Journal of Health, Population and Nutrition

RESEARCH

Open Access

Development and validation of prediction model for vitamin D deficiency in Chinese college students (a dynamic online nomogram predicting vitamin D deficiency for Chinese college students)

Yingyi Luo[1], Chunbo Qu[2], Guyanan Li[2], Qiannan Di[2], Shangzhen Ding[1], Ruoyou Jiang[1], Ruotong Wang[1], Siyuan Wang[1] and Lixin Na[2*]

Abstract

Objective This study aims to develop a model for predicting vitamin D deficiency in Chinese college students using easily accessible clinical characteristics.

Methods Data were derived from a cross-section study of the Vitamin D status in Chinese college students in September, 2020. Totally 1,667 freshmen from 26 provinces, autonomous districts or municipalities were analyzed. A LASSO regression model was used to select predictors and the significant factors were used to construct the logistic regression model expression and the nomogram. The prediction model was subjected to100 bootstrap resamples for internal validation to assess its predictive accuracy. Calibration and discrimination were used to assess the performance of the model. A dynamic online nomogram was conducted to make the model easy to use. The clinical use was evaluated by a decision curve analysis.

Results Gender, region of original residence, milk and yogurt intake, puffed foods intake, outdoor activity duration, UV protection index and "taken calcium or vitamin D supplements within 3 months" were identified as significant predictors of vitamin D deficiency among Chinese college students. The model demonstrated good calibration with a 100 bootstraps analysis. The C-index was 0.677 and the bias-adjusted C-index was 0.668 in internal validation with 100 bootstrap resamples. The decision curve analysis showed a threshold probability between 0.5 and 0.8, using the model added more benefit than considering all patients are deficient or not deficient.

Conclusions The performance of this vitamin D deficiency prediction model is commendable, and the dynamic online nomogram was proved to be a user-friendly screening tool for identifying high-risk subjects among Chinese college students. However, external validation is imperative to ensure the model's generalizability.

Keywords Vitamin D, Vitamin D deficiency, Prediction model, Dynamic nomogram, Lasso regression, Chinese

*Correspondence:
Lixin Na
nalixin2003@163.com

[1]Medical Technology College, Shanghai University of Medicine and Health Sciences, Shanghai, China
[2]Public Health College, Shanghai University of Medicine and Health Sciences, Shanghai, China

Introduction

Vitamin D is a fat-soluble vitamin which is not only critical for bone health but also associated with non-skeletal diseases such as cardiovascular disease [1], cancer [2], metabolic syndrome (MetS) [3], infection [4] and mortality [5]. However, vitamin D deficiency is still a globally public health problem not only in children [6], pregnant women [7] and old people [3] but also in general population [8].

In China, the prevalence of vitamin D deficiency is around 40–75% at the cut-off of 20 ng/mL in people with different age and characteristics [3, 6, 9, 10]. Additionally, our previous observational study reported the prevalence of vitamin D deficiency (< 20 ng/mL) and insufficiency (20 ~ < 30 ng/mL) were 64.4% and 30.2% respectively in Chinese college students who came from 26 provinces, autonomous districts or municipalities of China. Therefore, the monitoring of vitamin D levels and giving timely intervention in Chinese college students are very necessary to prevent the skeletal and non-skeletal diseases related to vitamin D deficiency [11].

Although vitamin D status can be evaluated by serum 25(OH)D concentration, its measurement is costly and not easily applicable to the healthy subjects. Therefore, a screening strategy that could predict whether an individual is likely to be at risk of vitamin D deficiency would be of much epidemiological value. Over the past years, various models have been developed and validated for the prediction of vitamin D deficiency [12–16], whereas the predictor variables in most of the models were limited for the information about diet, UV protection and physical activity were incomplete and most of the models were developed for patient in the hospital or special population and in different countries. There was still no vitamin D prediction model for Chinese population.

The purpose of this study was to develop a validated model for predicting vitamin D deficiency in Chinese college students using the demographic characteristics, lifestyle information about the diet, physical activity, and UV protection which could be easily identified. A simple and useful dynamic online nomogram would be conducted to predict vitamin D deficiency for Chinese college students. Therefore, the model would help us find participants with high risk of vitamin D deficiency and reduce the requests for serum vitamin D measurements and unnecessary supplementation.

Methods

Study design and data source

The data of this study were obtained from a cross-section study of the Vitamin D status in Chinese college students. The participants of this study were freshmen of Shanghai University of Medicine and Health Sciences, started the school in 2020 September. They came from twenty-six different provinces of China. A total of 3,385 students with completed physical examination and serum 25(OH) D_3 determination were recruited into the study in September, 2020. And 1,925 students completed the questionnaires on demographic information and lifestyles at the same time of the physical examination and serum 25(OH)D_3 determination. After excluding 257 students who had extreme values for total energy intake (< 500 or > 4500 kcal/d) and 1 case with obvious logic error in data, 1,667 subjects were eventually included in our analysis.

This study protocol was approved by the Ethics Committee of the Ethics Committee of Shanghai University of Medicine and Health Sciences. Approval Number:2019-CSHS-SUMHS-05-230119197901252329. A written informed consent was provided by all participants. The model was developed and validated in accordance with the recommendations established in the Transparency Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD).

Outcomes

The outcome variable was vitamin D deficiency which was defined as a 25(OH)D_3 below 20 ng/ml (50 nmol/L). Vitamin D deficiency is defined as a serum 25(OH) D_3 < 20 ng/mL according to the criteria of the guidelines on vitamin D deficiency from Endocrine Society Clinical Practice and the Consensus of the Chinese Society of Osteoporosis and Bone Mineral Research [17, 18]. Fasting venous blood samples were collected in September and serum 25(OH)D_3 concentration was measured using a high performance liquid chromatograph (Agilent 1100; Agilent Technologies Inc., Santa Clara, CA, USA) and a mass spectrometer (API4000Q trap; AB SCIEXLLC., Redwood City, CA, USA). The lower limits of 25(OH) D_3 for detection was 1.6ng/mL. The test sensitivity was assessed with the inter-batch coefficient of variation (CV) of 5.85% and between batches CV of 6.18%.

Predictor variables

We selected 30 variables including the demographic information, lifestyle information about the diet, physical activity, and UV protection of the participants as candidate predictors on the basis of pathogenesis of vitamin D deficiency and the possible influencing factors found in previous studies [17, 19–21]. An online questionnaire was used to collect the information. The questionnaires were provided in class units and were explained by trained class instructors before the formal investigation. The same IP address can only be answered once.

The information of daily food consumption was collected through an online simple food frequency questionnaire (FFQ) which was designed by Gao J and the reliability and validity were reasonable [22]. A total of 21 food groups that were common consumed in China were included in our FFQ questionnaire. For each food group, the participants were asked how frequently they consumed in the past 12 months, followed by the amount of consumption with five common amounts to choose. The pictures of the standard amount of food were presented when the participants answered questions online. The daily food consumption was calculated by multiplying food frequency by food amount. Before the survey started, 100 participants were recruited to complete our simple FFQ twice with a two-week interval and also finished FFQ146, which included 146 food items produced by Chinese Center for Disease Control and Prevention. The reliability coefficients of food intakes between two simple FFQs ranged from 0.56 to0.87 ($P<0.01$). The reliability coefficients of food intakes between the simple FFQ and FFQ146 ranged from 0.31 to 0.72 ($P<0.01$). The correlation coefficients of food intakes between simple FFQ and 3d 24 h dietary records ranged from 0.37 to 0.65($P<0.01$). To facilitate practical application of the model, we classified each of the daily food consumption of the 21 food groups into two groups according to the integer units which close to median or usual amount of the dietary intake, as shown in Table 1.

The information of physical activities which including vigorous physical activity, moderate physical activity and outdoor activity was also got from the online questionnaire. For each physical activity, the participants were asked the average frequency and duration per-time every week in the past 12 months, the average duration of physical activities per day was calculated as frequency per week multiplied by duration per-time and then divided by 7 days. To facilitate practical application of the model, we classified each kind of physical activity into three groups according to the activity time per day.

From the online questionnaire we also got the information of the UV protection of the participants. The UV protection measures including sun hat, sun umbrella, sun-protective clothing and sunscreen was asked. Each question had 5 choices including almost no use, occasional use in summer, frequent use in summer, always use in summer and use in all seasons. For each of the four questions, the choice of almost no use, occasional use in summer, frequent use in summer, always use in summer and use in all seasons was calculated as 1 point, 2 point,3point, 4 point, and 5 point, respectively. The UV protection index was calculated as the total scores of the four questions. Therefore, the UV protection index ranged from 4 to 20, the higher the score, the better the sun protection.

The information of height and weight was obtained through the physical examination of the students, and body mass index (BMI, kg/m^2) was calculated. There were 60 missing data in the BMI variable and multiple imputation with chained equations was used to replace the missing data. In the imputation model, all candidate predictor variables, along with the outcome indicator, were included. Five imputations were carried out and the imputation methods consisted of predictive mean matching for continuous predictors and logistic regression for binary predictors. Rubin's rules were used to combine the results across the imputed datasets [23]. Then BMI was classified into two groups with the cut point of 24 kg/m^2 according to the Chinese criteria for overweight and obesity [24, 25].

The information of the calcium or vitamin D supplements intake within 3 months was also obtained from the online questionnaire.

Sample size consideration

The number of events per variable (EPV) can be used to guide sample size and when developing prediction models for binary outcome, a rule of thumb for the required sample size is to ensure at least 10 events for each predictor parameter [26]. The number of events is the number of observations in the smallest outcome category of the binary outcome. In our study, the number of events was 542 which is the number of participants with serum $25(OH)D_3 \geq 20ng/mL$. The number of candidate predictors is 34 which including 4 dumb variables. Therefore, the EPV in our study was 15.94 and the sample was adequate for the development of prediction model.

Statistical analysis

The category variables were expressed as frequency (percentage) and continuous variables as mean ± SD. Restricted cubic splines were applied to explore the relationship between age, UV protection index and outcome variable. A least absolute shrinkage and selection operator (LASSO) regression model was performed to select predictors of vitamin D deficiency in Chinese college students and the significant factors were used to construct the logistic regression model expression and the prediction nomogram. The performance of the prediction model was assessed by discrimination and calibration. The discriminative ability of the model was determined by the area under the receiver operating characteristic (ROC) curve, which ranges from 0.5 (no discrimination) to 1 (perfect discrimination). The calibration of the prediction model was performed by a visual calibration plot comparing the predicted and actual probability of having a vitamin D below the predefined threshold. In addition, the prediction model was subjected to100 bootstrap resamples for internal validation to assess its predictive accuracy. A decision curve analysis was also built in order to determine a net-benefit threshold of prediction. An interactive web-based dynamic nomogram application was built with Shiny, version 1.7.4. All computations were conducted in the R environment, version R version 4.2.3

Table 1 Characteristics of the participants ($n = 1667$)

Factor		Category	Proportion in category N (%)	Serum 25(OH) $D_3 < 20$ng/mL N (%)	*P* value
Demographic information	Age	16～17	31(1.9)	23 (74.2)	0.178
		18～19	1492(89.5)	1014 (68.0)	
		20～26	144(8.6)	88 (61.1)	
	Gender	Male	383(23.0)	184 (48.0)	< 0.001
		Female	1284(77.0)	941 (73.3)	
	Region of original residence	Countryside	398(23.9)	237 (59.5)	< 0.001
		Town	197(11.8)	122 (61.9)	
		City	1072(64.3)	766 (71.5)	
	BMI categories*	>=24	311(19.4)	186 (59.8)	0.002
		< 24	1296(80.6)	895 (69.1)	
Food items/day	Refined rice, steamed	≥ 200 g/d	1099(65.9)	722 (65.7)	0.030
		< 200 g/d	568(34.1)	403 (71.0)	
	Refined rice, porridge	≥ 25 g/d	787(47.2)	509 (64.7)	0.021
		< 25 g/d	880(52.8)	616 (70.0)	
	Refined wheat products	≥ 50 g/d	960(57.6)	633 (65.9)	0.116
		< 50 g/d	707(42.4)	492 (69.6)	
	Desserts	≥ 25 g/d	741(44.5)	511 (69.0)	0.250
		< 25 g/d	926(55.5)	614 (66.3)	
	Whole grains	≥ 10 g/d	1049(62.9)	712 (67.9)	0.660
		< 10 g/d	618(37.1)	413 (66.8)	
	Tubers	≥ 25 g/d	725(43.5)	480 (66.2)	0.328
		< 25 g/d	942(56.5)	645 (68.5)	
	Milk and yogurt	≥ 200 ml/d	867(52.0)	558 (64.4)	0.005
		< 200 ml/d	800(48.0)	567 (70.9)	
	Eggs	≥ 50 g/d	587(35.2)	379 (64.6)	0.061
		< 50 g/d	1080(64.8)	746 (69.1)	
	Livestock meat	≥ 100 g/d	847(50.8)	561 (66.2)	0.267
		< 100 g/d	820(49.2)	564 (68.8)	
	Poultry meat	≥ 50 g/d	844(50.6)	557 (66.0)	0.188
		< 50 g/d	823(49.4)	568 (69.0)	
	River fish, shrimp and crab	≥ 10 g/d	1035(62.1)	701 (67.7)	0.786
		< 10 g/d	632(37.9)	424 (47.1)	
	Sea fish, shrimp and crab	≥ 10 g/d	796(47.8)	550 (69.1)	0.180
		< 10 g/d	871(52.2)	575 (66.0)	
	Soybeans and Soy products	≥ 25 g/d	790(47.4)	527 (66.7)	0.520
		< 25 g/d	877(52.6)	598 (68.2)	
	Nuts and seeds	≥ 5 g/d	911(54.6)	617 (67.7)	0.817
		< 5 g/d	756(45.4)	508 (67.2)	
	Dark vegetables	≥ 100 g/d	786(47.2)	508 (64.6)	0.019
		< 100 g/d	881(52.8)	617 (70.0)	
	Light color vegetables	≥ 50 g/d	957(57.4)	624 (65.2)	0.021
		< 50 g/d	710(42.6)	501 (70.6)	
	Mushrooms	≥ 12 g/d	897(53.8)	593 (66.1)	0.195
		< 12 g/d	770(46.2)	532 (69.1)	
	Fruits	≥ 100 g/d	835(50.1)	548 (65.6)	0.105
		< 100 g/d	832(49.9)	577 (69.4)	
	Puffed foods	≥ 10 g/d	940(56.5)	672 (71.5)	< 0.001
		< 10 g/d	727(43.6)	453 (62.3)	
	Candy and chocolate	≥ 6 g/d	755(45.3)	528 (69.9)	0.052
		< 6 g/d	912(54.7)	597 (65.5)	

Table 1 (continued)

Factor		Category	Proportion in category	Serum 25(OH) $D_3 < 20$ng/mL	P value
			N (%)	N (%)	
	Sugary drinks	≥ 100 ml/d	757(45.4)	516 (68.2)	0.590
		< 100 ml/d	910(54.6)	609 (66.9)	
	Taken calcium or vitamin D supplements within 3 months	Yes	132(7.9)	78 (59.1)	0.032
		no	1535(92.1)	1047 (68.2)	
Physical activity/day	Vigorous physical activity	≥ 2 h/d	476(28.6)	291 (61.1)	0.002
		1 ~ < 2 h/d	543(32.6)	382 (70.3)	
		< 1 h/d	648(38.9)	452 (69.8)	
	Moderate physical activity	≥ 2 h/d	267(16.0)	179 (67.0)	0.932
		1 ~ < 2 h/d	406(24.4)	277 (68.2)	
		< 1 h/d	994(59.6)	669 (67.3)	
	Outdoor activity	≥ 1 h/d	623(37.4)	380 (61.0)	< 0.001
		0.5 ~ < 1 h/d	527(31.6)	376 (71.3)	
		< 0.5 h/d	517(31.0)	369 (71.4)	
UV protection	UV protection index	4 ~ 8	942(56.5)	581 (61.7)	< 0.001
		9 ~ 12	468(28.1)	343 (73.3)	
		13 ~ 16	200(12.0)	156 (78.0)	
		17 ~ 20	57(3.4)	45 (78.9)	

*60 data missing in variable BMI categories

(2023-03-15). Results with $P < 0.05$ were considered statistically significant.

Results

Study population and characteristics

The data of this study were obtained from a cross-section study of the Vitamin D status in Chinese college students. The flow of participants enrollment was shown in Fig. 1. A total of 1667 participants with completed data were included in this analysis, with 383 males (23.0%) and 1284 females (77.0%). The participants came from twenty-six provinces, autonomous districts or municipalities of China with an average age of 18.55 ± 0.90 years. Levels of serum 25(OH)D_3 of the participants ranged from 5.20 to 64.83 ng/mL with the mean of 18.06 ± 6.34 ng/mL. There were 1125 (67.5%) participants had serum 25(OH)$D_3 < 20$ng/mL. The characteristics of the participants including the demographic information, lifestyle information about the diet, physical activity, and UV protection were shown in Table 1.

Development of prediction model

Thirty candidate variables including the demographic information, lifestyle information about the diet, physical activity, and UV protection of the participants were included in the original model. Restricted cubic splines are a flexible statistical tool used to model potential non-linear relationships between continuous predictors. After applying Restricted cubic splines, the shape of the relationships between age, UV protection index, and vitamin D deficiency was assessed. It was indicated that the relationships were approximately linear within the observed ranges of these variables. Therefore, age and UV protection index were entered into the original model as continuous variables for the selecting of predictor variables. Since we considered 30 variables as our candidate predictors, in order to reduce the issue of multicollinearity among independent variables, LASSO regression was applied to filter the variables for the final prediction model. A ten-fold cross-validation was applied to select the penalty term, lambda(λ). To make the model more concise, we selected the largest lambda value within one standard error of the minimum error (lambda.1se), which corresponds to the dashed line on the right side in Fig. 2A. log(λ) = – 3.558202 (λ = 0.02848985) when the error of the most regularized and parsimonious model is acceptably minimized. Six variables including gender, region of original residence, milk and yogurt intake, puffed foods intake, outdoor activity duration and UV protection index were remained in the final model. A cross-validated error plot of the LASSO regression model is shown in Fig. 2A and a coefficient profile plot was shown in Fig. 2B. Additionally, "taken calcium or vitamin D supplements" had always been considered as a way for preventing vitamin D deficiency, therefore the variable of "taken calcium or vitamin D supplements within 3 months" was included in the final model according to expert advice. Finally, a multivariable logistic regression model including seven predictive variables was established to predict vitamin D deficiency in Chinese college students. The final predictors with their β coefficients and *OR* value for vitamin D deficiency were shown in Table 2.

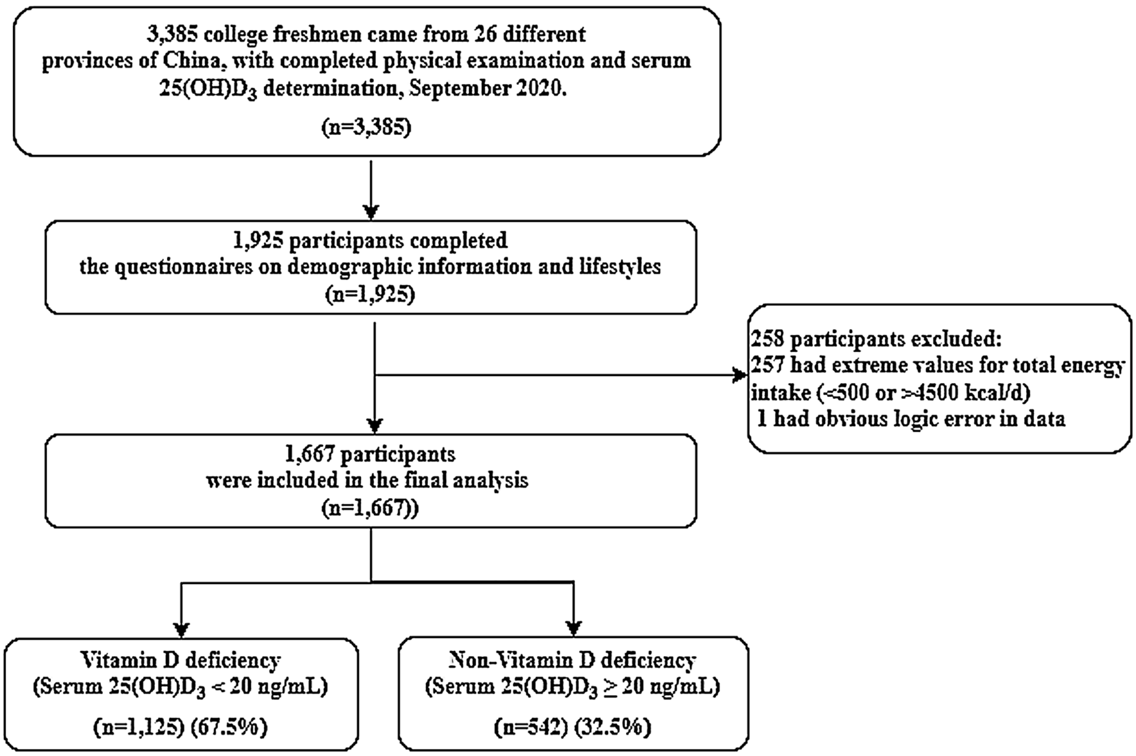

Fig. 1 Flowchart for participants enrollment

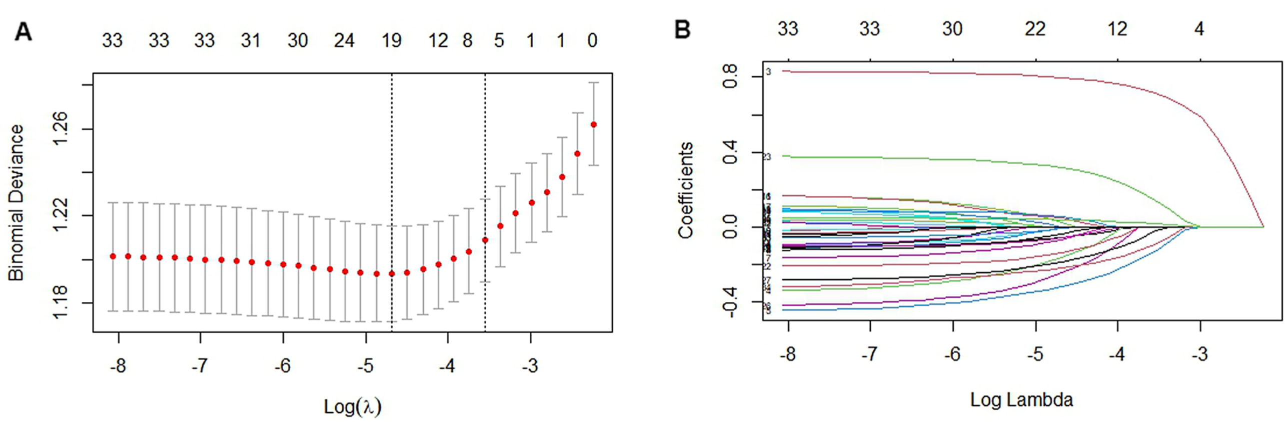

Fig. 2 **A** cross-validated error plot of the LASSO regression model. **B** Lasso coefficient profile plot

The final model was then developed as a simple-to-use nomogram as shown in Fig. 3A and available online (https://sumhs-lyy.shinyapps.io/vitd/) as screenshotted in Fig. 3B.

Performance of the prediction model

The internal validation and calibration of the prediction model was performed with a 100 bootstraps analysis. The C-index for the prediction model was 0.677 and the bias-adjusted C-index was 0.668 with a 100 bootstraps analysis. The correspondent ROC curve was shown in Fig. 4. The Area Under the Curve (AUC) was 0.677 (95% CI 0.6497–0.7044). The calibration plot of the prediction model was shown in Fig. 5 and it demonstrated a good correlation between observed and predicted vitamin D deficiency with a mean absolute error of 0.01.

Decision curve analysis

The decision curve analysis for the prediction model was shown in Fig. 6. Decision curve analysis calculates a clinical net benefit for a prediction model in comparison to default strategies of treating all or no patients. Net benefit is calculated across a range of threshold probabilities, defined as the minimum probability of disease at which further intervention would be warranted. The decision curve of this model indicated that utilizing the model to

Table 2 Predictors with their Β coefficients and OR values in the prediction model of vitamin D deficiency for Chinese college students

Predictors		β-value	OR	95%CI	*P* value
Gender	Male	Ref	Ref	Ref	Ref
	Female	0.846	2.329	1.772 ~ 3.065	< 0.001
Region of original residence	City	Ref	Ref	Ref	Ref
	Town	-0.329	0.720	0.518 ~ 1.006	0.052
	Countryside	-0.457	0.633	0.491 ~ 0.817	< 0.001
Milk and yogurt	< 200 ml	Ref	Ref	Ref	Ref
	≥ 200 ml	-0.315	0.730	0.587 ~ 0.906	0.005
Puffed foods	< 10 g	Ref	Ref	Ref	Ref
	≥ 10 g	0.434	1.543	1.244 ~ 1.915	< 0.001
Taken calcium or vitamin D supplements within 3 months	No	Ref	Ref	Ref	Ref
	Yes	-0.400	0.670	0.458 ~ 0.988	0.041
Outdoor activity	< 0.5 h/day	Ref	Ref	Ref	Ref
	0.5 ~ < 1 h/day	-0.086	0.918	0.693 ~ 1.214	0.547
	≥ 1 h/day	-0.351	0.704	0.541 ~ 0.915	0.009
UV protection index	By 1 degree	0.047	1.048	1.014 ~ 1.084	0.006

predict vitamin D deficiency and subsequently implementing an intervention or vitamin D supplementation provided greater net benefit compared to any default strategy, whether assuming all subjects are deficient and receiving intervention or assuming all subjects are not deficient and receiving no intervention, when the threshold probability of vitamin D deficiency ranged between 0.5 and 0.8.

Sensitivity analysis

In order to explore the reliability of the variable selection of the prediction model, univariate logistic regression analysis and multivariate logistic regression analysis were also applied to filter the variables for the final prediction model. Firstly, univariate logistic regression analysis was performed to identify risk factors for vitamin D deficiency from the thirty candidate variables. Secondly, risk factors with P values < 0.10 in the univariate analysis were included in a multivariate analysis. Multivariate logistic regression analysis was performed to identify independent risk factors, and a stepwise method was used to identify the useful combination of factors that could most precisely predict Vitamin D deficiency in Chinese college students. Finally, seven variables including gender, region of original residence, milk and yogurt intake, puffed foods intake, outdoor activity duration, UV protection index and "taken calcium or vitamin D supplements within 3 months" were remained in the final model which was consistent with the result of the LASSO regression and expert advice as shown in Table 3.

Discussion

In this study, we developed an on-line prediction model for vitamin D deficiency based on easy-to-assess individual characteristics of the participants in Chinese college students. The on-line prediction model used information on individual characteristics associated with vitamin D status including gender, diet, physical activity and UV protection that could feasibly be gathered by questionnaire survey. Our prediction model showed good discriminatory ability, calibration. It could be used as a tool to screen college students for vitamin D deficiency, thus avoiding unnecessary blood tests and supplementation.

The sample of this study was 1667 and they came from twenty-six different provinces of China. The prevalence of vitamin D deficiency was 67.5% according to the criteria of serum $25(OH)D_3 < 20$ng/mL. It means in China, the vitamin D statuses of college students were not better than that of children, pregnant women and old people [27–29]. Therefore, there is necessary to focus on screening and intervention for vitamin D deficiency in this population group.

To our knowledge, this is the first vitamin D deficiency prediction model to be developed and validated based on Chinese population. In the previous, there were also some vitamin D deficiency prediction model developed in Europe, Japan and New Zealand [13, 14, 30, 31]. Due to the differences of geographic latitude and the lifestyles of diet and UV protection, these models were not very suitable to Chinese. Furthermore, most of the vitamin D deficiency prediction model based on special population like old people or clinical patients [12, 14–16], therefore the character variables were not very suitable to healthy young population. Viprey M et al. developed a prediction model of hypovitaminosis D from 2488 general adult, but the participants were outpatients seen at the hospital who may be not representative for the real general population [32]. The study of Horton-French K et al. showed the prevalence and predictors of vitamin D deficiency in Australian adolescents and young adults based

Fig. 3 **A** Nomogram for the prediction of vitamin D deficiency as defined by a level below 20 ng/mL. **B**. The online dynamic nomogram accessible at https://sumhs-lyy.shinyapps.io/vitd/

on Australian Health Survey [33], but they didn't develop and validate the prediction model.

In our model, seven independent and significant predictive variables including gender, region of original residence, milk and yogurt intake, puffed foods intake, outdoor activity duration, UV protection index and "taken calcium or vitamin D supplements within 3 months" were selected according to LASSO regression model and expert advice. Sensitivity analysis indicated this result was consistent with the result of multivariate logistic regression analysis with stepwise method.

These characteristics had also been identified as predictors of vitamin D deficiency in previous analyses [13, 14, 31], and therefore our findings were consistent with the reported studies. The highlight of our study was the research approach regarding UV protection. We used UV protection index which was an accumulated scores calculated from 4 five-tier-questions about UV protection measures to comprehensively consider the cumulative effects of various UV protection measures rather than the effect of one of them. Another highlight of our study was that we considered 21 candidate variables about dietary information, and then "milk and yogurt intake", "puffed foods intake" were remained in the final prediction model according to LASSO regression. "Milk and yogurt intake ≥ 200 ml/day" was a protective factor for

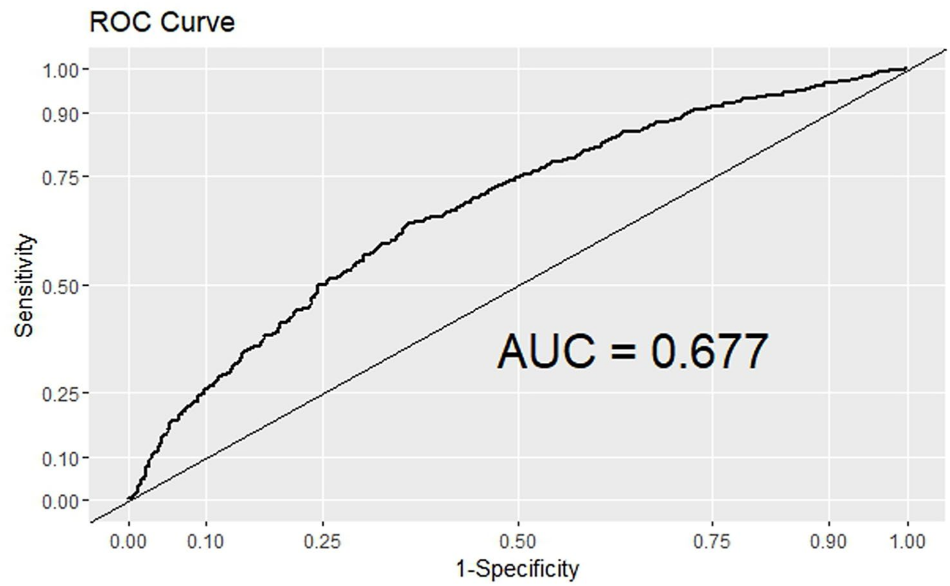

Fig. 4 ROC curve of the vitamin D deficiency model showing an AUC of 0.677

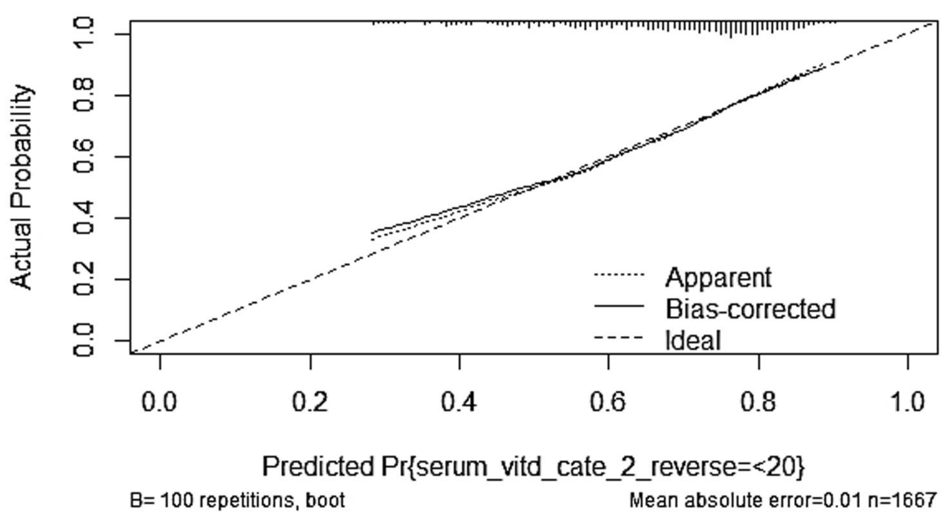

Fig. 5 Calibration plot of the vitamin D deficiency model with mean absolute error of 0.01 with a 100 repetitions bootstrap

vitamin D deficiency. The result was consistent with the US National Health and Nutrition Examination Survey (NHANES 2001–2010) which demonstrated a significant association between milk consumption and serum vitamin D status in US population [34]. The reason may be that some of the milk and yogurt were fortified with vitamin D. Another notable finding was the association of vitamin D deficiency with "puffed foods intake ≥ 10 g/day" which was not involved in other related research. Since puffed foods were popular snacks for Chinese students, we included it in our FFQ questionnaire, and the reason for this association was unclear. Further epidemiological studies on the relationship between puffed foods and vitamin D deficiency should be developed in the future.

The performance of the vitamin D deficiency prediction model was assessed by discrimination and calibration. The internal validation and calibration of the prediction model was performed with a 100 bootstraps analysis and the AUC was 0.677. The AUC in other studies ranged from 0.64 to 0.78 [13, 15, 31, 32] which was similar to our study. The calibration plot showed that the prediction model had a good correlation between observed and predicted vitamin D deficiency.

In previous studies, most researchers had only evaluate the performance of the predictive models and had not visualised the probabilities of vitamin D deficiency [31,

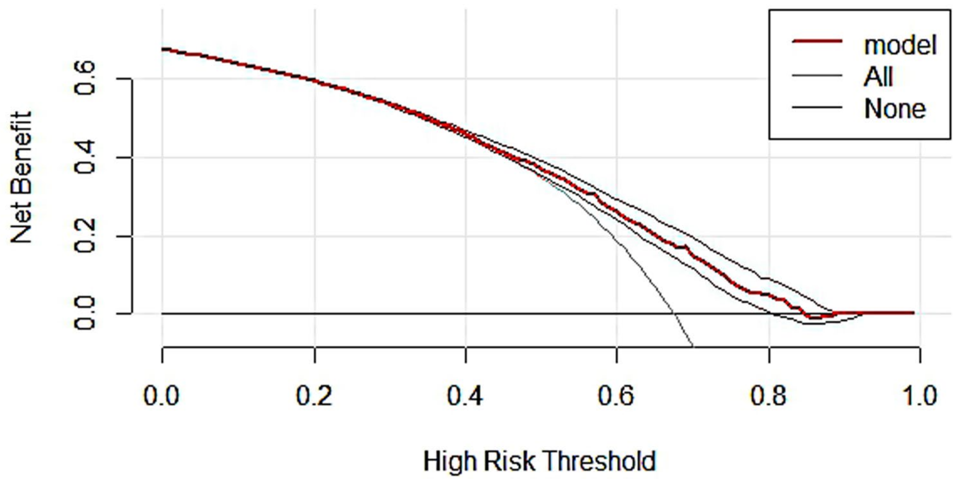

Fig. 6 Decision curve analysis of the vitamin D deficiency model

32]. Some researchers used risk scores which summed from each individual predictive item to estimate the probability of vitamin D deficiency [13]. The highlight of our study was that we made a dynamic online nomogram with a user-friendly digital interface to predict vitamin D deficiency for objective population which making it a simple-to-use tool to screen subjects for vitamin D deficiency. Furthermore, we also studied the clinical usefulness of the vitamin D deficiency prediction model based on a decision curve analysis. The decision curve was also used in the study of George Bou Kheir to estimate the clinical net benefit for a prediction model of severe vitamin D deficiency at ICU admission [15]. The decision curve of our model showed that if the threshold probability of the Vitamin D deficiency was between 0.5 and 0.8, using this model to predict vitamin D deficiency and therefore an intervention or a vitamin D supplementation added more net benefit than any default strategy It means that the application of the prediction model can be of clinical value for decision-making in the prevention of vitamin D deficiency in Chinese college students.

The strength of our study is that the participants were freshmen who came from 26 provinces, autonomous districts or municipalities of China which make it a relatively representative sample of Chinese college students. Another strength is that we considered 30 variables as our candidate predictors and then we used LASSO regression to select the prominent independent variables influencing the vitamin D status which reduced the issue of multicollinearity among independent variables. Third, the predictors in our model could be easily and rapidly obtained through simple questions and the dynamic online nomogram with a user-friendly digital interface made it a simple-to-use tool to screen subjects for vitamin D deficiency. We have to mention some limitations of our study. First, the data of our study was collected from a cross-section study which was done in the late of summer, therefore the effect of seasonal factors on vitamin D status was not included in the model. Future studies should employ longitudinal or cohort designs to validate the model and capture dynamic changes in vitamin D status over time. Second, our subjects were restricted to Chinese college students, therefore the results may not be generalizable to population with different age groups and other characteristics. Third, we should perform the external validation for the model in another population of college students in the further.

Conclusion

In conclusion, a prediction model for vitamin D deficiency in Chinese college students was developed. The model showed good discriminatory ability and calibration. The seven predictors in the model were easy to obtain and the dynamic online nomogram made it a screening tool that can be easily used in practice therefore, avoiding unnecessary blood tests and vitamin supplementation. Further research is needed to perform external validation of the model and find the most appropriate way to use this model in the decision-making process for blood tests or prescribing vitamin D supplementation.

Table 3 Predictors in the prediction model selected by univariate logistic regression analysis and multivariate logistic regression analysis

Factors			**Univariate analysis**			**Multivariate analysis**		
			OR	**95%CI**	***P* value**	**OR**	**95%CI**	***P* value**
Basic characteristics	Age	By 1 year	0.879	0.787 ~ 0.983	0.023			
	Gender	Female / Male	2.967	2.345 ~ 3.757	< 0.001	2.329	1.772 ~ 3.065	< 0.001
	Region of original residence	Town / City	0.650	0.474 ~ 0.894	0.008	0.720	0.518 ~ 1.006	0.052
		Countryside / City	0.588	0.463 ~ 0.748	< 0.001	0.633	0.491 ~ 0.817	< 0.001
	BMI categories	BMI ≥ 24 / BMI < 24	0.681	0.530 ~ 0.877	0.003			
Food items/day	Refined rice, steamed/day	>=200 g/<200 g	0.784	0.628 ~ 0.976	0.030			
	Refined rice, porridge/day	>=25 g/<25 g	0.785	0.639 ~ 0.963	0.021			
	Refined wheat products/day	>=50 g/<50 g	0.846	0.686 ~ 1.042	0.116			
	Desserts/day	>=25 g/<25 g	1.129	0.918 ~ 1.389	0.250			
	Whole grains/day	>=10 g/<10 g	1.049	0.848 ~ 1.295	0.660			
	Tubers/day	>=25 g/<25 g	0.902	0.734 ~ 1.109	0.328			
	Milk and yogurt/day	>=200ml/<200 ml	0.742	0.603 ~ 0.912	0.005	0.730	0.587 ~ 0.906	0.005
	Eggs/day	>=50 g/<50 g	0.816	0.660 ~ 1.010	0.061			
	Livestock meat/day	>=100 g/<100 g	0.890	0.725 ~ 1.093	0.267			
	Poultry meat/day	>=50 g/<50 g	0.871	0.710 ~ 1.070	0.188			
	River fish, shrimp and crab/day	>=10 g/<10 g	1.030	0.833 ~ 1.271	0.786			
	Sea fish, shrimp and crab/day	>=10 g/<10 g	1.151	0.937 ~ 1.414	0.181			
	Soybeans and Soy products/day	>=25 g/<25 g	0.935	0.761 ~ 1.148	0.520			
	Nuts and seeds/day	>=5 g/<5 g	1.025	0.834 ~ 1.258	0.817			
	Dark vegetables/day	>=100 g/<100 g	0.782	0.637 ~ 0.960	0.019			
	Light color vegetables/day	>=50 g/<50 g	0.782	0.634 ~ 0.963	0.021			
	Mushrooms/day	>=12 g/<12 g	0.873	0.710 ~ 1.072	0.195			
	Fruits/day	>=100 g/<100 g	0.844	0.687 ~ 1.036	0.105			
	Puffed foods/day	>=10 g/<10 g	1.517	1.234 ~ 1.865	< 0.001	1.543	1.244 ~ 1.915	< 0.001
	Candy and Chocolate/day	>=6 g/<6 g	1.227	0.998 ~ 1.510	0.052			
	Sugary drinks/day	>=100ml/<100 ml	1.058	0.861 ~ 1.301	0.590			
	Taken calcium or vitamin D supplements within 3 months	Yes/no	0.673	0.469 ~ 0.972	0.033	0.670	0.458 ~ 0.988	0.041
Physical activity/day	Vigorous physical activity/day	1 ~ < 2 h/<1 h	1.029	0.802 ~ 1.320	0.823			
		>=2 h/<1 h	0.682	0.532 ~ 0.875	0.003			
	Moderate physical activity/day	1 ~ < 2 h/<1 h	1.043	0.816 ~ 1.338	0.738			
		>=2 h/<1 h	0.988	0.743 ~ 1.321	0.935			
	Outdoor activity/day	0.5 ~ < 1 h /<0.5 h	0.999	0.763 ~ 1.306	0.993	0.918	0.693 ~ 1.214	0.547
		≥ 1 h /<0.5 h	0.627	0.489 ~ 0.804	< 0.001	0.704	0.541 ~ 0.915	0.009
UV protection	UV protection index	By 1 degree	1.101	1.069 ~ 1.134	< 0.001	1.048	1.014 ~ 1.084	0.006

Author contributions

Lixin Na: Conception and design of the work, substantively revised the manuscript. Yingyi Luo: Data curation, Investigation, drafted the work, prepared figures, wrote the main manuscript text. Chunbo Qu: Supervision, Project administration. Guyanan Li: Data curation, Investigation. Qiannan Di: Data curation, Investigation. Shangzhen Ding: Data curation, Investigation. Ruoyou Jiang: Data curation, Investigation. Ruotong Wang: Data curation, Investigation.Siyuan Wang: Data curation, Investigation. All authors critically reviewed the manuscript and approved the final version submitted for publication.

Funding

This research was funded by "Shanghai Public Health System Construction Three-Year Action Plan, grant number GWVI-11.1-43" and "High level local university projects, grant number A1-2601-24-201001-1".

Data availability

No datasets were generated or analysed during the current study.

Declarations

Ethical approval

The present study was approved by the Ethics Committee of Shanghai University of Medicine and Health Sciences (Protocol code: 2019-CSHS-SUMHS-05-230119197901252329/10-May-2019) and was conducted in accordance with the ethical principles set out in the Helsinki Declaration of 1975.

Informed consent

All participants provided informed consent to participate in the study.

Competing interests

The authors declare no competing interests.

Received: 17 September 2024 / Accepted: 8 April 2025
Published online: 21 April 2025

References

1. Sofianopoulou E, Kaptoge SK, Afzal S, Jiang T, Gill D, Gundersen TE, et al. RETRACTED: estimating dose-response relationships for vitamin D with coronary heart disease, stroke, and all-cause mortality: observational and Mendelian randomisation analyses. Lancet Diabetes Endocrinol. 2021;9(12):837–46.
2. Moukayed M, Grant WB. The roles of UVB and vitamin D in reducing risk of cancer incidence and mortality: A review of the epidemiology, clinical trials, and mechanisms. Reviews Endocr Metabolic Disorders. 2017;18(2):167–82.
3. Liu L, Cao Z, Lu F, Liu Y, Lv Y, Qu Y et al. Vitamin D deficiency and metabolic syndrome in elderly Chinese individuals: evidence from CLHLS. Nutr Metabolism. 2020;17(1).
4. Pereira M, Dantas Damascena A, Galvão Azevedo LM, de Almeida Oliveira T, da Mota Santana J. Vitamin D deficiency aggravates COVID-19: systematic review and meta-analysis. Crit Rev Food Sci Nutr. 2020;62(5):1308–16.
5. Park D, Lee J, Park CY, Shin M-J. Low vitamin D status is associated with increased risk of mortality in Korean men and adults with hypertension: A Population-Based cohort study. Nutrients. 2022;14(9).
6. Hu Y, Chen J, Wang R, Li M, Yun C, Li W et al. Vitamin D nutritional status and its related factors for Chinese children and adolescents in 2010–2012. Nutrients. 2017;9(9).
7. Reddy SV, Kanatani KT, Nakayama T, Adachi Y, Hamazaki K, Onishi K et al. High frequency of vitamin D deficiency in current pregnant Japanese women associated with UV avoidance and hypo-vitamin D diet. PLoS ONE. 2019;14(3).
8. Deplanque X, Wullens A, Norberciak L. Prévalence et facteurs de risque de L'insuffisance En vitamine D Chez L'adulte Sain Entre 18 et 65 Ans Dans Le Nord de La France. La Revue De Médecine Interne. 2017;38(6):368–73.
9. Jiang W, Wu D-B, Xiao G-B, Ding B, Chen E-Q. An epidemiology survey of vitamin D deficiency and its influencing factors. Medicina Clínica. 2020;154(1):7–12.
10. Yun C, Chen J, He Y, Mao D, Wang R, Zhang Y, et al. Vitamin D deficiency prevalence and risk factors among pregnant Chinese women. Public Health Nutr. 2015;20(10):1746–54.
11. Luo Y, Qu C, Zhang R, Zhang J, Han D, Na L. Geographic location and ethnicity comprehensively influenced vitamin D status in college students: a cross-section study from China. J Health Popul Nutr. 2023;42(1).
12. Kuo Y-T, Kuo L-K, Chen C-W, Yuan K-C, Fu C-H, Chiu C-T et al. Score-based prediction model for severe vitamin D deficiency in patients with critical illness: development and validation. Crit Care. 2022;26(1).
13. Kuwabara A, Tsugawa N, Mizuno K, Ogasawara H, Watanabe Y, Tanaka K. A simple questionnaire for the prediction of vitamin D deficiency in Japanese adults (Vitaimn D deficiency questionnaire for Japanese: VDDQ-J). J Bone Miner Metab. 2019;37(5):854–63.
14. Merlijn T, Swart KMA, Lips P, Heymans MW, Sohl E, Van Schoor NM, et al. Prediction of insufficient serum vitamin D status in older women: a validated model. Osteoporos Int. 2018;29(7):1539–47.
15. Bou Kheir G, Khaldi A, Karam A, Duquenne L, Preiser J-C. A dynamic online nomogram predicting severe vitamin D deficiency at ICU admission. Clin Nutr. 2021;40(10):5383–90.
16. Sohl E, Heymans MW, de Jongh RT, den Heijer M, Visser M, Merlijn T, et al. Prediction of vitamin D deficiency by simple patient characteristics. Am J Clin Nutr. 2014;99(5):1089–95.
17. Holick MF, Binkley NC, Bischoff-Ferrari HA, Gordon CM, Hanley DA, Heaney RP, et al. Evaluation, treatment, and prevention of vitamin D deficiency: an endocrine society clinical practice guideline. J Clin Endocrinol Metabolism. 2011;96(7):1911–30.
18. W X, Z Z, H L. Consensus recommendations on vitamin D and their analogs (Chinese). Chin J Osteoporos Bone Min Res. 2018;11(1):1–19.
19. Benedik E. Sources of vitamin D for humans. Int J Vitam Nutr Res. 2022;92(2):118–25.
20. Pfotenhauer KM, Shubrook JH, Vitamin D, Deficiency. Its role in health and disease, and current supplementation recommendations. J Osteopath Med. 2017;117(5):301–5.
21. Luo Y, Qu C, Zhang R, Zhang J, Han D, Zhang Q et al. Diet, physical activity, and UV protection comprehensively influenced vitamin D status in college students: a cross-section study from China. J Health Popul Nutr. 2023;42(1).
22. Gao J, Fei J, Jiang L, Yao W, Lin B, Guo H. Assessment of the reproducibility and validity of a simple food-frequency questionnaire used in dietary patterns studies. Acta Nutrimenta Sinica. 2011;33(5):452–6.
23. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol. 2009;9(1).
24. Gao M, Wei Y, Yu J, Yu C, Gao Y, Bian Z, et al. The cut-off points of body mass index and waist circumference for predicting metabolic risk factors in Chinese adults. Zhonghua Liuxingbingxue Zazhi. 2019;40(12):1533–40.
25. Chen A, Zhou W, Hou J, Nevill A, Ding Y, Wan Y, et al. Impact of older age adiposity on incident diabetes: A Community-Based cohort study in China. Diabetes Metabolism J. 2022;46(5):733–46.
26. Wynants L, Bouwmeester W, Moons KGM, Moerbeek M, Timmerman D, Van Huffel S, et al. A simulation study of sample size demonstrated the importance of the number of events per variable to develop prediction models in clustered data. J Clin Epidemiol. 2015;68(12):1406–14.
27. Yang C, Mao M, Ping L, Yu D. Prevalence of vitamin D deficiency and insufficiency among 460,537 children in 825 hospitals from 18 provinces in Mainland China. Medicine. 2020;99(44).
28. Yang C, Jing W, Ge S, Sun W. Vitamin D status and vitamin D deficiency risk factors among pregnancy of Shanghai in China. BMC Pregnancy Childbirth. 2021;21(1).
29. Wang Q, Yu D, Wang J, Lin S. Association between vitamin D deficiency and fragility fractures in Chinese elderly patients: a cross-sectional study. Annals Palliat Med. 2020;9(4):1660–5.
30. Valer-Martinez A, Sayon-Orea C, Martínez Hernandez JA, De la Fuente-Arrillaga C, Pérez de Rojas J, Barcones F, et al. Forecasting levels of serum 25-hydroxyvitamin D based on dietary intake, lifestyle and personal determinants in a sample of Southern Europeans. Br J Nutr. 2023;130(10):1814–22.
31. Narang RK, Gamble GG, Khaw K-T, Camargo CA, Sluyter JD, Scragg RKR et al. A prediction tool for vitamin D deficiency in new Zealand adults. Archives Osteoporos. 2020;15(1).
32. Viprey M, Merle B, Riche B, Freyssenge J, Rippert P, Chakir M-A et al. Development and validation of a predictive model of hypovitaminosis D in general adult population: SCOPYD study. Nutrients. 2021;13(8).
33. Horton-French K, Dunlop E, Lucas RM, Pereira G, Black LJ. Prevalence and predictors of vitamin D deficiency in a nationally representative sample of Australian adolescents and young adults. Eur J Clin Nutr. 2021;75(11):1627–36.
34. Torres-Gonzalez M, Cifelli CJ, Agarwal S, Fulgoni VL. Association of milk consumption and vitamin D status in the US population by ethnicity: NHANES 2001–2010 analysis. Nutrients. 2020;12(12).

Publisher's note