Factors determining the serum 25-hydroxyvitamin D response to vitamin D supplementation: data mining approach

Zahra Amiri, Mina Nosrati, Payam Sharifan, Sara Saffar Soflaei, Susan Darroudi, Hamideh Ghazizadeh, Maryam Mohammadi Bajgiran, Fahimeh Moafian, Maryam Tayefi, Elahe Hasanzade, Mahdi Rafiee, Gordon Ferns, Habibollah Esmaily, Mahnaz Amini, Majid Ghayour-Mobarhan

Publication date
10-06-2023

Published in
BioFactors

Link to external publisher version
https://doi.org/10.1002/biof.1770

Factors determining the serum 25-Hydroxyvitamin D response to vitamin D supplementation; data mining approach

Zahra Amiri[1,2]*, Mina Nosrati[2,3]*, Payam Sharifan[2,4,5]*, Sara Saffar Soflaei[2], Susan Darroudi[2], Hamideh Ghazizadeh[2,5], Maryam Mohammadi Bajgiran[2], Fahimeh Moafian[1,2], Maryam Tayefi[6], Elahe Hasanzade[5], Mahdi Rafiee[5], Gordon A. Ferns[7], Habibollah Esmaily[8], Mahnaz Amini[9]**, Majid Ghayour-Mobarhan[2,3,4]**

1. Department of Pure Mathematics, Center of Excellence in Analysis on Algebraic Structures (CEAAS), Ferdowsi University of Mashhad, P.O. Box 1159, Mashhad 91775, IRAN

2. International UNESCO center for Health-Related Basic Sciences and Human Nutrition, Mashhad University of Medical Sciences, Mashhad, Iran

3. Metabolic Research Center, School of Medicine, Mashhad University of Medical Sciences, Iran

4. Department of Nutrition, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.

5. Student Research Committee, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.

6. Norwegian Center for e-health Research, University hospital of North Norway, Tromsø, Norway

7. Division of Medical Education, Brighton & Sussex Medical School, Falmer, Brighton, Sussex, UK

8. Social Determinants of Health Research Center, Mashhad University of Medical Sciences, Mashhad, Iran

9. Allergy Research Center, Mashhad University of Medical Sciences, Mashhad, Iran

****Corresponding Authors:**

****Majid Ghayour-Mobarhan,** MD, PhD; Metabolic Syndrome Research Center, Mashhad University of Medical Sciences, Mashhad, Iran. Postal code: 99199-91766. Tel: +98 513 800 2288, Fax: +98 513 800 2287. Email: ghayourm@mums.ac.ir

**** Mahnaz Amini.**

Running title: Data mining and vitamin D supplementation

Conflict of interest: The authors have no conflict of interest to disclose.

Abstract

Backgrounds: Vitamin D supplementation has been shown to prevent vitamin D deficiency, but various factors can affect the response to supplementation. Data mining is a statistical method for pulling out information from large databases. We aimed to evaluate the factors influencing serum 25-Hydroxyvitamin D levels in response to supplementation of vitamin D using a random forest (RF) model.

Methods: Data was extracted from the Survey of Ultraviolet Intake by Nutritional Approach (SUVINA) study. Vitamin D levels were measured at baseline and at the end of study to evaluate the responsive. We examined the relationship between 76 potential influencing factors on vitamin D response using RF.

Result: We found several features that were highly correlated to the serum vitamin D response to supplementation by RF including anthropometric factors (body mass index (BMI), free fat mass (FFM), fat percentage, waist to hip ratio (WHR)), liver function tests (serum Gamma glutamyl transferase (GGT), total bilirubin, total protein), hematological parameters (mean corpuscular volume (MCV), mean corpuscular hemoglobin concentration (MCHC), hematocrit), measurement of insulin sensitivity (homeostatic model assessment of insulin resistance).

Conclusion: BMI, total bilirubin, FFM, and GGT were found to have a positive relationship and HOMA, MCV, MCHC, fat percentage, total protein and WHR were found to have a negative correlation to vitamin D concentration in response to supplementation. The accuracy of RF in predicting the response was 93% compared to logistic regression, for which the accuracy was 40%, in the evaluation of the correlation of the components of the dataset to serum vitamin D.

Key words: vitamin D, random forest, data mining

Abbreviation list:

AKI (acute kidney injury), ALP (alkaline phosphatase), ALT (alanine aminotransferase), AST (aspartate aminotransferase), BMI (body mass index), BMR (basal metabolic rate), BUN (blood urine nitrogen), CHD (chronic heart disease), CKD (chronic kidney disease), CPK (creatine phosphokinase), DBP (diastolic blood pressure), DII (normal dietary inflammatory index), DM (diabetes mellitus), FBG (fasting blood glucose), FFM (fat free mass), FM (fat mass), GGT(gamma-glutamyl transferase), HC (hip circumference), HC (hematocrit), HDL-C (high density lipoprotein-cholesterol), HEI (healthy eating index), HOMA (homeostatic model assessment for insulin resistance), HMG-CoAR (3hydroxy-3- methyl- glutaryl-coenzyme A reductase), Hs-CRP (high sensitivity C-reactive protein), HTN (hypertension), LDL-C (low density lipoprotein-cholesterol), MAC (mid arm circumference), MCH (mean corpuscular hemoglobin), MCHC (mean corpuscular hemoglobin concentration), MetS (metabolic syndrome), MCV (mean corpuscular volume), PAB (pro-oxidant-oxidant balance), PAL (physical activity level), PMS (premenstrual syndrome), PSQI (Pittsburg sleep quality index), PSST (premenstrual symptoms screening tool), QoL (quality of life), QUICKI (quantitatile insulin sensitivity check index), RBC (red blood cell), RDW (red cell distribution width), RLS (restless leg syndrome), SBP (systolic blood pressure), SUVINA (survey of ultraviolet intake by nutritional approach), TBW (total body water), T2DM (type 2 diabetes mellitus), VDR (vitamin D receptor), WBCs (withe blood cell), WC (waist circumference), WHR (waist to hip ratio).

Introduction

Approximately one billion people suffer from vitamin D deficiency, globally (1). Vitamin D has important roles for normal skeletal and extra-skeletal function. Vitamin D required for maintaining optimum health and normal growth (2) and deficiency is associated with a number of diseases such as cancer, inflammatory and autoimmunity disease (3-7).

Vitamin D deficiency is an important public health problem in developing countries and its prevalence has been reported to be approximately 79% in Iranian adults (8). Various factors can be affected on vitamin D status. Studies have been shown there is a relationship between hematological factors and vitamin D (9-11). It has been shown that vitamin D status has an influence on erythropoiesis, and the receptor for vitamin D is expressed on cells within the bone marrow that include accessory and stromal cells (11). It has been reported that levels of vitamin D can influence production of serum levels of cytokines that can due to augmentation of white blood cells (WBCs) (12). Several demographic and biological factors, such as body mass index (BMI) and body fat percent (13, 14), aging (15), ethnicity (16) have been investigated in association to response to supplementation of vitamin D.

Several studies have shown that vitamin D levels are inversely correlated with BMI. Obese individuals have lower serum levels of vitamin D amount 57% (17). This may be because vitamin D is a fat soluble vitamin and increase degree of adiposity due to stick vitamin D in a large pool of body fat (17).

Data mining is a statistical method for extracting information from large databases to find meaningful pattern, trends, association of interest (18, 19). Recently, researchers have used data mining models such as random forest, decision tree to predict correlated risk factors of CHD (chronic heart disease) (18) and diabetes mellitus (20). A review study was conducted by Saberi-Karimian et al to investigate learning machine methods in several diseases. They observed that RF has the best prediction for CKD (chronic kidney disease) (21). Data mining can identify new determinants and also reveal associations among factors which can indicate patterns and develop predictions according to new factors (20). RF is one of the most popular and powerful supervised machine learning models. The "Forest" it builds, is an ensemble of decision trees. Xu et al, used RF model to T2DM (type 2 diabetes mellitus) risk prediction, they found that RF model can effectively predict risk of diabetes (22). Lin et al to predict mortality of patients with AKI (acute kidney injury) using RF. they found that RF has the best accuracy to prediction among the other models (23).

Whereas few studies have been performed in association between biochemical factors and response to vitamin D supplementation, in this study, for the first time, we aimed to evaluate the factors influencing in serum 25-Hydroxyvitamin D levels in response to supplementation of vitamin D by random forest model in an Iranian population.

Method

Study design

Data for this study was obtained from the SUVINA (survey of ultraviolet intake by nutritional approach) randomized control trial. The protocol of the SUVINA study has been published elsewhere (24). This study was performed in the City of Mashhad in Iran from January 2019 to March of 2019 for ten weeks and comprised 306 adults (both case and control groups). Individuals were selected from MUMS (Mashhad University of Medical Sciences) students, staff and relatives who were eligible. The inclusion criteria were: subjects without any renal and liver malignancies who were 30 to 50 years old. Abdominal obesity was defined using the International Diabetes Federation criteria (waist circumference: ≥ 80 cm for females, ≥ 94 cm for males) (25). The exclusion criteria were: the use of any medication, supplement containing vitamin D, pregnancy, lactation, smoking habit, drinking alcohol, weight changes (more than 5 kg during previous year), liver or renal diseases, under 30 and over 50 ages, any specific diet such as veganism, and unwillingness to complete the trial. During the trial 17 participants were excluded from the study and finally, 289 participants finished the trial. All participants provided written informed consent. Demographic and anthropometric information collected from each subject at the beginning of the study by questionnaire. Each subjects completed a validated questionnaire that included on: clinical history (medication and diet history), depression, stress, anxiety, physical activity, PMS (premenstrual syndrome), sleeping insomnia, sleep quality, RLS (restless legs syndrome), sleep apnea and quality of life. Anthropometric data included: weight, height, WC (waist circumference), HC (hip circumference), WHR by a flexible measuring tape, weight was measured with light clothes and standing height was assessed with a wall stadiometer. All measurements performed by an experienced person. BMI, FM (fat mass), FFM, TBW (total body water) was assessed using

bio-impedance body analyzer (*TANITA BC 418* model). Blood pressures (BP) (systolic and diastolic) were measured from right arm of participants while they were in sitting down on the chair. They rested for 15 min then BP was measured by an experienced person. Our study was registered by Research Ethics Committee of MUMS with registration number of IRCT20101130005280N27.

Modeling

Random Forest is one of the most popular and powerful supervised machine learning algorithm. The "Forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. Generally, random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. One of the biggest advantages of random forest is its versatility. It can be used for both regression and classification tasks, and it’s also easy to view the relative importance it assigns to the input features. Random forest is also a very useful algorithm because the default hyper-parameters it uses often produce a good prediction result. Understanding the hyper-parameters is pretty straightforward, and there is also not that many of them. We are trying to predict effect of different factors in serum vit.D levels after supplementation by RF.

X is the set of all the above features apart from except "Vitamin D" and Y is "Vitamin D" as the target. Using thertiles, we categorized the target column into three categories, with 33% in class 0, 33% in class 1, and 33% in class 2. Then a ratio of 85/15 is used for data splitting such that 85% goes to training set and 15% to the testing set.

In the random forest model a low correlation between models is the key, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. The reason for this wonderful effect is that the trees protect each other from their individual errors. While some trees may be wrong, many other trees will be right, so as a group the trees

are able to move in the correct direction. Therefore, the prerequisites for random forest to perform well are:

1. There needs to be some actual signal in our features so that models built using those features do better than random guessing.

2. The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

Here we used the 3.8.6 version of Phyton software for data analysis by random forest algorithm.

Our dataset contained 76 features in 142 subjects (as an intervention group), and included:

anthropometric data (weight, height ,body mass index (BMI), waist to hip ratio (WHR), waist circumference (WC), hip circumference (HC), mid arm circumference (MAC), neck circumference), body composition data (free fat mass (FFM), body fat mass, body fat percentage, trunk fat, total body water (TBW)), hematological data (white blood cell (WBC), red blood cell (RBC), hemoglobin, hematocrit, mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), platelets, red cell distribution width (RDW), mean platelet volume (MPV)), liver function tests (aspartate aminotransferase (AST), alanine aminotransferase (ALT), alkaline phosphatase (ALP), gamma-glutamyl transferase (GGT), creatine phosphokinase(CPK), Direct bilirubin, Total bilirubin, albumin, total protein)), renal function tests (uric acid, creatinine, blood urine nitrogen (BUN)), measurement of insulin sensitivity data (insulin, homeostatic model assessment for insulin resistance (HOMA-IR), fasting blood glucose (FBG), quantitative insulin sensitivity check index (QUICKI)), lipid profile data (cholesterol, triglyceride, high density lipoprotein-cholesterol (HDL-C), low density lipoprotein-cholesterol (LDL-C)) micronutrients data (calcium, phosphorus, iron, magnesium, vitamin D), inflammatory tests (pro-oxidant-oxidant balance (PAB), high sensitivity C-reactive protein (Hs-CRP)), demographic data (age, sex, systolic blood pressure (SBP), diastolic blood pressure (DBP), insomnia, restless legs syndrome (RLS), apnea, education level, metabolic syndrome (MetS), depression, anxiety, stress, diabetes mellitus (DM), hypertension (HTN)) and other data including premenstrual symptoms screening tool (PSST) Pittsburg sleep quality index (PSQI), quality of life (QoL), healthy eating index (HEI), unhealthy-dietary pattern, antioxidant-dietary pattern, basal metabolic rate (BMR), normal dietary inflammatory index (DII), physical activity level (PAL), Epworth.

Feature importance

Another great quality of the random forest algorithm is that it is easy to measure the relative importance of each feature on the prediction. Scikit-learn provides a goodgreat tool for this that measures a feature's importance by looking at how much the tree nodes that use that feature reduce impurity across all trees in the forest.*Hyper-parameter tuning*

Some hyper-parameters in random forest such as n-estimators and max-depth is used to increase the predictive power of the model and the others like random-state increase the speed of the model.

Result

First, we detected highly correlated features in the training subset of X, I mean those features which have more than 70% correlation with each other in X-train. We have found 13 correlate d groups out of 75 total features and put them in 13 correlated groups (table 1).

Feature importance

We obtain the feature importance of each feature of correlated groups by using the feature importance property of the random forest classifier for extracting the 13 top important features from the training set (table 2).

Note that, "hematocrit" is the mosttop important feature in both groups 9 and 10. We selected the 12 most important features of each group and dropped other features from training and testing subsets. The correlation among these selected features is shown in the figure 1.

Hyper parameter tuning

We perform the tuning of these hyper-parameters as n-estimators=100, max-depth=7, and random-state=42, which improve the performance of the model.

Sequentially, by training the random forest classifier model on the training dataset and evaluating the model on the testing dataset, subject to the same feature selection scheme, it gave us 93% accuracy.

The pair plot of three important features with respect to the target (Vitamin D) were shown in figure 2.

Logistic regression

We also trained our training dataset by Logistic Regression model. After dropping those features which have more than 70 percent correlation, i.e., X-train was contained important features and those features which have less than 70 percent correlation, we evaluated the model on tasting dataset it gave us 62% accuracy

Moreover, it gave us a 44% accuracy when X-train and X-test just contains the important features.

In comparison between logistic regression and random forest algorithm, random forest has a better performance than logistic regression on our dataset. As you can see it has high accuracy on selected features while logistic regression has low performance with very weak accuracy on these features.

Discussion

The novelty of this investigation was that it used a new method of the random forest algorithm for evaluating the factors associated with 25-Hydroxyvitamin D serum levels in response to supplementation of vitamin D using the random forest model. Our results show that BMI, total bilirubin, FFM and GGT factors have a positive correlation to serum vitamin D in response to supplementation. In other words, increasing in each of these variables related to increasing the response to vitamin D supplementation and elevation of serum levels of vitamin D. While, other factors including cholesterol, HOMA, hematocrit, MCV, MCHC, fat percentage, total protein and WHR had a negative association with serum vitamin D changes associated with supplementation.

One of an important result from this study was the high accuracy of the model for factors determining the response to vitamin D supplementation, approximately 93% using random forest model. A study conducted by Sambasivam et al. to assess the severity of vitamin D deficiency, used 11 different algorithms of machine learning models (26). Their findings were similar to our findings they have indicated that the random forest method had highest accuracy (96%) in a vitamin D related context which to predict severity of vitamin D deficiency (26).

Investigations to evaluate different factors affecting the 25(OH)D concentration in response to vitamin D supplementation have been limited. The effect of BMI, fat percentage and some other factors such as age, supplement type, vitamin D baseline, intake of calcium, ethnicity have been well examined but no findings about biochemical factors (27). Many factors exist which can affect the way subjects respond to supplemental vitamin D. Our study in one of investigation to assess biochemical factors affecting 25(OH)D concentration in response to vitamin D supplementation by high accuracy (93%) with random forest model.

One study investigated the response to supplementation of vitamin D (400 IU/d) for one month in non-obese and obese children. They found that response to vitamin D supplementation was different between non-obese and obese subjects. Significant effectiveness of treatment was observed in non-obese, those with 25(OH)D <30 ng/ml at the base line and in obese, those with 25(OH)D ≤20 ng/ml at the base line. They found a positive correlation between intake of weight based vitamin D (supplement plus diet) and changes in serum levels of 25(OH)D in obese group (β= 0.056, p= 0.81) but it was not significant. The associations of serum 25(OH)D in response to dosage of vitamin D supplement based on weight were not significant in obese and non-obese groups. However, in non-obese group, taking of vitamin D from supplement and diet was higher ($P < 0.001$) based on weight, but changes in serum levels of 25(OH)D was low (p= 0.54). They proposed that the body weight is not a key factor for serum levels maintenance of 25(OH)D (28). But other studies have shown a negative association (29, 30).

No investigation have been performed investigating the association between bilirubin and serum 25(OH)D concentration in response to supplementation of vitamin D. It has been shown that individuals with high levels of total bilirubin have low concentration of vitamin D (31), also, subjects with vitamin D deficiency may response differently with vitamin D supplementation (32), so it has concluded that individuals with high levels of total bilirubin may be more absorption of vitamin D consequence of supplementation and have been higher levels of vitamin D.

Ghayour-Mobarhan et al. carried out a study to evaluate association between high dose of vitamin D supplementation (50,000 IU/week) and serum markers improvement of liver function in 988 girls between 12-18 years for 9 weeks. Individuals divided two groups including subjects with abnormal and subjects with normal liver markers. The markers of liver function were measured at the baseline and at the end of study. Serum concentration of vitamin D elevated significantly at the end of study. Treatment response to supplementation of vitamin

D about total bilirubin ($p<0.001$) and GGT ($p=0.001$) were significant in abnormal group but not significant in normal group (33).

Gonoodi K et al, studied 940 persons between 12 to 18 years who received 50000 IU/W for 9 weeks. FFQ filled out to dietary intakes and vitamin D serum levels was measured before and after the supplementation. They used from decision tree method to evaluate the influence of numerous biomarkers in vitamin D levels by supplementation. They found that serum LDL-C, HDL-C, triglyceride, age and WHR correlated with serum vitamin D levels with response to supplementation of vitamin D using a decision-tree model. Also, they have shown this model had specificity, sensitivity and accuracy 75.8%, 59.4% and 69.3% respectively (34).

Higher body fat percent was reported to be inversely correlated with concentration of 25(OH)D in response to supplementation of vitamin D (29, 30). Blum et al. carried out a study in subjects aged ≥65 years who took 700 IU vitamin D supplement daily. They found that the changes in concentration of 25(OH)D were inversely associated with central body fat and waist circumference (14) that this is consistent to our findings. Several studies showed that decrease body fat percent for each 1 unit, 25(OH)D changes will be elevate by 0.7 nmol/L (35). The mechanism pathway for this event is that vitamin D is a fat soluble vitamin that stores in fat tissue. The increasing fat tissue likely due to trap vitamin D in there (17). Consequently, it concluded that with decrease of FM percent and increase of FFM, concentration of vitamin D in response to supplementation may be increased.

Results of studies demonstrated that between vitamin D and total cholesterol has an inversely correlation independent of sex, age, smoking, physical activity level, drinking, BMI, energy intake and cholesterol intake (36, 37) Li et al conducted a study of 1172 human participants and an animal study included 16 male rats and in 2008. Lipid profiles measured by assay kit and vitamin D measured by ELISA kit. They found that deficiency of vitamin D was associated with increased cholesterol synthesis. The mechanism for this was related to the activity of VDR (vitamin D receptor) decreased by vitamin D deficiency and reduced insulin-induced gene-2 (Insig-2) expression and it has inhibited activation of SREBP-2 and due to elevate of HMG-CoAR (3hydroxy-3- methyl- glutaryl- coenzyme A reductase) expression consequently cholesterol synthesis from liver (36).

Al-Daghri et al assessed the effect of vitamin D supplementation in 204 patients with type 2 diabetes (38). The participants received 2000 IU supplement of vitamin D daily for twelve months. The vitamin D concentration was evaluated at the first and end of the study. Findings

have been indicated that a negative correlation has between vitamin D supplementation and HOMA. Vitamin D elevated HOMA β-cell function (38). Moreover, vitamin D supplementation has a reverse association to hematocrit and MCHC and a positive relation to MCV in one study (39). Arnberg et al investigated 312 infants aged 9 months old. Blood samples were collected between nine am and three pm. The parents were ordered don't feed the infant 2 hours prior to blood collecting. They reported that vitamin D plasma concentration had an inverse relationship with TG, TC, BMI and WC (37). But there are no studies on the factors affecting serum 25(OH)D in response to vitamin D supplementation.

The association between concentration of 25(OH)D and vitamin D supplementation is not consistent and is affected by many features. Some of these factors including basal serum concentration of 25(OH)D, BMI, fat percent are well established. While the findings for other features areunclear. The mechanisms by which these factors may influence the response to supplement are not well identified so further studies should be considered: (1) to confirm significance of these features, (2) to identify mechanistic pathways which these factors by those exhibit their roles (3) to determine the required dose of supplementation for well- being, (4) to examined VDR genes polymorphisms in various population.

The strengths of our investigation included the large population sample, and we used RF model to predict associations. Limitation our study was the population selected from Iranian people that they located in MASHHAD, so, findings cannot be attributed to other populations.

Conclusion

Data mining has a very important role in medicine. In this study, we observed effect of BMI, total bilirubin, FFM, GGT, cholesterol, HOMA, hematocrit, MCV, MCHC, fat percentage, total protein and WHR on vitamin D concentration in response to supplementation by random forest model with a high accuracy. Since few studies have been performed on the effect of these parameters on vitamin D, further studies should be done in this field.

Acknowledgement

Research reported in this publication was supported by Elite Researcher Grant Committee under award number [957705] from the National Institutes for Medical Research Development (NIMAD), Tehran, Iran

Legends:

Figure 1. The correlation among features and vitamin D as our study target. According to the random Forest model, 12 factors are highly correlated with vitamin D. As shown in the heat map, the highlight areas indicate a strong correlation and the lowlight areas indicate a weak correlation of the factors with vitamin D. The negative and positive signs indicate the inverse and direct relationships, respectively. Hematocrit is the top important feature in both groups 9 and 10.

Figure 2. The pair plot of three important features with respect to vitamin D change in response to vitamin D supplementation

Table 1. Highly correlated characteristics in the dataset

Group	Feature 1	Feature 2	Correlation
1	BMI	TBW	0.991
		FFM	0.858
2	SEX	PSST	0.973
		WHR	0.798
		Hematocrit	0.742
		Height	0.741
		Neck	0.718
3	Albumin	Total Protein	0.921
4	Insulin	HOMA	0.913
		QUIKI	0.837
5	MCV	MCH	0.9
6	AST	ALT	0.892
		Gamma.GT	0.712
7	LDL-C	Cholesterol	0.879
8	Weight	WC	0.849
		MAC	0.785
		BMI	0.775
		HC	0.765
		Wrist	0.708
9	RBC	Hematocrit	0.836
10	Hemoglobin	Hematocrit	0.816
11	Percent FAT	Body Fat Mass	0.786
12	MCHC	MCH	0.762
13	Total bilirubin	Direct bilirubin	0.718

BMI (body mass index), TBW (total body water), FFM (fat free mass), PSST (premenstrual symptoms screening tool), WHR (waist to hip ratio), HOMA (homeostatic model assessment for insulin resistance), QUICKI (quantitative insulin sensitivity check index), MCV(mean corpuscular volume), MCH (mean corpuscular hemoglobin), AST (aspartate aminotransferase), ALT (alanine aminotransferase), Gamma.GT(gamma-glutamyl transferase), LDL-C (low density lipoprotein-cholesterol), WC (waist circumference), MAC (mid arm circumference), HC (hip circumference), RBC(red blood cell), MCHC (mean corpuscular hemoglobin concentration).

Table 2. The importance of each characteristic of the correlated groups.

Group		Feature Importance
1	FFM	0.338
	BMI	0.336
	TBW	0.324
2	WHR	0.26
	Hematocrit	0.228
	Height	0.215
	Neck	0.159
	PSST	0.124
	Sex	0.01
3	Total Protein	0.537
	Albumin	0.462
4	HOMA	0.337
	QUIKI	0.335
	Insulin	0.327
5	MCV	0.513
	MCH	0.486
6	Gamma-GT	0.406
	ALT	0.313
	AST	0.28
7	Cholesterol	0.517
	LDL	0.482
8	BMI	0.222
	Weight	0.185
	WC	0.176
	HC	0.154
	MAC	0.139
	Wrist	0.12
9	Hematocrit	0.5
	RBC	0.499
10	Hematocrit	0.54
	Hemoglobin	0.459
11	percent FAT	0.522
	Body Fat Mass	0.477
12	MCH	0.521
	MCHC	0.478
13	Total bilirubin	0.583
	Direct bilirubin	0.416

FFM (fat free mass), BMI (body mass index), TBW (total body water), WHR (waist to hip ratio), PSST (premenstrual symptoms screening tool), HOMA (homeostatic model assessment for insulin resistance), QUICKI (quantitative insulin sensitivity check index), MCV(mean corpuscular volume), MCH (mean corpuscular hemoglobin), Gamma.GT(gamma-glutamyl transferase), ALT (alanine aminotransferase), AST (aspartate aminotransferase), LDL-C (low density lipoprotein-cholesterol), WC (waist circumference), HC (hip circumference), MAC (mid arm circumference), RBC (red blood cell), MCHC (mean corpuscular hemoglobin concentration).

References

1. Holick MF. Vitamin D deficiency. New England Journal of Medicine. 2007;357(3):266-81.
2. Hoffmann MR, Senior PA, Mager DR. Vitamin D supplementation and health-related quality of life: a systematic review of the literature. Journal of the Academy of Nutrition and Dietetics. 2015;115(3):406-18.
3. Holick MF. Sunlight and vitamin D for bone health and prevention of autoimmune diseases, cancers, and cardiovascular disease. The American journal of clinical nutrition. 2004;80(6):1678S-88S.
4. Abe J, NAKAMURA K, TAKITA Y, NAKANO T, IRIE H, NISHII Y. Prevention of immunological disorders in MRL/l mice by a new synthetic analogue of vitamin D3: 22-oxa-1α, 25-dihydroxyvitamin D3. Journal of nutritional science and vitaminology. 1990;36(1):21-31.
5. Merlino LA, Curtis J, Mikuls TR, Cerhan JR, Criswell LA, Saag KG. Vitamin D intake is inversely associated with rheumatoid arthritis: results from the Iowa Women's Health Study. Arthritis & Rheumatism: Official Journal of the American College of Rheumatology. 2004;50(1):72-7.
6. Mahon B, Gordon S, Cruz J, Cosman F, Cantorna M. Cytokine profile in patients with multiple sclerosis following vitamin D supplementation. Journal of neuroimmunology. 2003;134(1-2):128-32.
7. Zhu Y, Mahon BD, Froicu M, Cantorna MT. Calcium and 1α, 25-dihydroxyvitamin D3 target the TNF-α pathway to suppress experimental inflammatory bowel disease. European journal of immunology. 2005;35(1):217-24.
8. Bonakdaran S, Fakhraee F, Karimian MS, Mirhafez SR, Rokni H, Mohebati M, et al. Association between serum 25-hydroxyvitamin D concentrations and prevalence of metabolic syndrome. Advances in medical sciences. 2016;61(2):219-23.
9. Bella LM, Fieri I, Tessaro FH, Nolasco EL, Nunes FP, Ferreira SS, et al. Vitamin D modulates hematological parameters and cell migration into peritoneal and pulmonary cavities in alloxan-diabetic mice. BioMed research international. 2017;2017.
10. Zhou S, LeBoff MS, Glowacki J. Vitamin D metabolism and action in human bone marrow stromal cells. Endocrinology. 2010;151(1):14-22.
11. Menart-Houtermans B, Rütter R, Nowotny B, Rosenbauer J, Koliaki C, Kahl S, et al. Leukocyte profiles differ between type 1 and type 2 diabetes and are associated with metabolic phenotypes: results from the German Diabetes Study (GDS). Diabetes Care. 2014;37(8):2326-33.
12. Takiishi T, Ding L, Baeke F, Spagnuolo I, Sebastiani G, Laureys J, et al. Dietary supplementation with high doses of regular vitamin D3 safely reduces diabetes incidence in NOD mice when given early and long term. Diabetes. 2014;63(6):2026-36.
13. Zwart SR, Mehta SK, Ploutz-Snyder R, Bourbeau Y, Locke JP, Pierson DL, et al. Response to vitamin D supplementation during Antarctic winter is related to BMI, and supplementation can mitigate Epstein-Barr virus reactivation. The Journal of nutrition. 2011;141(4):692-7.
14. Blum M, Dolnikowski G, Seyoum E, Harris SS, Booth SL, Peterson J, et al. Vitamin D 3 in fat tissue. Endocrine. 2008;33(1):90-4.
15. Chen JS, Sambrook PN, March L, Cameron ID, Cumming RG, Simpson JM, et al. Hypovitaminosis D and parathyroid hormone response in the elderly: effects on bone turnover and mortality. Clinical endocrinology. 2008;68(2):290-8.
16. Gallagher JC, Peacock M, Yalamanchili V, Smith LM. Effects of vitamin D supplementation in older African American women. The Journal of Clinical Endocrinology & Metabolism. 2013;98(3):1137-46.
17. Wortsman J, Matsuoka LY, Chen TC, Lu Z, Holick MF. Decreased bioavailability of vitamin D in obesity. The American journal of clinical nutrition. 2000;72(3):690-3.

18. Tayefi M, Tajfard M, Saffar S, Hanachi P, Amirabadizadeh AR, Esmaeily H, et al. hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm. Computer methods and programs in biomedicine. 2017;141:105-9.
19. Benke K, Benke G. Artificial intelligence and big data in public health. International journal of environmental research and public health. 2018;15(12):2796.
20. Esmaeily H, Tayefi M, Ghayour-Mobarhan M, Amirabadizadeh A. Comparing three data mining algorithms for identifying the associated risk factors of type 2 diabetes. Iranian biomedical journal. 2018;22(5):303.
21. Saberi-Karimian M, Khorasanchi Z, Ghazizadeh H, Tayefi M, Saffar S, Ferns GA, et al. Potential value and impact of data mining and machine learning in clinical diagnostics. Critical Reviews in Clinical Laboratory Sciences. 2021:1-22.
22. Xu W, Zhang J, Zhang Q, Wei X, editors. Risk prediction of type II diabetes based on random forest model. 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB); 2017: IEEE.
23. Lin K, Hu Y, Kong G. Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model. International journal of medical informatics. 2019;125:55-61.
24. Sharifan P, Bagherniya M, Bajgiran MM, Safarian M, Vatanparast H, Eslami S, et al. The efficacy of dairy products fortified with nano-encapsulated vitamin D3 on physical and mental aspects of the health in obese subjects; the protocol of the SUVINA trial. Translational Metabolic Syndrome Research. 2021;4:1-9.
25. Alberti KGMM, Zimmet P, Shaw J. Metabolic syndrome—a new world-wide definition. A consensus statement from the international diabetes federation. Diabetic medicine. 2006;23(5):469-80.
26. Sambasivam G, Amudhavel J, Sathya G. A Predictive Performance Analysis of Vitamin D Deficiency Severity Using Machine Learning Methods. IEEE Access. 2020;8:109492-507.
27. Mazahery H, Von Hurst PR. Factors affecting 25-hydroxyvitamin D concentration in response to vitamin D supplementation. Nutrients. 2015;7(7):5111-42.
28. Rajakumar K, Fernstrom JD, Holick MF, Janosky JE, Greenspan SL. Vitamin D status and response to vitamin D3 in obese vs. non-obese African American Children. Obesity. 2008;16(1):90-5.
29. Barger-Lux M, Heaney R, Dowell S, Chen T, Holick M. Vitamin D and its major metabolites: serum levels after graded oral dosing in healthy men. Osteoporosis International. 1998;8(3):222-30.
30. Giusti A, Barone A, Pioli G, Girasole G, Razzano M, Pizzonia M, et al. Heterogeneity in serum 25-hydroxy-vitamin D response to cholecalciferol in elderly women with secondary hyperparathyroidism and vitamin D deficiency. Journal of the American Geriatrics Society. 2010;58(8):1489-95.
31. Akinlade K, Atere A, Rahamon S, Olaniyi J, Ogundeji P. Vitamin D, bilirubin and urinary albumin-creatinine ratio in adults with sickle cell anaemia. Arch Basic Appl Med. 2014;2:77-82.
32. Talwar SA, Aloia JF, Pollack S, Yeh JK. Dose response to vitamin D supplementation among postmenopausal African American women. The American journal of clinical nutrition. 2007;86(6):1657-62.
33. Tavakoli H, Rostami H, Avan A, Bagherniya M, Ferns GA, Khayyatzadeh SS, et al. High dose vitamin D supplementation is associated with an improvement in serum markers of liver function. BioFactors. 2019;45(3):335-42.
34. Gonoodi K, Tayefi M, Bahrami A, Amirabadi Zadeh A, Ferns GA, Mohammadi F, et al. Determinants of the magnitude of response to vitamin D supplementation in adolescent girls identified using a decision tree algorithm. BioFactors. 2019;45(5):795-802.
35. Mazahery H, Stonehouse W, Von Hurst P. The effect of monthly 50 000 IU or 100 000 IU vitamin D supplements on vitamin D status in premenopausal Middle Eastern women living in Auckland. European journal of clinical nutrition. 2015;69(3):367-72.
36. Li S, He Y, Lin S, Hao L, Ye Y, Lv L, et al. Increase of circulating cholesterol in vitamin D deficiency is linked to reduced vitamin D receptor activity via the Insig-2/SREBP-2 pathway. Molecular nutrition & food research. 2016;60(4):798-809.
37. Arnberg K, Østergård M, Madsen A, Krarup H, Michaelsen K, Mølgaard C. Associations between vitamin D status in infants and blood lipids, body mass index and waist circumference. Acta Paediatrica. 2011;100(9):1244-8.

38. Al-Daghri NM, Mohammed AK, Al-Attas OS, Ansari MGA, Wani K, Hussain SD, et al. Vitamin D receptor gene polymorphisms modify cardiometabolic response to vitamin D supplementation in T2DM patients. Scientific reports. 2017;7(1):1-10.
39. Jastrzebska M, Kaczmarczyk M, Suárez AD, Sánchez GFL, Jastrzebska J, Radziminski L, et al. Iron, hematological parameters and blood plasma lipid profile in vitamin D supplemented and non-supplemented young soccer players subjected to high-intensity interval training. Journal of nutritional science and vitaminology. 2017;63(6):357-64.