% CI: 60%-67%) in the test data. For children residing in urban areas, the corresponding AUROC was AUC = 67% (95% CI: 58–76%), as opposed to those in rural areas, AUC = 63% (95% CI: 59–67%).

Conclusion The derived child stunting diagnostic prediction model could be useful as a first screening tool to identify children more likely to be stunted. The identified children could then receive necessary nutritional interventions.

Keywords Child stunting, Area under-receiver operator curve (AUROC), Prediction, MDHS, Model

*Correspondence:
Halima S. Twabi
htwabi@unima.ac.mw
Full list of author information is available at the end of the article

Background

Child malnutrition remains a public health burden in most developing countries, particularly in Sub-Saharan Africa (SSA) and South Asia. An important measurement of childhood malnutrition is stunting, which is defined as poor linear growth, indicated by height for age less than -2 standard deviations from the World Health Organization (WHO) 2006 child growth standards median [1] SSA has the highest burden of stunted children with over one-third of children in 2019 [2]. Worldwide, over 3.1 million children die annually, either directly or indirectly, because of malnutrition and many more suffer from impaired growth as a result of malnutrition [3]. The problem of childhood stunting is a focus of some global initiatives including the Global Nutrition Targets for 2025 and the Sustainable Development Goals for Zero Hunger [4]. The World Health Organization (WHO) considers stunting a public health problem when the prevalence of stunting among children under five years of age is greater than 20% [5]. The prevalence of child stunting was 39% in 2019, thus placing Malawi as a country with a huge health problem regarding child stunting [6].

Childhood stunting affects child morbidity and mortality adversely resulting in poor childhood development and educational performance and increases the risks of infections, which may contribute negatively to adult health and economic productivity [7, 8]. Timely and accurate detection of children who are most likely to be stunted may ensure the prevention of detrimental health outcomes. It may help in delivering tailor-made interventions for the optimum use of available resources. Achieving this goal necessitates the development of diagnostic or clinical prediction models, also known as clinical prediction rules, prognostic models, or risk scores, depending on the study's aim and design. Prognostic prediction models are designed to calculate the probabilities of specific patient outcomes over time by considering a variety of clinical and non-clinical factors. Diagnostic prediction models estimate an individual's probability of having a specific health condition (often a disease) at a given time [9, 10]. The aim is to tailor policy decisions to meet the needs of individuals [11–13].

A risk prediction model is a mathematical tool that combines different predictor variables to estimate the probability of occurrence of an outcome of interest [14, 15]. The determination of specific parameter values for the predictor variables, derived from data the model has not seen (the training data), is crucial for calculating a risk score. For prediction models, the choice of the modelling approach is guided by the nature of the predicted outcome. A linear regression method is employed when dealing with continuous outcomes. However, when the outcome of interest is binary, as in the case of predicting stunting status (stunted or not), a logistic regression model is employed. This logistic regression model effectively assesses the likelihood of a binary outcome occurring based on the given set of predictor variables. A Cox proportional hazards regression model is used for a scenario where the outcome is related to the timing of an event, like the occurrence of a particular event over time enabling predictions about the timing of particular events of interest. For count data, a Poisson or negative binomial model has been proposed [16].

Regression modelling faces challenges like overfitting and non-convergence, mainly when dealing with large datasets containing numerous variables. Variable selection methods have been developed to address these issues to identify a concise set of influential predictors from a larger pool. Standard methods, including backward elimination, forward selection, and stepwise selection, are preferred due to their straightforward algorithms [17]. Recent advancements have introduced penalised selection procedures such as the least absolute shrinkage and selection operator (LASSO) and adaptive LASSO. These have gained popularity for enhancing model prediction and inference by simplifying the model. Moreover, tree-based methods like Random Forests and Boruta have been employed to improve accuracy in selecting the most relevant variables from a vast pool, addressing limitations observed in test-based variable selection procedures [18]. Alternative approaches, such as the firefly algorithm, are employed for counting and skewed outcomes using Poisson or Negative Binomial and gamma distribution, respectively [16, 19]. Various simulation studies have demonstrated the effectiveness of these variable selection methods under different conditions [18, 20, 21]. Applying these methodologies to a survey-based study in Malawi, this paper aims to use the variable selection approaches to identify the best predictors that can be used to develop and validate a child stunting diagnostic prediction model.

Determinants of stunting are well-known and studied. However, few studies in Sub-Saharan Africa have developed and validated risk-predictive models to identify children likely to be stunted [22]. One such study was conducted by Hasegawa et al. [23], who developed a screening tool to predict malnutrition among young children in Zambia [24]. The study, however, used data collected from a restricted area (one health facility) from a rural location, limiting the model's generalizability. The Lives Saved Tool is another predictive model used to estimate the impact of specified changes in key interventions on stunting among children under five years [25]. The Lives tool falls short of predicting which children are stunted or not. Hanieh et al. [25], also developed and externally validated an early life predictive model to predict the risk of stunting in preschool children in Vietnam at three years of age. This stunting prediction model has limitations as it was developed and validated in a rural setting, leaving out urban locations and other areas. Also, the prediction was not based on all ages below five years.

Malawi is a developing country and suffers a shortage of skilled community health workers (HW) and resources to timely detect stunting among children. Thus, modelling techniques emerge as valuable tools that could aid community health workers in promptly identifying children at risk of stunting. This would inform the formulation of proper policies to overcome the challenge of child stunting. Numerous studies have previously been conducted to identify predictors of stunting in Sab-Saharan Africa, specifically in Malawi however, very few studies have been conducted to combine these predictors into a mathematical model to predict stunting. This study aimed to derive and validate a diagnostic risk (score) model that could help identify children who are stunted in Malawi using a nationally representative cross-sectional sample.

Variable selection methods

Using variable selection algorithms, namely forward selection, backward elimination, stepwise selection; LASSO; random forest, and judgmental selection, different sets of variables were selected from the list of candidate predictors to identify relevant predictors of stunting. The research used more than one variable selection method to compare the predictive ability of each model fitted using the different sets of variables selected by different methods. The following sections provide a brief description of the methods used.

Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO is a linear regression analysis method that reduces both the sum of squares of errors and the sum of the absolute values of regression coefficients. The regression coefficient $\widehat{\beta}$, is determined by minimizing the following formula:

$$\min\boldsymbol{\beta}\left\{\frac{1}{2}\sum\nolimits_{i=1}^{n}\left(y_i - X_i^T\boldsymbol{\beta}\right) + \lambda\sum\nolimits_{j=1}^{p}\left|\beta_j\right|\right\}$$

$\boldsymbol{\beta}$ represents the vector of regression coefficients.

y_i is the observed response for the i-th observation.

X_i is the vector of predictor variables for the i-th observation.

λ is the regularization parameter that controls the value of the penalty term.

LASSO is an algorithm that has a built-in variable selection method. The use of LASSO as a feature selection technique can be seen from the fact that decreasing the values of λ in the equation below leads to shrinkage of regression coefficients and some of these even become zero.

Random forest

Boruta [26], is a non-linear Random Forest (RF)-based variable selection method. It performs feature selection by identifying important features from many potential predictors. It is based on the random forest process. Random shuffling of the original variables creates shadow features. Variable importance for each shadow feature is computed and the highest score becomes the threshold for selection. The shadow features are permuted at random for each new RF iteration. A hit is defined as the number of times a variable has a higher relevance score than maximal importance of random variables. Using the properties of binomial distribution with probability of success $p=0.5$, it is easy to argue that for N number of RF shuffles, the expected number of hits is E(N) = 0.5N, while the standard deviation is 0.25N. If the importance of the original feature is higher than the threshold, the variable is kept in the model, otherwise, it is discarded. The method comes to a halt when only the most important variables remain in the test or when the maximum number of iterations with some uncertain features has been reached. The Boruta algorithm was implemented in the random forest's R package (Boruta).

Forward, backward and stepwise selection methods

Forward selection starts with an empty model and adds variables one at a time. At each step, the algorithm evaluates all potential predictors and selects the one that results in the greatest increase in a chosen criterion, Akaike Information Criterion (AIC). Backward elimination begins with a model that incorporates all potential predictor variables. At each step, the method evaluates the impact of removing each variable and selects the one whose removal results in the least deterioration in the chosen criterion. Stepwise selection alternates between forward and backward steps. At each step, it evaluates both adding and removing variables and chooses the action that optimizes the selected criterion.

The AIC is used as a stopping rule for the forward, backward, and stepwise variable selection methods. AIC is given as;

$$AIC = 2K - 2\ln L\left(\widehat{\beta}\right)$$

Where K is the number of estimated parameters in the candidate model and $L\left(\widehat{\beta}\right)$ is the estimate from the log-likelihood function. AIC quantifies the comparative information content of a model by utilizing maximum likelihood estimates and counting the number of parameters involved in the model, as indicated in the abovementioned formula. It is employed to assess and distinguish various potential models, helping in the identification of the best-fit model that is consistent with the given data. It is also used as a stopping rule in variable selection methods. The model giving the smallest AIC over the set of models considered is selected as the best model.

Judgement variable selection method

Numerous methods for variable selection have been proposed; however, there is no consensus on a single approach that consistently performs well under all circumstances. Therefore, for each dataset, the technique for variable selection should be carefully chosen [27]. In statistical analysis, prior knowledge derived from scientific literature is considered the primary basis for determining the inclusion or exclusion of covariates. However, such information may not always be accessible for all research questions [28]. The judgement variable selection method relies on field expertise acquired through reviewing relevant literature and consulting with experts.

Data source

The study used data on under-five children extracted from the 2015–16 Malawi Demographic and Health Survey (MDHS). The MDHS is a nationally representative survey conducted by the National Statistical Office in collaboration with the demographic health survey program funded by the United States Agency for International Development (USAID). The MDHS collected up-to-date information on mothers' demographic and health information on child nutrition. The MDHS collected anthropometric data for the under-five children in selected households. The study analyzed data on 5149 children who had a stunting outcome. Details of the sampling procedure for the MDHS can be obtained in the 2015–16 MDHS report [29].

Outcome and predictor variables

The outcome variable of this study was stunting and was calculated based on the anthropometric indicator (height-for-age) among under-five children. The height-for-age z-score of the children was calculated using growth standards published by the WHO in 2006 [1]. The height-for-age z-score is a metric used to assess linear growth retardation and cumulative growth deficits in children. Children with height-for-age Z-score below minus two standard deviations (-2SD) from the median of the WHO reference population were considered stunted [30]. To identify determinants of childhood stunting, we conducted a subject review of the available literature PubMed and Google scholar databases was conducted for relevant articles between August to December 2021 to see what has been studied and found about the determinants of childhood stunting in sub-Saharan Africa. Several searches were performed with the search terms "Determinants of stunting AND Africa" or "Risk factors of stunting AND Africa" or "Predictors of stunting AND Africa". All duplicate articles and those not done in Sub-Saharan Africa were eliminated from the results.

Statistical methods

All model development and analysis were performed with R software (Version 3.6.2, R Foundation). The data were randomly partitioned into a training set (80%) and a testing set (20%). Six multivariate logistic regressions were trained using the sets of variables selected by the six variable selection algorithms on the training set (80%), and the models were validated on the remaining test set (20%). Another logistic regression model was also fitted based on the risk factors commonly selected by all variable selection algorithms. These models were compared for their discriminative ability and predictive performance. The R package `glmnet` statistical software (R Foundation) was used to perform the logistic regression. The probability cut-off points (discriminative value) that define positive and negative test results were estimated using the SpEqualSe criterion implemented in the `OptimalCutpoints` package in R using the training data set (80%). SpEqualSe criterion is the method for computing the optimal cut-off point, which minimizes the absolute deference of sensitivity and specificity [31]. Most often, research studies utilize a default cut point of 0.5 [23, 32, 33]. For comparison purposes, For the sake of comparison, the study also used the cut-off point of 0.5 derived from the Bayes rule. The estimated probability cut-off points were then utilized to derive several performance measures, including the Area under the Operating Curve (AUROC). The AUROC was employed as a metric for assessing the model's discriminative ability. In the formation of the receiver operating characteristic (ROC) curve at various thresholds, the "true positive rate" (TPR) was plotted against the "false positive rate." AUC values range from 0.50 to 1. A greater AUC indicates a higher predictive capacity. A model with an AUC of 1.0 is a perfect discriminator, 0.90 to 0.99 is considered excellent, 0.80 to 0.89 is good, 0.70 to 0.79 is fair, and 0.51 to 0.69 is considered poor. Sensitivity was calculated using the following formula: Sensitivity = True Positives / (True Positives + False Negatives). Specificity was calculated using the following formula: Specificity = True Negatives / (True Negatives + False Positives).

Results

Distribution of stunting in the data set

Overall, the prevalence of child stunting was 35.6% (95% CI: 34.2% -36.9%). In the training set, the prevalence was 35.9% (34.4%—37.4%) and 34.5% (95% CI: 31.5%—37.6%) in the testing data set.

Potential predictor variables

The study reviewed 28 papers and identified 68 potential predictors of child stunting from various sources, including [30, 34, 35] particularly in the context of Sub-Saharan Africa, with 27 found in the Malawi Demographic and Health Survey (MDHS). Additional sources are provided in the Appendix section. The predictive variables for child stunting identified in the MDHS included demographic factors; mother's education level, ethnicity, child's age, type of residence, child's gender, household head's age, marital status, number of under-five children in the household, mother's age, mother's body mass index, family size, religion, and region. Economic factors; household wealth index and mother's occupation. Obstetric and child morbidity variables; childbirth weight, birth order, mode of delivery, diarrhoea episodes, child anaemia level, preceding birth interval, place of delivery, number of births, type of delivery assistance, cough or fever episodes, and distance to the nearest health facility.

Predictor selection

Boruta (Random Forest) selected 10 variables, which are; the type of birth (single/multiple), age of the child, wealth index, birth weight of the child, location (rural/urban), distance to health facility, age of household head, birth order, body mass index of the mother and size of the household (see Fig. 1). In Fig. 1, the confirmed important variables are those shown in green and those in red represent variables that are not important.

The variables selected by LASSO included location, wealth index, maternal age, age of household head, age of the child, household size, body mass index of the mother, distance to a health facility, number of under-five children, religion, maternal education, type of birth, birth order of the child, region, diarrhoea, maternal occupation, anaemia, delivery assistance, sex of the child, and sex of household head, (see Table S1).

The risk factors selected by backward, forward, and stepwise variable selection algorithms are indicated in Table 1.

All variable selection methods selected five common factors: household wealth index, child age, household size, type of birth (singleton/multiple births), and birth weight. The top five most important variables according to variable importance ranking in the best model were the age of a child, birth weight of a child, type of birth, wealth index, and sex of the child (see Fig. 2).

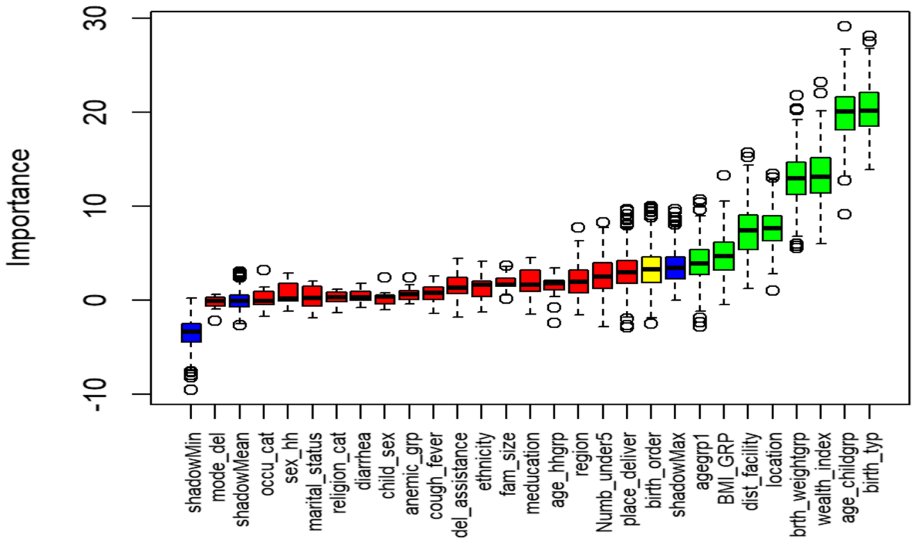

Fig. 1 Selected variables: Random Forest, Boruta. *(birth_typ = birth type, age_childgrp = age of child, wealth_index = wealth index, brth_weightgrp = child birth weight, location = location, dist_facility = distance to a health facility, birth_order = birth order, age_hhgrp = age of household head, BMI_GRP = body mass index, fam_size = family size, agegrp1 = mother's age, place_deliver = place of delivery, Numb_under5 = number of under five children, meducation = mother's education level, del_assistance = delivery assistance, marital_status = marital status, child_sex = sex of child, sex_hh = sex of household head, religion_cat = religion of mother, anemic_grp = anemia, mode_del = mode of delivery, occu_cat = mother's occupation, cough_fever = cough/fever)*

Table 1 Variables selected by automated methods

Backward	Forward	Stepwise
Age of child	Age of child	Age of child
Type of birth	Birth weight	Birth weight
Wealth index	Type of birth	Type of birth
Body mass index of the mother	Wealth index	Wealth index
Maternal education	Body mass index of the mother	Body mass index of the mother
Sex of the child	Ethnicity	Ethnicity
Number of under-five children	Sex of the child	Sex of the child
Diarrhoea	maternal occupation	maternal occupation
Distance to a health facility	Distance to a health facility	Distance to a health facility
Household size	Location	Location
Delivery assistance	Diarrhoea	Diarrhoea
Age of household head	Number of under-five children	Number of under-five children
Marital status		

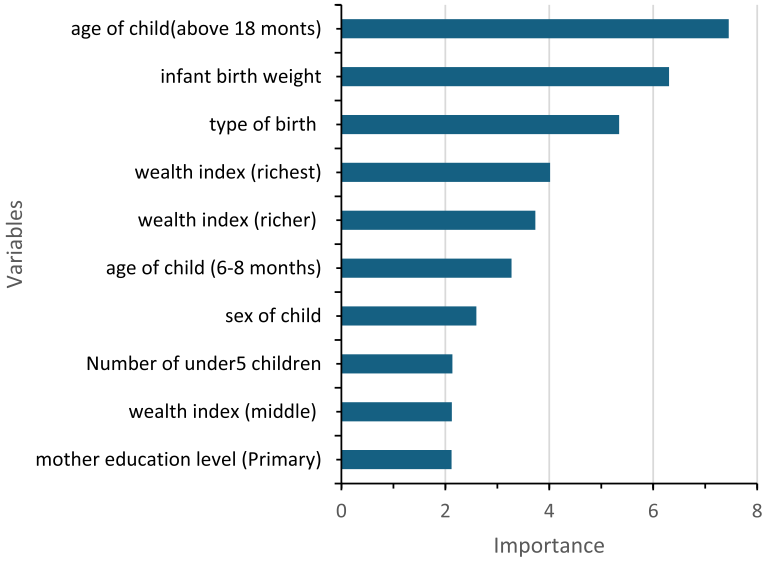

Fig. 2 Variable importance plot for the best model (Judgement model)

Model evaluation and performance

The best cut-off point was at probability (stunting) equal to 0.37. The model based on risk factors determined by judgment outperformed the other models AUC = 64% (95% CI: 60%-67%) and 65% (95% CI: 64%-67%) in the testing set and training set, respectively (Tables 2 and 3). Based on the risk factors identified commonly by all variable selection methods, the predictive performance was 62% (95% CI: 59.0%-66.0%). The sensitivity indicates that 61% of the children, who were stunted, were correctly classified as being stunted and the specificity of 60% indicates that children who were not stunted were correctly classified as not being stunted. PPV for all the models were high for the cutoff points of 36% and 37% ranging between 0.71 to 0.74 (Tables 2 and 3). In the study, the performance estimates based on the cutoff point of 0.5 for the test data do well in terms of specificity but all have some poor sensitivity and low PPV's ranging between 0.40 to 0.57 (see Table 4). The estimated cutoff points for each model seek an equilibrium between sensitivity and specificity. This is important for this study as the interest is in detecting a stunted child than finding a non-stunted child. Hence, more attention is on the sensitivity than the specificity.

Calibration plots illustrating the analysis results for each model are presented in Fig. 3. The calibration performance varied across models, with some models

Table 2 Model performance measures using estimated cutoff points on the training set

	Cut point	AUC (95% CI)	Sensitivity	Specificity	Misclassification error	Accuracy	NPV	PPV
Model 1	0.36	0.66(0.64–0.68)	0.62	0.62	0.39	0.61	**0.74(0.72–0.76)**	**0.45(0.43–0.47)**
Model 2	0.36	0.66(0.65–0.68)	0.62	0.62	0.38	0.62	**0.74(0.72–0.76)**	**0.45(0.43–0.47)**
Model 3	0.36	0.66(0.65–0.68)	0.62	0.62	0.38	0.62	**0.74(0.72–0.76)**	**0.45(0.43–0.47)**
Model 4	0.37	0.65(0.64–0.67)	0.61	0.61	0.39	0.61	**0.73(0.72–0.76)**	**0.42(0.40–0.44)**
Model 5	0.37	0.67(0.65–0.69)	0.63	0.63	0.37	0.63	**0.73(0.71–0.75)**	**0.44(0.42–0.47)**
Model 6	0.37	0.65(0.64–0.67)	0.61	0.60	0.39	0.61	**0.74(0.72–0.75)**	**0.45(0.42–0.47)**
Model 7	0.37	0.64(0.62–0.66)	0.60	0.61	0.39	0.61	**0.72 (0.70–0.75)**	**0.42 (0.39–0.45)**

Model 1 was constructed using the variables selected by backward variable selection algorithms. Model 2 was constructed using the variables selected by forward variable selection algorithms. Model 3 was constructed using the variables selected by stepwise variable selection algorithms. Model 4 was constructed using the variables selected by random forest variable selection algorithms. Model 5 was constructed using the variables selected by LASSO variable selection algorithms. Model 6 was constructed using the variables selected by judgment. Model 7 was constructed using the variables that were common to the 6 models

PPV Positive Predictive Value, *NPV* Negative Predictive Value

Table 3 Model performance measures using estimated cutoff points on the test data

	Cut-off point	AUC (95% CI)	Sensitivity	Specificity	Misclassification error	Accuracy	PPV	NPV
Model 1	0.36	0.63(0.59–0.66)	0.57	0.59	0.42	0.58	**0.73 (0.67–0.78)**	**0.37 (0.33–0.40)**
Model 2	0.36	0.62(0.59–0.66)	0.57	0.59	0.42	0.58	**0.74 (0.69–0.78)**	**0.40 (0.36–0.43)**
Model 3	0.36	0.62(0.59–0.66)	0.57	0.59	0.42	0.58	**0.74 (0.69–0.78)**	**0.40 (0.36–0.43)**
Model 4	0.37	0.62(0.59–0.66)	0.58	0.61	0.40	0.60	**0.74 (0.69–0.78)**	**0.39 (0.36–0.043)**
Model 5	0.37	0.62(0.59–0.67)	0.54	0.62	0.41	0.59	**0.72 (0.68–0.76)**	**0.40 (0.37–0.45)**
Model 6	0.37	0.64(0.60–0.67)	0.61	0.60	0.40	0.60	**0.74 (0.71–0.78)**	**0.44 (0.40–0.48)**
Model 7	0.37	0.62(0.59–0.66)	0.59	0.59	0.41	0.59	**0.71 (0.69–0.76)**	**0.41(.038–0.47)**

Model 1 was constructed using the variables selected by backward variable selection algorithms. Model 2 was constructed using the variables selected by forward variable selection algorithms. Model 3 was constructed using the variables selected by stepwise variable selection algorithms. Model 4 was constructed using the variables selected by random forest variable selection algorithms. Model 5 was constructed using the variables selected by LASSO variable selection algorithms. Model 6 was constructed using the variables selected by judgment. Model 7 was constructed using the variables that were common to the 6 models

PPV Positive Predictive Value, *NPV* Negative Predictive Value

Table 4 Model performance measures using a cutoff point of 0.5 on test data

	Cut point	AUC (95% CI)	Sensitivity	Specificity	Misclassification error	Accuracy	PPV	NPV
Model 1	0.5	0.63(0.59–0.66)	0.34	0.81	0.35	0.65	0.40 (0.35–0.45)	0.68 (0.65–0.72)
Model 2	0.5	0.59(0.56 -0.63)	0.27	0.83	0.37	0.64	0.41 (0.34–0.48)	0.67 (0.64–0.70)
Model 3	0.5	0.59(0.56 -0.63)	0.27	0.83	0.37	0.64	0.41 (0.34–0.48)	0.67 (0.64–0.70)
Model 4	0.5	0.57(0.53 -0.61)	0.25	0.8	0.39	0.61	0.45 (0.38–0.53)	0.68 (0.64–0.71)
Model 5	0.5	0.62(0.59–0.66)	0.18	0.91	0.34	0.66	0.46 (0.38–0.54)	0.68 (0.64–0.71)
Model 6	0.5	0.64(0.60–0.67)	0.17	0.93	0.33	0.67	0.57 (0.47–0.66)	0.68 (0.65–0.71)
Model 7	0.5	0.63(0.59–0.66)	0.15	0.93	0.34	0.66	0.48 (0.39–0.55)	0.68(0.65–0.70)

Model 1 was constructed using the variables selected by backward variable selection algorithms. Model 2 was constructed using the variables selected by forward variable selection algorithms. Model 3 was constructed using the variables selected by stepwise variable selection algorithms. Model 4 was constructed using the variables selected by random forest variable selection algorithms. Model 5 was constructed using the variables selected by LASSO variable selection algorithms. Model 6 was constructed using the variables selected by judgement. Model 7 was constructed using the variables that were common to the 6 models

PPV Positive Predictive Value, *NPV* Negative Predictive Value

demonstrating good calibration at specific probability ranges, as indicated by the calibration slopes and gradients, while others showed poor discrimination. For instance, the judgment model showed good calibration for predicted probabilities between 0.3 and 0.4, whereas the LASSO model performed well for

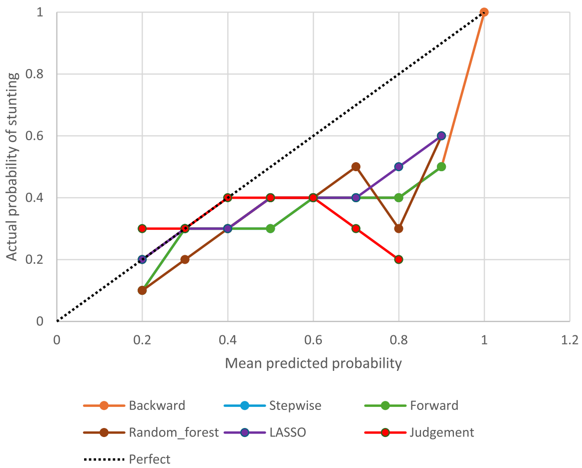

Fig. 3 Calibration plots for the fitted models

probabilities ranging from 0.2 to 0.3.The selected model has provided a predictive tool that displays a good ability to discriminate between stunted children and those not stunted, particularly in children residing in urban areas (AUC = 67% (95% CI: 58–76%) in children living in urban versus AUC = 63% (95% CI: 59–67) in children living in rural areas) (Table 5).

Figure 4 presents the receiver operating characteristic curve for the models. The findings of this study show that the six prediction models have a better discrimination ability compared to a random classifier as indicated by the ROC curves in Fig. 3.

Discussion

The study set out to develop and validate a child stunting prediction score based on the best predictive model in Malawi. The study aimed to develop a model that could accurately classify a child as stunting using important factors available in the data. The model was based on predictor variables obtained from fitting a multivariate logistic regression model to child stunting using data from the 2015–16 Malawi Demographic and Health Survey (MDHS). Using six variable selection methods, namely backwards, forward, stepwise, Boruta based on the random forest, LASSO, and judgment, we identified nine easily measured key predictors of child stunting.

The best-performing model was based on the predictors selected using the judgment method, and these included the age of the child, the weight of the child at birth, type of birth, sex of the child, wealth index category of the household, number of under-five children in the household, location, and maternal education. The discriminative ability of our model was different for the type

Table 5 Performance of the selected model after adjusting for sex and residence

	Sex of a child		Residence	
	Female	Male	Urban	Rural
AUC (95% CI)	0.64 (0.59–0.70)	0.63 (0.58 -0.68)	0.67(0.58–0.76)	0.63 (0.59–0.67)
Sensitivity (%)	0.86	0.79	0.62	0.7
Specificity (%)	0.26	0.27	0.59	0.45

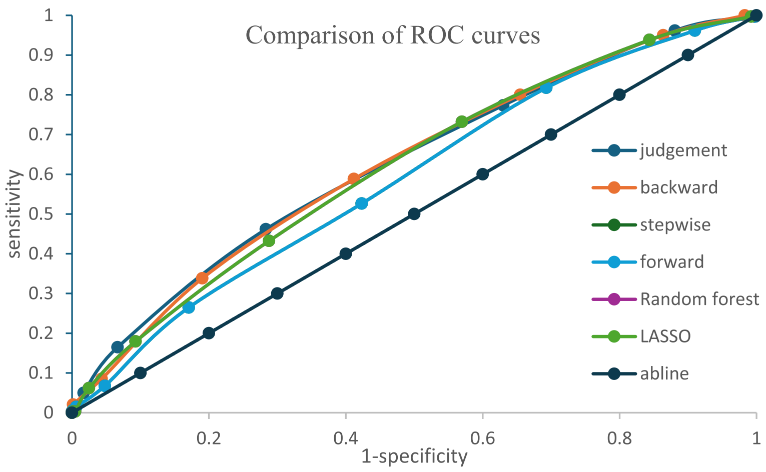

Fig. 4 Comparing discrimination of the models fitted using variables selected by the different methods

of residence of a child. Our model has shown better predictive performance for children residing in urban areas than those living in rural areas. The discriminative performance of our model was better than that of Hasegawa et al. [23] (AUC = 0.67, 95% CI: 0.64, 0.69), which used data from one health facility. The predictors of stunting identified by our best-performing model were consistent with existing knowledge of determinants of stunting, such as child demographics and wealth indicators.

Our study demonstrated that the best-performing predictive model was based on variables selected using the judgmental variable selection method. In contrast, other studies focusing on childhood stunting in Bangladesh and Rwanda have reported that the gradient-boosting classifier, followed by the random forest algorithm, achieved the lowest classification error in predicting stunting [22, 36, 37]. Some studies have also used the logistic and probit models for stunting prediction due to their common usage in predicting binary outcomes [38]. Other machine learning methods, such as Elastic net, regularized random forests and gradient-boosted feature selection, have been shown as the most effective methods for predicting stunting and other binary outcomes [39]. These help remove less important variables from large datasets with numerous variables for the derived models.

The performance of our final diagnostic risk model for child stunting was comparatively lower than that reported in other studies that developed risk prediction models for child stunting [40, 41]. According to the literature, numerous factors are associated with stunting; however, our study was limited to variables available in the dataset. A similar study by Haque et al. [42], which utilized a randomized cluster design to predict child stunting, also found that the predictive performance of their models, including logistic and probit regression and non-parametric decision trees, was suboptimal. They attributed this to the absence of crucial factors such as environmental and biological variables in their data [42]. Of note, our study calculated the metrics for assessing the performance of the predictive model using cut-points derived based on the SpEqualSe method implemented in the `OptimalCutpoins` package in R [31]. This method is based on the principles of balancing sensitivity and specificity with the assumption that the costs of false positives and false negatives are equal. The study acknowledges that other methods are used in choosing the probability cut points- such as Youden's Index (J). The Youden's Index (J) is defined as the sum of sensitivity and specificity minus one ($J_c = SE_c + SP_c - 1$). Xu et al., [43] used Youden's Index to determine the optimal value for predicting Acute Kidney Injury (AKI) in their model [44]. Both methods are data-driven and produce similar estimates. However, using the data-driven methods of choosing optimal cut points in studies with small sample sizes may identify inaccurate optimal cutoff points and overstate accuracy estimates [45]. Using pre-specified cutoff points when available would improve the validity of a classification model [46]. These pre-specified cutoff points are the ones that are predetermined by using previous studies, and they are always not available. Our study used a big sample size, which might have avoided what Bhandari et al., [45] observed. By using the model's optimal cutoff point, the research assesses the performance of a model at one point, but it is important to assess the performance of a model at different cutoff points.

The strength of our approach is that from a large range of candidate predictors from nationally representative data (MDHS), the research was able to identify a small set of key variables routinely measured at the primary healthcare level in many countries or that could be easily obtained. Even though factors affecting stunting that have been reported in the literature vary by many attributes such as type of study, region, sample size, and the ones mentioned above, considerable key findings have emerged that provide support for predictive variables that our model has identified.

The study had some limitations. Firstly, some variables were not captured in the MDHS, and some had missing values, as such, they were not used in developing our predictive model which may have affected the accuracy of our diagnostic predictive model. In addition, the study did not consider the clustering and weighting of the MDHS data, which may have affected the estimated probabilities of being stunted by not being representative. Although the majority of the variables employed were binary, maintaining a linear relationship between the log odds of stunting and the binary indicator points, it is important to note the potential for non-linear relationships with the log odds of stunting for certain categorical variables with more than two levels. Perhaps a major limitation of this study was the use of cross-sectional data, which restricts our ability to establish temporal relationships between predictors and stunting. The predictors (for instance socio-economic status) and the outcome (stunting) were collected at the same time, it was not possible to determine if the exposures preceded the outcome. Due to the cross-sectional nature of the data that we used, the model we have produced is a diagnostic prediction model to identify children their current stunting status. However, it could be used on children who are not yet stunted to calculate their risk of being stunted in the future. Future research with prospective cohort designs is needed to develop accurate prognostic models for stunting. The study did not consider least-angle regression (LARS) due to its complexity in implementation in R as another variable selection method for nonlinear models as another variable selection method for nonlinear models, which is considered a limitation of this study. Another limitation is that the study assumed a logistic link function provides a better predictive model. However, the study could have used a probit link function or complementary log–log link function. Lastly, the literature search was limited to PubMed and Google Scholar, while other databases such as Web of Science and Scopus could have been utilized and would have provided additional studies on child stunting in Sub-Saharan Africa.

Conclusion

The study has shown the viability of deriving child stunting diagnostic models that could be used to assess current child stunting status. Timely identification of children that are more likely to be stunted may help prevent the future impact of stunting and alleviate the disease burden in low-resource settings.

Appendix

Additional sources

[47–65]

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02283-6.

Supplementary Material 1.

Acknowledgements

We thank the DHS Program for providing access to the Malawi Demographic and Health Survey data set.

Authors' contributions

This paper is part of a Master of Science degree in Biostatistics thesis of JM in Department of Mathematical Sciences, University of Malawi under the supervision of SOM and HST. SOM conceived the ideas for the master's thesis, including the review and analysis plans, and oversaw the work undertaken in this paper. HST helped with checking the codes, data analyses and reviewed the work. JM wrote the original manuscript, which HST and SOM critically edited and revised. All authors have approved the final version for submission.

Funding

Not applicable.

Availability of data and materials

The data generated and analysed from this study are available upon request from the corresponding author. The original 2015/16 Malawi DHS data set is available on the DHS Program website: https://dhsprogram.com/.

Declarations

Ethics approval and consent to participate

The Malawi Health Research Committee determined that ethical approval was not deemed necessary in this study because the study used data from a research study already approved by an ethical research committee. The MDHS study was ethically approved by Malawi Health Research Committee, Institutional Review Board of ICF Macro, Centre for Disease and Control (CDC) in Atlanta, GA, USA and Prevention IRB. Informed consent was obtained from all eligible persons.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Mathematical Sciences, University of Malawi, Zomba, Malawi. [2]Department of Statistics, University of Pretoria, Pretoria, South Africa.

Received: 21 August 2023 Accepted: 12 July 2024
Published online: 08 August 2024

References

1. WHO. WHO Child growth standards based on length/height, weight and age. WHO Multicentre Growth Reference Study Group, Acta Paediatr Suppl, John and Wiley Sons Ltd; 2006.
2. W. UNICEF. Worldbank. 2019.
3. Shinsugi C, Matsumura M, Karama M, Tanaka J, Changoma M, Kaneko S. Factors associated with stunting among children according to the level of food insecurity in the household: a cross-sectional study in a rural community of Southeastern Kenya. BMC Public Health. 2015;15:441.
4. U. Nations. The 17 Goals. Available online: (accessed on 14 February 2024). 2019. Available: https://sdgs.un.org/goals. Accessed 14 Feb 2024.
5. Uwiringiyimana V, Veldkamp A, Ocke MC. Predictors of stunting with particular focus on complementary feeding practices: a cross-sectional study in the northern province of Rwanda. Nutrition. 2019;60:11–8.
6. Mtenda PAM. Association of low birth weight with undernutrition in preschool-aged children in Malawi. Nutrition. 2019;18:51.
7. Aguayo VM, Nair R, Badgaiyan N, Krishna V. Determinants of stunting and poor linear growth in children under 2 years of age in India: an in-depth analysis of Maharashtra's comprehensive nutrition survey. Matern Child Nutr. 2016;1:121–40.
8. Stewart CP, Iannotti L, Dewey KG, Michaelsen KF, Onyango AW. Contextualising complementary feeding in a broader framework for stunting prevention. Matern Child Nutr. 2013;9:27-45.
9. Hendriksen JM, Geersing GJ, Moons KG, de Groot JA. Diagnostic and prognostic prediction models. J Thromb Haemost. 2013;11:129–41.
10. Van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. J Clin Epidemiol. 2021;132:142–5.
11. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014;35(29):1925–31.
12. Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: an overview. J Thorac Dis. 2019;11(Suppl 4):S574–84.
13. Ranapurwala SI, Cavanaugh JE, Young T, Wu H, Asa CP, Ramirez MR. Public health application of predictive modeling: an example from farm vehicle crashes. Inj Epidemiol. 2019;6(31):1.
14. Grant SW, Collins GS, Nashef SAM. Statistical primer: developing and validating a risk prediction model. Eur J Cardiothorac Surg. 2018;54(2):203–8.
15. Janes H, Pepe MS, Gu W. Assessing the value of risk predictions by using risk stratification tables. Ann Intern Med. 2008;149(10):751–60.
16. Zakariya A, Algamal ZY. Variable selection in count data regression model based on firefly algorithm. Stat Optim Inf Comput Sch J. 2019;7(2):520–9.
17. Chowdhury MZ, Turin TC. Variable selection strategies and its importance in clinical prediction modelling. Fam Med Community Health. 2020;8(1):e000262.
18. Bag S, Gupta K, Dep S. A review and recommendations on variable selection methods in regression models for binary data. 2022. Available: https://api.semanticscholar.org/CorpusID:246016122. Accessed 15 Feb 2024.

19. Alkhateeb A, Agamal Z. Variable selection in gamma regressionmodel usingchaoticfireflyalgorithmwithapplicationin chemometrics. Electron J Appl Stat Anal. 2021;14(1):266–76.
20. Kipruto E, Sauerbrei W. Comparison of variable selection procedures and investigation of the role of shrinkage in linear regression-protocol of a simulation study in low-dimensional data. PLoS ONE. 2022;17(10):e0271240.
21. Hou D, Zhou W, Zhang Q, Zhang K, Fang J. A comparative study of different variable selection methods based on numerical simulation and empirical analysis. PeerJ Comput Sci. 2023;9:e1522.
22. Ndagijimana S, Kabano IH, Masabo E, Ntaganda JM. Prediction of stunting among under-5 children in Rwanda using machine learning techniques. J Prev Med Public Health. 2023;56(1):41–9.
23. Hasegawa J, Ito YM, Yamauchi T. Development of a screening tool to predict malnutrition among children under two years old in Zambia. Glob Health Action. 2017;10(1):1339981.
24. Ito YM, Yamauchi T, Hasegawa J. Development of a screening tool to predict malnutrition among children under two years old in Zambia. Glob Health Action. 2017;10(1):1339981.
25. Hanieh S, Braat S, Simpson JA, Ha TT, Tran TD, Tuan T, Fisher J, Biggs BA. The stunting tool for early prevention: development and external validation of a novel tool to predict the risk of stunting in children at 3 years of age. BMJ Glob Health. 2019;4:e001801.
26. Kursa MB, Jankowski A, Rudnicki WR. Boruta–a system for feature selection. Fundamenta Informaticae. 2010;101(4):271-85.
27. Farideh BK, Azra R, Fereidonn A, Farzad H, Ewout S, Davood K. A tutorial on variable selection for clinical prediction models: feature selection methods in data-mining could improve the results, J. J Clin Epidemiol. 2015;71:17.
28. Walter S, Tiemeier H. Variable selection: current practice in epidemiological studies. Eur J Epidemiol. 2009;24:733–6.
29. National Statistical Office. 2015–16 Malawi demographic & health survey. Rockville: NSO and ICF; 2017.
30. Akombi JB, Agho EK, Astell-Burt T, Hall JJ, Renzaho AM. Stunting and severe stunting among children under-5 years in Nigeria: a multilevel analysis. BMC Pediatr. 2017;17:1–6.
31. López-Ratón M, Rodríguez-Álvarez MX, Cadarso-Suárez C, Gude-Sampedro F. OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests. J Stat Softw. 2014;61(8):1–36.
32. Mukuku O, Mutombo AM, Kamona LK, Lubala TK, Mawaw PM, Aloni MN, Wembonyama SO, Luboya ON. Predictive model for the risk of severe acute malnutrition in children. J Nutr Metab. 2019;2019:1–7.
33. van den Brink DA, de Meij T, Brals D, Bandsma RHJ, Thitiri J, Ngari M, Mwalekwa L, de Boer NKH, Wicaksono A, Covington JA, van Rheenen PF, Voskuijl WP. Prediction of mortality in severe acute malnutrition in hospitalized children by faecal volatile organic compound analysis: proof of concept. Sci Rep. 2020;10:18785.
34. McDonald CM, Kupka R, Manji KP, Okuma J, Bosch RJ, Aboud S, Kisenge R, Spiegelman D, Fawzi WW, Duggan CP. Predictors of stunting, wasting and underweight among Tanzanian children born to HIV-infected women. Eur J Clin Nutr. 2012;66:1265–76.
35. Nshimyiryo A, Hedt-Gauthier B, Mutaganzwa C, Kirk CM, Beck K, Ndayisaba A, Mubiligi J, Kateera F, El-Khatib Z. Risk factors for stunting among children under five years: a cross-sectional population-based study in Rwanda using the 2015 demographic and health survey. BMC Public Health. 2019;19:1.
36. Khan JR, Tomal JH, Raheem E. Model and variable selection using machine learning methods with applications to childhood stunting in Bangladesh. Inform Health Soc Care. 2021;46(4):425–42.
37. Vu NU. Childhood stunting prediction in Bangladesh a machine learning approach (Doctoral dissertation, Tilburg University). 2022.
38. Ahmed SM, Brintz BJ, Pavlinac PB, Shahrin L, Huq S, Levine AC, Nelson EJ, Platts-Mills JA, Kotloff KL, Leung DT. Derivation and external validation of clinical prediction rules identifying children at risk of linear growth faltering. Elife. 2023;12:e78491.
39. Sanchez-Pinto LN, Venable LR, Fahrenbach J, Churpek MM. Comparison of variable selection methods for clinical predictive modelling. Int J Medi Inform. 2018;116:10–7.
40. Shen H, Zhao H, Jiang Y. Machine learning algorithms for predicting stunting among under-five children in Papua New Guinea. Children. 2023;10(10):1638.
41. Ndagijimana S, Kabano IH, Masebo E, Mtaganda JM. Prediction of stunting among under-5 children in Rwanda using machine learning techniques. J Prev Med Public Health. 2023;56(1):41.
42. Haque MA, Choudhury N, Wahid BZ, Ahmed ST, Farzana FD, Ali M, Naz F, Siddiqua TJ, Rahman SS, Faruque AS, Ahmed T. A predictive modelling approach to illustrate factors correlating with stunting among children aged 12–23 months: a cluster randomised pre-post study. BMJ. 2023;13(4):e067961.
43. Xu FB, Cheng H, Yue T, Ye N, Zhang HJ, Chen YP. Derivation and validation of a prediction score for acute kidney injury secondary to acute myocardial infarction in Chinese patients. BMC Nephrol. 2019;20:1-11.
44. Xu T, Wang J, Fang Y. A model-free estimation for the covariate-adjusted Youden index and its associated cut-point. Stat Med. 2014;33(28):4963-74. https://doi.org/10.1002/sim.6290.
45. Bhandari PM, Lewis B, Neupane D, Patten SB, Shrier I, Thombs BD. Data-driven methods distort optimal cutoffs and accurate estimates of depression screening tools: a simulation study using individual participant data. J Clin Epidemiol. 2021;137:137–47.
46. Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. J Clin Epidemiol. 2006;59(8):798–801.
47. Bukusuba J, Kaaya AN, Atukwase A. Predictors of stunting in children aged 6 to 59 months: a case-control study in Southwest Uganda. Food Nutr Bull. 2017;38(4):542–53.
48. Nkurunziza S, Meessen B, Van Geertruyden J-P, Korachais C. Determinants of stunting and severe stunting among Burundian children aged 6–23 months: evidence from a national cross-sectional household survey. BMC Pediatr. 2017;17(1):1–4.
49. Dake SK, Solomon FB, Bobe TM, Tekle HA, Tufa EG. Predictors of stunting among children 6–59 months of age in Sodo Zuria District, South Ethiopia: a community based cross-sectional study. BMC Nutr. 2019;5:1-7.
50. Semali A. Prevalence and determinants of stunting in under-five children in central Tanzania: remaining threats to achieving Millennium Development Goal 4. BMC Public Health. 2015;15:1–6.
51. Garcia Cruz LM, González Azpeitia G, Reyes Súarez D, Santana Rodríguez A, Loro Ferrer JF. Factors associated with stunting among children aged 0 to 59 months from the Central Region of Mozambique. Nutrients. 2017;9(5):491.
52. Habimana S, Biracyaza E. Risk factors of stunting among children under 5 years of age in the Eastern and Western Provinces of Rwanda: analysis of Rwanda demographic and health survey 2014/2015. Pediatr Health Med Ther. 2019;10:115–30.
53. Haile D, Azage M, Mola T, Rainey R. Exploring spatial variations and factors associated with childhood stunting in Ethiopia: a spatial and multilevel analysis. BMC Pediatr. 2016;16:1–4.
54. Kofi JO. Predictors of childhood stunting in Ghana: a cross-sectional survey of the association between stunting among children under age five and maternal bio-demographic and socioeconomic characteristics in Ghana 2014 (Dissertation). 2018.
55. Kismul H, Acharya P, Mapatano MA, Hatløy A. Determinants of childhood stunting in the Democratic Republic of Congo: further analysis of demographic and health survey 2013–14. BMC Public Health. 2018;18:1–4.
56. Berhe K, Seid O, Gebremariam Y, Berhe A, Etsay N. Risk factors of stunting (chronic undernutrition) of children aged 6 to 24 months in Mekelle City, Tigray Region, North Ethiopia: an unmatched case-control study. PLoS ONE. 2019;14(6):e0217736.
57. Fikadu T, Assegid S, Dube L. Factors associated with stunting among children of age 24 to 59 months in Meskan district, Gurage Zone, South Ethiopia: a case-control study. BMC Public Health. 2014;14:800.
58. Sema B, Azage M, Tirfie M. Childhood stunting and associated factors among irrigation and non-irrigation user northwest, Ethiopia: a comparative cross-sectional study. Ital J Pediatr. 2021;47:102.
59. Tariku A, Biks GA, Derso T, Wassie MM, Abebe SM. Stunting and its determinant factors among children aged 6–59 months in Ethiopia. Ital J Pediatr. 2017;43(1):112.
60. Gebru KF, Haileselassie WM, Temesgen AH, Seid AO, Mulugeta A. Determinants of stunting among under-five children in Ethiopia: a multilevel mixed-effects analysis of 2016 Ethiopian demographic and health survey data. BMC Pediatr. 2019;19:176.
61. Keino S, Plasqui G, Ettyang G, Van Den Borne B. Determinants of stunting and overweight among young children and adolescents in sub-Saharan Africa. Food Nutr Bull. 2014;35(2):167–78.

62. Afolabi F, Palamuleni ME. Determinants of stunting among under-five children in Malawi. African Population Conference. 2019.
63. Nsereko E, Mukabutera A, Iyakaremye D, Umwungerimwiza YD, Mbarushimana V, Nzayirambaho M. Early feeding practices and stunting in Rwandan children: a cross-sectional study from the 2010 Rwanda demographic and health survey. Pan Afr Med J. 2018;29:1–6.
64. Krasevec J, An X, Kumapley R, Bégin F, Frongillo EA. Diet quality and risk of stunting among infants and young children in low- and middle-income countries. Matern Child Nutr. 2017;13:20.
65. Ayelign A, Zerfu T. Household, dietary and healthcare factors predicting childhood stunting in Ethiopia. Heliyon. 2021;7(4):e06733.

Publisher's Note