Diabetes & Metabolic Syndrome: Clinical Research & Reviews 13 (2019) 1773–1777

Contents lists available at ScienceDirect

Diabetes & Metabolic Syndrome: Clinical Research & Reviews

journal homepage: www.elsevier.com/locate/dsx

Original Article

An assessment of the risk factors for vitamin D deficiency using a decision tree model

Kayhan Gonoodi [a, 1], Maryam Tayefi [b, c, 1], Maryam Saberi-Karimian [d], Alireza Amirabadi zadeh [e], Susan Darroudi [d], Seyed Kazem Farahmand [f], Zahra Abasalti [c], Alireza Moslem [g], Mohsen Nematy [a], Gordon A. Ferns [h], Saeid Eslami [i, j, **], Majid Ghayour Mobarhan [c, k, *]

[a] Department of Clinical Nutrition, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran
[b] Clinical Research Unit, Mashhad University of Medical Sciences, Mashhad, Iran
[c] Metabolic Syndrome Research Center, Mashhad University of Medical Sciences, Mashhad, Iran
[d] Student Research Committee, Metabolic Syndrome Research Center, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran
[e] Medical Toxicology and Drug Abuse Research Center (MTDRC), Birjand University of Medical Sciences, Birjand, Iran
[f] Department of Chinese and Complementary Medicine, School of Persian and Complementary Medicine, Mashhad University of Medical Science, Mashhad, Iran
[g] Department of Anesthesiology, School of Medicine, Sabzevar University of Medical Sciences, Sabzevar, Iran
[h] Brighton & Sussex Medical School, Division of Medical Education, Falmer, Brighton, Sussex, BN1 9PH, UK
[i] Pharmaceutical Research Center, Mashhad University of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran
[j] Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran
[k] Cardiovascular Research Center, Mashhad University of Medical Sciences, Mashhad, Iran

ARTICLE INFO

Article history:
Received 26 February 2019
Accepted 13 March 2019

Keywords:
Data mining
Decision tree
25OHD
Vitamin D deficiency

ABSTRACT

Background and objectives: Vitamin D (25-hydroxyvitamin D or 25OHD) has a key role in the pathogenesis of several chronic disorders. Vitamin D deficiency is a common global public health problem. We aimed to evaluate the risk factors associated with vitamin D deficiency using a decision tree algorithm.
Methods: A total of 988 adolescent girls, aged 12–18 years old, were recruited to the study. Demographic characteristics, serum biochemical factors, all blood count parameters and trace elements such as Zinc, Copper, Calcium and SOD were measured. Serum levels of vitamin D below 20 ng/ml were considered to be deficiency. 70% of these girls (618 cases) were randomly allocated to a training dataset for the constructing of the decision-tree. The remaining 30% (285 cases) were used as the testing dataset to evaluate the performance of decision-tree. In this model, 14 input variables were included: age, academic attainment of their father, waist circumference, waist to hip ratio, zinc, copper, calcium, SOD, FBG, HDL-C, RBC, MCV, MCHC, HCT. The validation of the model was assessed by constructing a receiver operating characteristic (ROC) curve.
Results: The results showed that serum Zn concentration was the most important associated risk factor for vitamin D deficiency. The sensitivity, specificity, accuracy and the area under the ROC curve (AUC) values were 79.3%, 64%, 77.8% and 0.72 respectively using the testing dataset.
Conclusions: The results suggest that the serum levels of Zn is an important associated risk factor for identifying subjects with vitamin D deficiency among Iranian adolescent girls.

1. Introduction

Vitamin D (25-hydroxyvitamin D or 25OHD) is a member of the steroid nuclear hormone superfamily, which has an essential role in the occurrence and development of skeletal disorders and calcium homeostasis [1]. Mainly, it has been shown that vitamin D has an essential role in growth and puberty of adolescents [2,3]. Most

* Corresponding author. Metabolic Syndrome Research Center, School of Medicine, Mashhad University of Medical Sciences, Mashhad, 99199-91766, Iran.
** Corresponding author. Pharmaceutical Research Center, Mashhad university of Medical Sciences, Mashhad, 99199-91766, Iran.
E-mail addresses: Eslamis@mums.ac.ir (S. Eslami), ghayourm@mums.ac.ir (M.G. Mobarhan).
1 Equal first author.

https://doi.org/10.1016/j.dsx.2019.03.020
1871-4021/

chronic conditions such as autoimmune diseases, malignancies, metabolic disorders and infectious diseases can be affected by Vitamin D levels [4]. Several chronic diseases including cardiovascular disease (CVD) have their origins in childhood and adolescence [5].

Vitamin D deficiency is a particular public health problem in developed [4,6], and developing countries [7]. It is highly prevalent worldwide [8], and it is reported that the prevalence is 79% of Iranian adults [7]. Knekt et al. have reported that sufficiency of vitamin D can protect against type 2 diabetes mellitus (T2DM) [9–11]. Although, this has not been a consistent finding. Others have reported that there is no relationship between serum 25OHD levels with metabolic syndrome (MetS) and CVD [7]. It has been shown that the serum levels of copper and zinc can be affected by gender, age and diet. Moreover, the serum concentrations of these trace elements may be influenced by CVD risk factors such as body mass index (BMI), physical activity level (PAL), serum HDL-C and CRP [12].

Previous studies have reported that there is an association between vitamin D and hematological factors [13–15]. It has been shown that erythropoiesis can be affected by vitamin D, because the vitamin D receptor is expressed in the bone marrow by stromal and accessory cells [14]. Aucella et al. have reported that hemoglobin (HB) and hematocrit (HCT) levels are significantly increased by taking vitamin D for a period of 4 months in patients with chronic kidney disease undergoing hemodialysis [16]. Moreover, it has been demonstrated that serum vitamin D levels can affect the production of systemic levels of cytokines and cause to an augmentation in white blood cell (WBC) count [17].

In the current study, we aimed to assess the risk factors associated with vitamin D deficiency by applying a decision tree model, in an Iranian large population.

2. Materials and methods

2.1. Subjects

This study was undertaken in the cities of Mashhad and Sabzevar, in northeastern Iran between January and April 2015. Participants were selected using a randomized clustering method and computer-generated random numbers. Written consent was obtained from the girls and their parents. We excluded girls with any auto-immune diseases, cancer, metabolic bone disease, hepatic or renal failure, cardiovascular disorders, malabsorption or thyroid, parathyroid or adrenal diseases. Subjects who were taking anti-inflammatory, anti-depressant, anti-diabetic, or anti-obesity drugs, vitamin D or calcium supplement use and hormone therapy within the last 6 months were also excluded. A total of 1026 adolescents aged 12–18 y old were screened; of whom, 988 met the inclusion criteria. The protocol was approved by the Ethics Committee of MUMS and all the subject gave informed written consent to participate in the study.

Fasting blood samples (14 h overnight fast) were obtained at baseline and after intervention by venipuncture of an antecubital vein into vacuum tubes. The sera were separated by centrifuging blood samples (Hettich model D-78532) and stored at −80 °C at the reference laboratory in Mashhad University of medical science for future analysis. An electrochemi-luminescence method (ECL, Roche, Basel, Switzerland) was used for the measurement of serum 25OHD concentrations. Complete blood count (CBC) was measured using the Sysmex autoanalyser system KX-21 N in whole blood samples. Serum biochemical parameters and trace elements included zinc, copper, calcium and SOD were assessed using Pars Azmoon kits (Tehran, Iran), as described previously [18].

In this study, subjects were classified as having deficiency of vitamin D less than 20 and the others were considered as normal group [19].

Data mining algorithms, particularly the decision tree, does not work with missing values in target variable, and so even if records had a single missing value, they were deleted from the dataset. Therefore after data cleaning, 903 subjects were included in the final data analysis. All the variables that were significantly different between vitamin D deficiency group and normal group were considered as input variables. All the demographic and biochemical markers, blood count parameters and trace elements which were significantly different between two groups, were considered as input variables. Target variable consisted of two classes. One class was regarded as vitamin D deficiency group and the other one was related to normal level of vitamin D group. In this model, 14 input variables included as age, academic degree of father, waist circumference, waist to hip ratio, zinc, copper, calcium, SOD, FBG, HDL, RBC, MCV, MCHC, HCT.

2.2. The decision tree

Data mining is a popular technique to extract unknown patterns or prediction rules. One of most popular data mining algorithm is decision tree. The decision-tree procedure is a non-parametric method which creates a tree-structured model [20]. It divides subjects into groups or predicts values of a target variable based on values of predictor variables. The main goal of decision tree is to make a predictive model for the target variable according to predictors [21]. The decision tree algorithms consist of three types of nodes, the root node, internal node, and end node (leaves) [20,22]. Decision-tree algorithms use splitting criteria to make branches from root node to leaves to form a tree. It means that the root node, that contains the entire dataset, is split into subgroups by using all predictor variables. This procedure creates the nodes repeatedly, up to form the homogenous subsets with respect to target variable. Therefore tree is constructed from root-node to leaves [21]. Reduction of node impurity is the main goal of splitting criteria. Splitting criteria provide a rate for each predictor variable. Variables which have the best rate of splitting criterion are chosen to remain in the model. Information Gain, Gini index and Gain ratio are the most popular and important splitting criteria [23]. The Classification And Regression Tree (CART) is one decision tree algorithm. CART is constructed by splitting subsets of data using all predictor variables. By this procedure, progressively all leaves are created. The CART algorithm creates a binary division of the tree and pruning a tree on the cost-complexity [24]. Also, the CART algorithm uses the Gini impurity index for selecting the best variable. The Gini index measures impurity:

$$\mathrm{Gini(D)} = 1 - \sum_{i=1}^{m} P_i^2$$

where P_i is the probability that a record in D belongs to class Ci and is estimated by $|Ci,D|/|D|$ [24]. The sum is computed over m classes. In the decision-tree, the first variable in the root node, is the most important variable and others can be classified according to their importance in association with target variable. It is common in data mining methods to divide the data set into two parts; a training data set, and the testing dataset. The model is constructed on training dataset and it is tested on testing dataset.

3. Statistical analysis

All statistical analyses were carried out using R version 3.4.2.

The training dataset included 70% of the data (618) were randomly chosen to construct the decision tree. The testing dataset, means the remaining 30% (285 cases), and these data were used to evaluate the performance of the model. A confusion matrix was applied to validate the performance. The accuracy, sensitivity, specificity and the receiver operating characteristics (ROC) curve were measured for comparison. A ROC graph is a technique for visualizing, organizing and selecting classifiers based on their performance [25]. The area under the ROC curve of the classifier can be described as the probability of the classifier to rank a randomly selected positive case higher than a randomly selected negative case, and the higher area under the ROC curve results the higher accuracy [22,26].

4. Results

The data were divided into a training dataset (70% of the total) and testing dataset (the remaining 30%). In this model, a decision tree was built on the training dataset (618 records). The testing dataset (285 records) were used to evaluate the model. The algorithm used the Gini index for selecting the variables, and the final tree was pruned. In this model, 14 variables such as age, academic degree of father, waist circumference, waist to hip ratio, zinc, copper, calcium, SOD, FBG, HDL, RBC, MCV, MCHC, HCT were considered as input variables which were significantly different between two groups.

The variables, serum zinc, FBG, HGB, serum calcium, MCV, HCT, and MCHC remained in the model. The final decision tree, with size 17, 9 leaves and 6 layers is shown in Fig. 1. The if-then rules created by model is shown in Table 1. The evaluation of the model was carried out using a confusion matrix on a testing dataset and is shown in Table 2. This model had an accuracy of 77.8%. Of the 25 individuals without vitamin D deficiency in the testing dataset, 16 were classified correctly using the decision-tree with a specificity of 64%. For the 260 cases with vitamin D deficiency in the testing dataset, the decision tree correctly classified 206 individuals, with sensitivity of 79.3%. A ROC curve was obtained by applying decision-tree on testing dataset is shown in Fig. 2. The tree showed that in a subgroup with serum zinc<88 (μg/dl), the probability of having vitamin D deficiency was 98% and in a subgroup who had serum zinc ≥ 88 (μg/dl) and FGB ≥ 98 (mg/dl) and HGB ≥ 14, the probability of having deficiency was 85%. In the other subgroup with serum zinc ≥ 88 (μg/dl), FBG ≥ 98 (mg/dl), HGB<14 and MCV> = 91, 78% of individuals were identified as subjects with normal level of vitamin D. In a subgroup of Zinc ≥ 88, FBG ≥ 98, HGB<14, MCV<91 and HCT> = 41, the probability of non-presence of deficiency was 55%. In a same subgroup, with HCT<41, subjects have deficiency with probability of 91%. In the subgroup with serum zinc ≥ 88, FBG<98, Calcium<9.8, subjects had deficiency with probability of 92%. In the same subgroup with Calcium ≥ 9.8, the role of MCV comes up. If MCV<92, the probability of having deficiency was 84%. If MCV ≥ 92, MCHC comes in the model. If MCHC<32, the subjects had no deficiency with probability of 70%. By MCHC ≥ 32, subjects had deficiency with probability of 86% (Table 1). The sensitivity, specificity, accuracy and under the ROC curve (AUC) values for this model was 79.3%, 64%, 77.8% and 0.72 respectively using the testing dataset.

5. Discussion

A major strength of the present study was that it has explored a new application of the decision tree model for examining and evaluating the predicators related to vitamin D deficiency among Iranian population, which has not been considered before. Moreover, this is the first model using blood count parameters beside the other factors as a component in the determination of vitamin D deficiency.

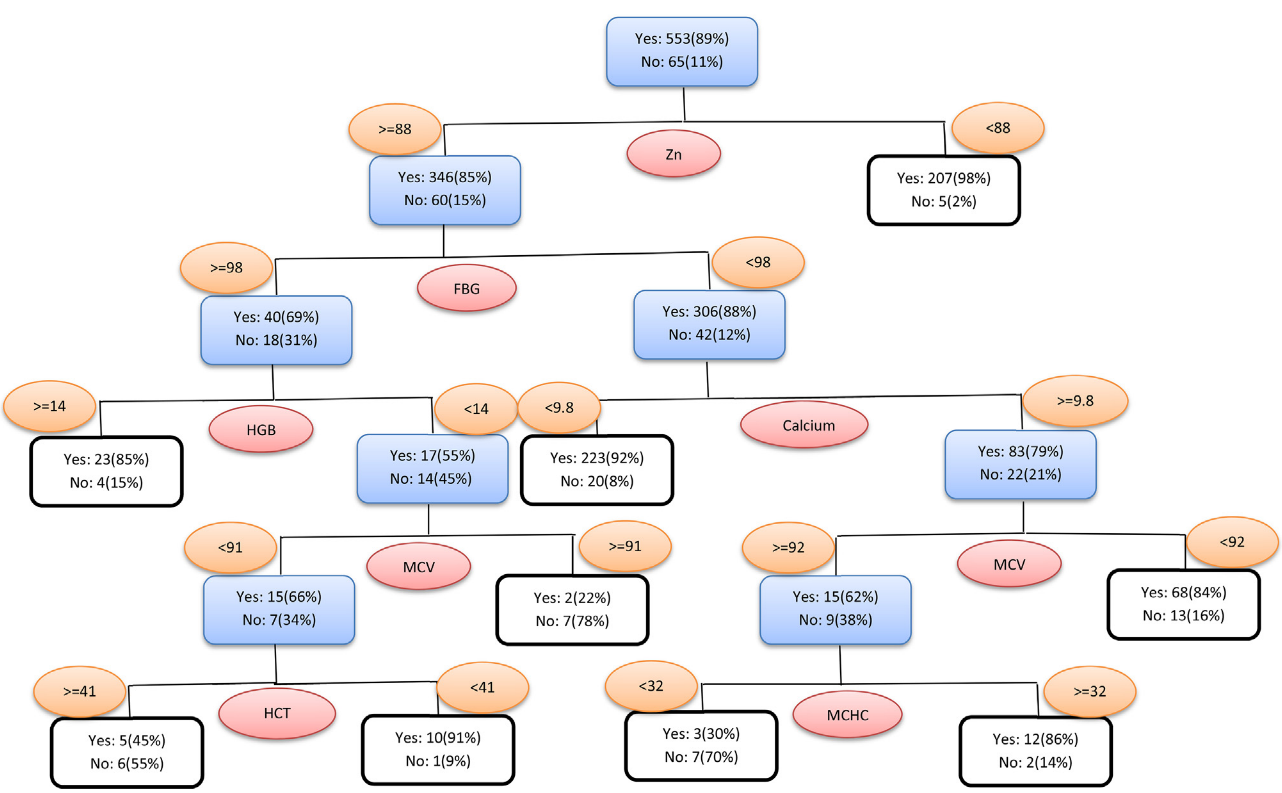

Fig. 1. Decision tree with training dataset: Zn; serum zinc (μg/dl), FBG; fasting blood glucose (mg/dl), calcium; (mmol/l).

Table 1
The 9 rules extracted through the decision tree model.

R1: IF serum Zn < 88 (μg/dl), THEN class: person with deficiency (207/212 or98%)
R2: IF serum Zn ≥ 88 (μg/dl) and FBG ≥ 98 and HGB ≥ 14, THEN class: person with deficiency (23/27 or 85%)
R3: IF serum Zn ≥ 88 (μg/dl) and FBG ≥ 98, HGB<14 and MCV> = 91, THEN class: person without deficiency (7/9 or 78%)
R4: IF serum Zn ≥ 88 (μg/dl) and FBG ≥ 98, HGB<14, MCV<90 and HCT> = 41, THEN class: person without deficiency (6/11 or 55%)
R5: IF serum Zn ≥ 88 (μg/dl) and FBG ≥ 98, HGB<14, MCV<90 and HCT<41, THEN class: person with deficiency (10/11 or 91%)
R6: IF serum Zn ≥ 88 (μg/dl) and FBG<98 and Calcium<9.8, THEN class: person with deficiency (223/243 or 92%)
R7: IF serum Zn ≥ 88 (μg/dl) and FBG<98 and Calcium ≥ 9.8 and MCV<92, THEN class: person with deficiency (68/81 or 84%)
R8: IF serum Zn ≥ 88 (μg/dl) and FBG<98, Calcium ≥ 9.8, MCV ≥ 92 and MCHC<32, THEN class: person without deficiency (7/10 or 70%)
R9: IF serum Zn ≥ 88 (μg/dl) and FBG<98, Calcium ≥ 9.8, MCV ≥ 92 and MCHC ≥ 32, THEN class: person with deficiency (12/14 or 86%)

R: abbreviation of rule.

Table 2
Confusion matrix of testing dataset.

Predicted outcome		
Person without deficiency	Person with deficiency	Actual outcome
54	206	Person with deficiency
16	9	Person without deficiency

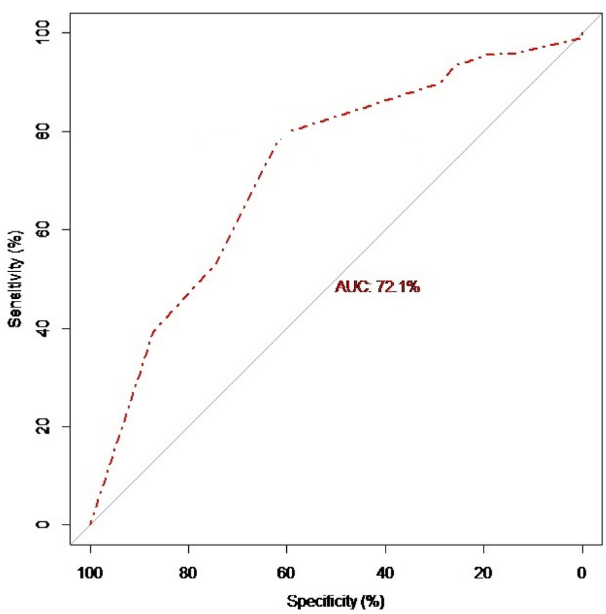

Fig 2. Roc curve of the decision tree model in testing dataset.

One interesting result of the current study was the involvement of serum zinc as an important associated factor of vitamin D deficiency which is consistent with previous studies [27,28]. The results of the CASPIAN-III study has shown that there was a significant association between low serum levels of zinc and 25OHD among Iranian adolescents aged 10–18 years old [27]. Moreover, Ziaei et al. have reported a statistically relationship between serum levels of zinc and 25OHD among Iranian pregnant women. Their results showed that 37% of pregnant women have vitamin D deficiency and 23% have zinc deficiency [28].

Another interesting finding, was the involvement of FBG as an important factor after Zn. If FBG was <98 (mg/dl), the range of Ca goes up to 9.8 (μg/dl). Then, hematological factors determine the other layers. In subjects with serum level of Ca ≥9.8, MCV and MCHC were the associated factor of vitamin D deficiency.

It has been determined that vitamin D deficiency can impair insulin secretion which has major roles in the pathogenesis of T2DM [9,10]. In spite of vitamin D receptor (VDR), it has been shown that vitamin D dependent calcium binding protein has also an important role in 25OHD's functions. The synthesis and secretion of insulin in islet beta cells can be induced by 25OHD, when it combined with the VDR and vitamin D dependent calcium binding protein [29].

Another point in the model was that for the subjects had FBG ≥98 (mg/dl), HB, MCV and HCT were the indicators of presence or non-presence of vitamin D deficiency. The results are consistent with previous studies reported that there is an association between 25OHD and hematological factors [13–17].

In future studies, it is aimed to develop prediction models with more sensitivity and specificity, which can be used for determining the presence or absence of vitamin D deficiency more accurately.

6. Conclusion

In current study, biochemical and hematological markers were used as input variables in a large Iranian adult girls and the results determined the factors associated to vitamin D deficiency. This study provides an easy to use classification rules for categorizing risk factors associated with vitamin D deficiency that can be useful to improve programs for its management.

Funding

This work was supported by Mashhad University of Medical Sciences, Iran.

Conflicts of interest

The authors declare no conflict of interests.

References

[1] Grant WB. Epidemiology of disease risks in relation to vitamin D insufficiency. Prog Biophys Mol Biol 2006;92:65–79.
[2] Cashman KD, Hill TR, Cotter AA, et al. Low vitamin D status adversely affects bone health parameters in adolescents. Am J Clin Nutr 2008;87:1039–44.
[3] Tylavsky FA, Ryder KM, Li R, et al. Preliminary findings: 25 (OH) D levels and PTH are indicators of rapid bone accrual in pubertal children. J Am Coll Nutr 2007;26:462–70.
[4] Lanham-New SA, Buttriss JL, Miles LM, Ashwell M, Berry JL, et al. Proceedings of the rank forum on vitamin D. Br J Nutr 2011;105:144–56.
[5] da Conceicao-Machado MEP, Silva LR, Santana MLP, et al. Hypertriglyceridemic waist phenotype: association with metabolic abnormalities in adolescents. J Pediatr 2013;89:56–63.
[6] Ginde AA, Liu MC, Camargo Jr CA. Demographic differences and trends of vitamin D insufficiency in the US population, 1988–2004. Arch Intern Med 2009;169:626–32.
[7] Bonakdaran Shokoufeh, Fakhraee Farzaneh, Saberi Karimian Maryam, Mirhafez Seyed Reza, Rokni Haleh, Mohebati Mohsen, et al. Association between Serum 25-hydroxyvitamin D concentrations and prevalence of metabolic syndrome. Adv Med Sci 2016 Feb 2;61(2):219–23. https://doi.org/10.1016/j.advms.2016.01.002.
[8] Holick MF. High prevalence of vitamin D inadequacy and implications for health. Mayo Clin Proc 2006;81:353–73.
[9] Mathieu C, Badenhoop K. Vitamin D and type 1 diabetes mellitus: state of the art. Trends Endocrinol Metabol 2005;16:261–6.

[10] Reis AF, Hauache OM, Velho G. Vitamin D endocrine system and the genetic susceptibility to diabetes, obesity and vascular disease. A review of evidence. Diabetes Metab 2005;31:318–25.
[11] Mathieu C, Badenhoop K. Vitamin D and type 1 diabetes mellitus: state of the art. Trends Endocrinol Metabol 2005;16:261–6.
[12] Ghayour-Mobarhan M, Taylor A, New SA, Lamb DJ, Ferns GA. Determinants of serum copper, zinc and selenium in healthy subjects. Ann Clin Biochem 2005 Sep;42(Pt 5):364–75.
[13] Bella LM, Fieri I, Tessaro FHG, Nolasco EL, Nunes FPB, Ferreira SS, et al. Vitamin D modulates hematological parameters and cell migration into peritoneal and pulmonary cavities in alloxan-diabetic mice. BioMed Res Int 2017;2017: 7651815. https://doi.org/10.1155/2017/7651815. Epub 2017 Apr 19.
[14] Zhou S, LeBoff MS, Glowacki J. Vitamin D metabolism and action in human bone marrow stromal cells. Endocrinology 2010;151(6):14–22.
[15] Menart-Houtermans B, Rütter R, Nowotny B, et al. Leukocyte profiles differ between type 1 and type 2 diabetes and are associated with metabolic phenotypes: results fromtheGerman Diabetes Study (GDS). Diabetes Care 2014;37(8):2326–33.
[16] Aucella F, Scalzulli RP, Gatta G, Vigilante M, Carella AM, Stallone C. Calcitriol increases burst-forming unit-erythroid proliferation in chronic renal failure. A synergistic effect with r-HuEpo. Nephron Clinical Practice 2003;95(4):c121–7.
[17] Takiishi T, Ding L, Baeke F, 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.
[18] Khayyatzadeh SS, Mirmoosavi SJ, Fazeli M, Abasalti Z, Avan A, Javandoost A, et al. High-dose vitamin D supplementation is associated with an improvement in several cardio-metabolic risk factors in adolescent girls: a nine-week follow-up study. Ann. Clin. Biochem. 2018;55(2):227–35.
[19] 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 Metab 2011;96(7):1911–30.
[20] Shi G. Chapter 5, decision trees. In: Shi G, editor. Data mining and knowledge discovery for geoscientists. Oxford: Elsevier; 2014. p. 111–38.
[21] Tayefi M, Esmaeili H, Saberi Karimian M, Amirabadi Zadeh A, Ebrahimi M, Safarian M, et al. The application of a decision tree to establish the parameters associated with hypertension. Comput Methods Progr Biomed 2017 Feb;139: 83–91. https://doi.org/10.1016/j.cmpb.2016.10.020. Epub 2016 Oct 24.
[22] Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett 2006;27(8): 861–74.
[23] 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. Comput Methods Progr Biomed 2017 Apr;141:105–9. https://doi.org/10.1016/j.cmpb.2017.02.001. Epub 2017 Feb 3.
[24] Han Jiawei, Jian Pei, Micheline Kamber. Data mining: concepts and techniques. Elsevier; 2011.
[25] Tayefi M, Saberi-Karimian M, Esmaeili H, Zadeh AA, Ebrahimi M, Mohebati M, et al. Evaluating of associated risk factors of metabolic syndrome by using decision tree. Comp. Clin. Pathol. 2018;27(1):215–23. https://doi.org/10.1007/s00580-017-2580-6.
[26] Ke W-s, Hwang Y, Lin E. Pharmacogenomics of drug efficacy in the interferon treatment of chronic hepatitis C using classification algorithms. Comput Biol Chem Adv Appl: AABC 2010;3:39.
[27] Shams B, Afshari E, Tajadini M, Keikha M, Qorbani M, Heshmat R, et al. The relationship of serum vitamin D and Zinc in a nationally representative sample of Iranian children and adolescents: the CASPIAN-III study. Med J Islam Repub Iran 2016 Oct 18;30:430. eCollection 2016.
[28] Ziaei S, Norrozi M, Faghihzadeh S, Jafarbegloo E. A randomised placebo controlled trial to determine the effect of Zinc supplementation on Vitamin D in pregnant women with haemoglobin. BJOG 2007;114(6):684–8.
[29] Dunlop TW, Vaisanen S, Frank C, Molnar F, Sinkkonen L, Carlberg C. The human peroxisome proliferator activated receptor delta gene is a primary target of 1alpha,25-dihydroxyvitamin D3 and its nuclear receptor. J Mol Biol 2005;349:248–60.