A PREDICTIVE PERFORMANCE ANALYSIS OF VITAMIN D DEFICIENCY USING A DECISION TREE MODEL

F. Osmani

Infectious Disease Research Center, Birjand University of Medical Sciences, Birjand, Iran

Abstract. *Background.* HBV infection is a major health problem which may be life-threatening. Vitamin D (VD) is involved in various pathophysiological mechanisms in a plethora of diseases. And also, there is a strong demand for the prediction of its severity using different methods. The study aims to evaluate performance of DT as one of the machine learning models in the prediction of severity in vitamin D deficiency. *Methods.* In total, data containing serum VD levels were collected from 292 CHB patients. The independent characteristics such as: age, sex, weight, height, zinc, BMI, body fat, sunlight exposure, and milk consumption were used for prediction of VD deficiency. 60% of them were allocated to a training dataset randomly. To evaluate the performance of decision-tree the remaining 40% were used as the testing dataset. The validation of the model was evaluated by ROC curve. *Results.* The prevalence of VD deficiency was high among patients (63.0%). The final experimentation results showed that DT classifier achieves better accuracy of 96 % and outperforms well on training and testing of VD dataset. Also, the areas under the ROC curve AUC is 0.78, when we applied DT algorithm with the significant variables by cross validation, the values of AUC = 0.78 and 85.3% accuracy were obtained. *Conclusion.* We concluded that the serum level of Zn is an important associated risk factor for identifying cases with vitamin D deficiency. Also, the risk of VD deficiency could be predicted with high accuracy using decision tree learning algorithm that could be used for antiviral therapy in CHB patients.

Key words: *vitamin D deficiency, decision tree, machine learning, hepatitis B virus, vitamin D, ROC curve.*

ПРОГНОЗНЫЙ АНАЛИЗ ЭФФЕКТИВНОСТИ ДЕФИЦИТА ВИТАМИНА D С ИСПОЛЬЗОВАНИЕМ МОДЕЛИ ДЕРЕВА РЕШЕНИЙ

Османи Ф.

Исследовательский центр инфекционных заболеваний, Бирджандский университет медицинских наук, г. Бирджанд, Иран

Резюме. *Актуальность.* Печень является основным местом синтеза витамина D (ВД), участвующего в различных патофизиологических механизмах при различных заболеваниях. Поэтому важно спрогнозировать степень дефицита ВД при помощи различных методов. Наше исследование было направлено на оценку эффективности дерева решений (DT) как одной из моделей машинного обучения для прогнозирования степени дефицита ВД. *Методы.* Всего было обследовано 292 пациента с ХГВ. У каждого из них определен уровень ВД в сыворотке. Для прогнозирования дефицита ВД использовались независимые характеристики, такие как возраст, пол, вес, рост, содержание цинка, индекс массы тела, жировые отложения, частота и продолжительность воздействия солнечного света и потребление молока. Информация 60% пациентов была внесена в обучающий набор данных случайным образом. Для оценки эффективности дерева решений результаты иссле-

Адрес для переписки:
Фрэштэх Османи
Иран, г. Бирджанд, Исследовательский центр инфекционных заболеваний, Бирджандский университет медицинских наук.
Тел.: +0915 163-76-53.
E-mail: dr.osmani68 @gmail.com

Contacts:
Freshteh Osmani
Iran, Birjand, Infectious Disease Research Center,
Birjand University of Medical Sciences.
Phone: +0915 163-76-53.
E-mail: dr.osmani68 @gmail.com

Для цитирования:
Османи Ф. Прогнозный анализ эффективности дефицита витамина D с использованием модели дерева решений // Инфекция и иммунитет. 2023. Т. 13, № 1. С. 141–146. doi: 10.15789/2220-7619-AOT-1642

Citation:
Osmani F. A predictive performance analysis of vitamin D deficiency using a decision tree model // Russian Journal of Infection and Immunity = Infektsiya i immunitet, 2023, vol. 13, no. 1, pp. 141–146. doi: 10.15789/2220-7619-AOT-1642

DOI: http://dx.doi.org/10.15789/2220-7619-AOT-1642

дований оставшихся 40% пациентов были использованы в качестве набора данных тестирования. Валидация модели оценивалась кривой ROC. *Результаты.* Распространенность дефицита ВД среди пациентов была высокой (63,0%). Окончательные результаты экспериментов показали, что классификатор DT обеспечивает точность 96% и превосходит по производительности при обучении и тестировании набора данных о ВД. Кроме того, площади под кривой ROC AUC составила (0,78) при применении алгоритма DT со значимыми переменными путем перекрестной проверки, с получением значения AUC = 0,78 и точности 85,3%. *Заключение.* Мы пришли к выводу, что уровень цинка в сыворотке крови является важным сопутствующим фактором риска для выявления случаев дефицита ВД. Кроме того, риск дефицита ВД можно предсказать с высокой точностью с использованием алгоритма обучения дерева решений. Полученные данные можно применять в ходе противовирусной терапии у пациентов с ХГВ.

Ключевые слова: *дефицит витамина D, дерево решений, машинное обучение, вирус гепатита B, витамин D, кривая ROC.*

Introduction

HBV infection is a major health problem which may be life-threatening due to its frequent severe complications. In the other hand, VD is an essential vitamin that has powerful influence on several parts of the human body. Nearly one billion people highly suffered from VDD across the globe [4].

Most chronic conditions such as autoimmune diseases and infectious diseases can be affected by VD levels. VDD is a public health problem and is highly prevalent worldwide, so that, it’s prevalence is reported 79% in Iranian adults [5].

There is strong evidence about the association between VD and various chronic liver diseases in different stages [29]. Previous studies have reported that there is an association between vitamin D and hematological factors [3]. Machine learning models will be useful in discovering new patterns of the etiology and thus preventive public health measures can be applied effectively. The traditional severity prediction of VDD have used questionnaires with statistical models such as Linear Regression (LR) [9].

In previous studies, the results were compared between the statistical models and they have not used the machine learning algorithms for the severity prediction. The traditional statistical model like LR is used to predict the severity of VDD but its performance is deprived due to its predictive performance limit and many parameters [7, 10]. Currently, the analysis of VD status is highly expensive, and it is identified using the biochemical methods. The research gap identified urges to condense the cumbersome analytical procedures in identifying VDD among the patients [28].

So, the main objective of this study is to evaluate the DT classifier in the prediction of severity in VDD. And also determine the associated risk factors related with VDD by using DT algorithm, in an Iranian CHB patients.

Method

Two hundred and ninety-two HBV-infected patients were enrolled for this cross-sectional study. Patients were selected randomly according to consent to participate. In this study, we used input parameters such as age, sex (male/female), weight (kgs), height (m), BMI (kg/m), grade and the activity of fibrosis, Sunlight Exposure/Day (hrs.).

Written consent was obtained from the all of patients. Patients with any auto-immune diseases, other viral hepatitis (HCV, HDV, and HIV) and other causes of liver disease, VD, calcium supplement use or injection in the last six months were excluded.

The inclusion criteria were: patients who were admitted to the infectious disease’s outpatient clinic with the diagnosis of CHB with the approval of the infectious specialist and willingness to participate in the study.

The laboratory tests were performed with 10 cc of venous blood was taken from patients (14 h overnight fast). The serum levels of VD were measured using a COBAS e411 analyzer, manufactured by Mannheim Roch diagnostic Gmbh in Germany, with the Elecsys kit (REF 0589413). CBC was measured in whole blood samples.

Total VD levels were measured in the serum samples, then, VD status was classified as normal (≥ 30 ng/ml), insufficient (20–29.9 ng/ml), and deficient (< 20 ng/ml) [27, 31].

Decision Tree (DT). DT Classifier is a well-known supervised ML tool that is used for solving classification problems and it has a tree-like model or graphs. The DT can capture the decision-making knowledge from the given data [12]. In DT that every branch indicates the output of the test set and every leaf node represents the particular label. The classification rules are represented by the path from the root node to the leaf node.

For our VDD severity modeling, each node in the tree predicts the deficiency severity and each branch indicates the states of the variable [19].

ROC curve. ROC curves are used to evaluate the performance of multiclass classifier problems. The ROC curve has false positive rate on X-axis and true positive rate on Y-axis. In the ROC curve topmost left edge of the plot considered to be the ideal point and the steepness of the curve also very important since the TPR value should maximize and the FPR should be minimized [20].

Statistical analysis. All statistical analyses were carried out using R version 3.4.2. The significance

Table. Characteristics of variables

Variables		Training dataset	Validation dataset	Pearson correlation coefficients	p value
Age		36±12	35±11	0.32	< 0.01
Sex	**Male**	72.7%	73.1%		
	Female	27.3%	26.9%	–0.03	0.008
BMI		24.20±4.21	25.32±2.84	0.10	< 0.001
AFP (U/L)		6.56±16.51	6.69±21.49	0.10	< 0.001
AST (U/L)		54.17±23.73	55.78±31.62	0.12	< 0.001
ALT (U/L)		61.84±36.89	61.84±38.19	0.06	0.008
Hemoglobin (Hb)		14.03±1.47	14.03±1.62	–0.02	0.005
Albumin (g/dL)		3.39±0.47	3.40±0.53	0.05	0.064
Platelet count (× 10^9/L)		216.48±53.64	214.55±55.5	0.07	< 0.12

in all of these tests was two-tailed with a 5% significant level. The ROC curve sensitivity, specificity were measured for comparison. Several types of DT learning techniques (CART) [15], C4.5 [26], were implemented on the datasets.

Results

Generally, 48.6% were male; with mean age 29±5.3; and 52.2% female with mean age (31.5±7.8). The data were divided into a training and testing dataset (60% vs 40%) respectively. A decision tree was built on the training dataset. The testing dataset were used to assess the model. Gini index was used for selecting the variables in the algorithm to achieve final tree. The training and test datasets were similar to each other roughly. The results showed that, age, BMI as potential predictors of VDD (P value < 0.001) (Table). Hence, these variables were used in DT model.

DT was learned for the training dataset by using variables with significant correlation with VDD (P value < 0.001).

The final DT, with size 17, 9 leaves and 6 layers is shown in Fig. 1

Fig. 2 showed the accuracy, ROC curve, sensitivity, specificity values for predicting VDD in training set. The areas under the ROC curves is (0.78),

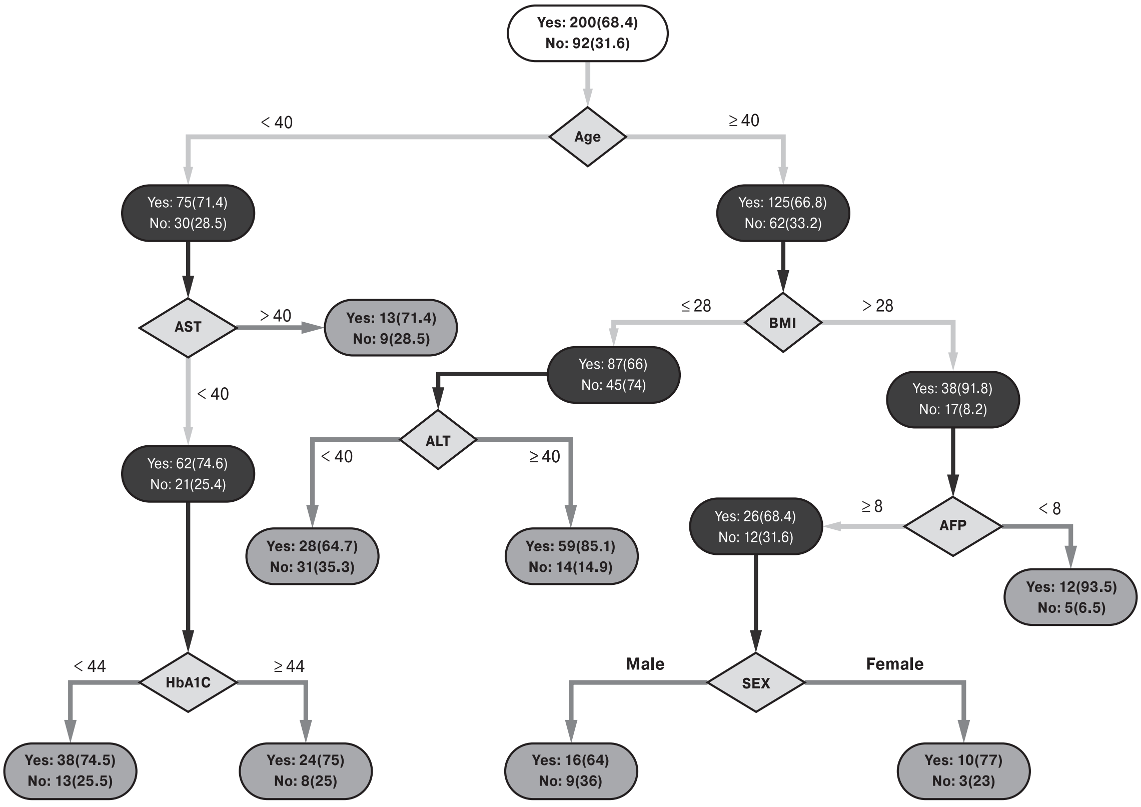

Figure 1. Decision tree with training dataset in CHB group (model 2)

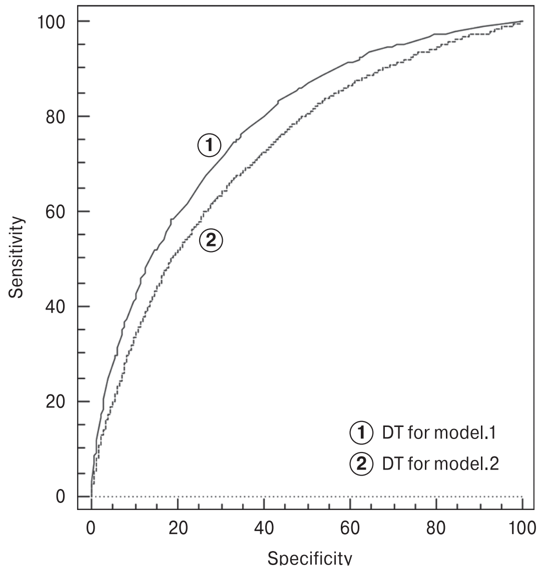

Figure 2. ROC curve of both decision tree models

When we applied DT algorithm with the significant variables by cross validation, the values of 0.78 ROC and 85.3% accuracy were obtained, which is similar to the obtained results of applying training and test sets one by one.

Discussion

The main objective of this study is to evaluate the performance of machine learning model in the prediction of VDD. The prediction accuracy was calculated and compared with the training and testing set. We have used various parameters in the severity prediction.

A main strength of the current study was that it has explored a new application of the DT model for examining and evaluating the predicators related to VDD among CHB Iranian population. Result of this study showed that the serum zinc as a potential factor of VDD which is similar with previous studies [21, 23].

This study shows that insufficiency of VD occurs more frequently among CHB patients.

A high prevalence of VDD (> 80 %) in chronic liver disease has been reported to be associated with liver disease progression [25]. VDD occurs rather often among the elderly population[22].

In the current study, deficiency and severe VDD were observed more frequently in patients with HBV-related liver disease and were found significantly associated with the end-stage of liver cirrhosis; so, our study is in line with previous studies which have indicated an association of VD levels with CHB [5]. The results of this research work proved that the machine learning models especially the random forest classifier accurately predict the severity of VDD [14]. This machine learning classifier will have a greater opportunity in the real-world medical domain which would assist experts to efficiently identify the severity of VDD.

The future direction of our research is to validate the model with a different type of VD datasets of all age groups.

In a period (1990–2010), the prevalence of VDD was studied in Iranian society and according to the results, in all regions; both sexes had moderate and significant VDD [2, 32]. A study showed that VDD existed in many CHB patients. Decreased liver function due to HBV-induced injuries to liver cells can be one of the causes of VDD in CHB [24, 30].

One of the interesting finding of this study was the pretense of serum zinc as a substantial factor for VDD which is in line with other researches [15].

A previous study has shown a significant correlation between serum level of VD and low serum levels of zinc among Iranian population aged 10–18 years old [6]. In addition, another study had reported a statistically association between serum level of VD and serum levels of zinc among Iranian pregnant women. Their findings showed 37% VDD and 23% of them had zinc deficiency [8].

Data mining analysis has the potential ability to select patients with VDD based on the possibility of response to treatment against a various factors. Moreover, it may provide a rationale to improve the efficacy of therapy. Similarly, CART analysis recognized several variables which were not associated with response by standard statistical model significantly [18]. The fitted DT model could identify few demographic characteristics such as age and sex as significant factors associated with VDD.

In this study, significant association was found between serum levels of vitamin D3 and BMI. In the different studies on the prevalence of VDD showed that VDD prevalence was various based on regions and Iranian population [1, 16]. In this study, however, VDD was not related with liver function parameters, probably due to that VD serum levels are affected by different factors [11, 13, 17].

One of the limitations of this study was influencing factors such as seasonal variation, diet and geographical habitation on Serum VD levels. We recommend more sensitivity and specificity prediction models, which be able to specify having VDD exactly.

Conclusion

This study provides an easy to use classification rules for classifying risk factors related with VDD that can be useful to improve programs for its management.

Acknowledgement

We would like to thank all the study subjects for their participation, Dentistry Clinical Research Development Center, Birjand University of Medical Sciences, Birjand, Iran for consulting.

Conflict of interest

The author declares that there is no conflict of interest.

Ethics approval

This study was approved by the ethics board committee of Birjand University of Medical Sciences, reference number: IR.BUMS.REC.1398.324.

References

1. Azarkar G., Doosti Z., Osmani F., Ziaee M. Analysis of risk factors for nonalcoholic fatty-liver disease in hepatitis B virus infection: a case-control study. *Hepat. Med., 2019, vol. 11: 153. doi: 10.2147/HMERS211106*
2. Bedossa P., Poynard T. An algorithm for the grading of activity in chronic hepatitis C. *Hepatology, 1996, vol. 24, no. 2, pp. 289–293. doi: 10.1002/hep.510240201*
3. Chen E.-Q., Shi Y., Tang H. New insight of vitamin D in chronic liver diseases. *Hepatobiliary Pancreat. Dis. Int., 2014, vol. 13, no. 6, pp. 580–585.*
4. Coussens A.K., Wilkinson R.J., Hanifa Y., Nikolayevskyy V., Elkington P.T., Islam K., Timms P.M., Venton T.R., Bothamley G.H., Packe G.E., Darmalingam M., Davidson R.N., Milburn H.J., Baker L.V., Barker R.D., Mein C.A., Bhaw-Rosun L., Nuamah R., Young D.B., Drobniewski F.A., Griffiths C.J., Martineau A.R. Vitamin D accelerates resolution of inflammatory responses during tuberculosis treatment. *Proc. Natl Acad. Sci. USA, 2012, vol. 109, no. 38, pp. 15449–15454. doi: 10.1073/pnas.1200072109*
5. Efe C., Kav T., Aydin C., Cengiz M., Imga N.N., Purnak T., Smyk D.S., Torgutalp M., Turhan T., Ozenirler S., Ozaslan E., Bogdanos D.P. Low serum vitamin D levels are associated with severe histological features and poor response to therapy in patients with autoimmune hepatitis. *Dig. Dis. Sci., 2014, vol. 59, no. 12, pp. 3035–3042. doi: 10.1007/s10620-014-3267-3*
6. Farnik H., Bojunga J., Berger A., Allwinn R., Waidmann O., Kronenberger B., Keppler O.T., Zeuzem S., Sarrazin C., Lange C.M. Low vitamin D serum concentration is associated with high levels of hepatitis B virus replication in chronically infected patients. *Hepatology, 2013, vol. 58, no. 4, pp. 1270–1276. doi: 10.1002/hep.26488*
7. Ghaziasadi A., Ziaee M., Norouzi M., Malekzadeh R., Alavian S.M., Saberfar E., Judaki M.A., Ghamari S., Khedive A., Namazi A., Rahimnia R., Jazayeri S.M. The prevalence of hepatitis B virus surface antigen (HBsAg) variations and correlation with the clinical and serologic pictures in chronic carriers from Khorasan Province, North-East of Iran. *Acta Med. Iran, 2012, vol. 50, no. 4, pp. 265–272.*
8. Han J., Pei J., Kamber M. Data mining: concepts and techniques: 3rd ed. *Burlington: Morgan Kaufmann Publishers, 2012, pp. 83–124.*
9. Hauge Matzen L, Christensen J., Hintze H., Schou S., Wenzel A. Diagnostic accuracy of panoramic radiography, stereo-scanography and cone beam C.T. for assessment of mandibular third molars before surgery. *Acta Odontol. Scand., 2013, vol. 71, no. 6, pp. 1391–1398. doi: 10.3109/00016357.2013.764574*
10. Heshmat R., Mohammad K., Majdzadeh S., Forouzanfar M., Bahrami A., Ranjbar Omrani G, Nabipour I., Rajabian R., Hossein-Nezhad A., Rezaei Hemami M., Keshtkar A., Pajouhi M. Vitamin D deficiency in Iran: a multi-center study among different urban areas. *Iran J. Public Health, 2008, vol. 37, no. 1, pp. 72–78.*
11. Hewison M. Vitamin D and immune function: autocrine, paracrine or endocrine? *Scand J. Clin. Lab. Invest., 2012, vol. 72, no. 243, pp. 92–102. doi: 10.3109/00365513.2012.682862*
12. Hoan N.X., Van Tong H, Le Huu Song C.G.M., Velavan T.P. Vitamin D deficiency and hepatitis viruses-associated liver diseases: a literature review. *World J. Gastroenterol., 2018, vol. 24, no. 4: 445. doi: 10.3748/wjg.v24.i4.445*
13. Kitson M.T., Roberts S.K. D-livering the message: the importance of vitamin D status in chronic liver disease. *J. Hepatol., 2012, vol. 57, no. 4, pp. 897–909. doi: 10.1016/j.jhep.2012.04.033*
14. Mahamid M., Nseir W., Elhija O.A., Shteingart S., Mahamid A., Smamra M., Koslowsky B. Normal vitamin D levels are associated with spontaneous hepatitis B surface antigen seroclearance. *World J. Hepatol., 2013, vol. 5, no. 6: 328. doi: 10.4254/wjh.v5.i6.328*
15. Mohamadkhani A., Katoonizadeh A., Poustchi H. Immune-regulatory events in the clearance of HBsAg in chronic hepatitis B: focuses on HLA-DP. *Middle East J. Dig. Dis., 2015, vol. 7, no. 1: 5.*
16. Osmani F. Effect evaluation of vitamin D level amongst patients with chronic hepatitis B. *Arch. Pathol. Clin. Res., 2019, vol. 3, pp. 20–21. doi: 10.29328/journal.apcr.1001014*
17. Osmani F. Problems with reporting accuracy in COVID-19 statistics in Iran. *Gastroenterol. Hepatol. Bed Bench, 2020, vol. 13, no. 4, pp. 275–277.*
18. Osmani F., Hajizadeh E., Rasekhi A., Akbari M.E. Analyzing relationship between local and metastasis relapses with survival of patients with breast cancer: a study using joint frailty model. *Int. J. Cancer Manag., 2018, vol. 11, no. 12: e81783. doi: 10.5812/ijcm.81783*
19. Osmani F., Ziaee M. Assessment of the risk factors for vitamin D3 deficiency in chronic hepatitis B patient using the decision tree learning algorithm in Birjand. *Inform. Med. Unlocked, 2021, vol. 23: 100519. doi: 10.1016/j.imu.2021.100519*
20. Parvaie P., Majd H.S., Ziaee M., Sharifzadeh G., Osmani F. Evaluation of gum health status in hemophilia patients in Birjand (a case-control study). *Am. J. Blood Res., 2020, vol. 10, no. 3: 54.*
21. Plum L.A., DeLuca H.F. Vitamin D, disease and therapeutic opportunities. *Nat. Rev. Drug Discov., 2010, vol. 9, no. 12, pp. 941–955. doi: 10.1038/nrd3318*
22. Remelli F., Vitali A., Zurlo A., Volpato S. Vitamin D deficiency and sarcopenia in older persons. *Nutrients. 2019, vol. 11, no. 12: 2861. doi: 10.3390/nu11122861*
23. Schillie S., Xing J., Murphy T., Hu D. Prevalence of hepatitis B virus infection among persons with diagnosed diabetes mellitus in the United States, 1999–2010. *J. Viral. Hepat., 2012, vol. 19, no. 9, pp. 674–676. doi: 10.1111/j.1365-2893.2012.01616.x*
24. Shoaei S.D., Sali S., Karamipour M., Riahi E. Non-invasive histologic markers of liver disease in patients with chronic hepatitis B. *Hepat. Mon., 2014, vol. 14, no. 2: e14228. doi: 10.5812/hepatmon.14228*

25. Tabrizi R., Moosazadeh M., Akbari M., Dabbaghmanesh M.H., Mohamadkhani M., Asemi Z., Heydari S.T., Akbari M., Lankarani K.B. High prevalence of vitamin D deficiency among Iranian population: a systematic review and meta-analysis. *Iran. J. Med. Sci., 2018 vol. 43, no. 2: 125.*
26. Tayefi M., Saberi-Karimian M., Esmaeili H., Zadeh A.A., Ebrahimi M., Mohebati M., Heidari-Bakavoli A., Azarpajouh M.R., Heshmati M., Safarian M., Nematy M., Parizadeh S.M.R., Ferns G.A., Ghayour-Mobarhan M. Evaluating of associated risk factors of metabolic syndrome by using decision tree. *Comp. Clin. Pathol., 2018, vol. 27, no. 1, pp. 215–223. doi: 10.1007/s00580-017-2580-6*
27. Torresi J., Tran B.M., Christiansen D., Earnest-Silveira L., Schwab R.H.M., Vincan E. HBV-related hepatocarcinogenesis: the role of signalling pathways and innovative ex vivo research models. *BMC Cancer, 2019, vol. 19, no. 1: 707. doi: 10.1186/s12885-019-5916-6*
28. Trépo E., Ouziel R., Pradat P., Momozawa Y., Quertinmont E., Gervy C., Gustot T., Degré D., Vercruysse V., Deltenre P., Verset L., Gulbis B., Franchimont D., Devière J., Lemmers A., Moreno C. Marked 25-hydroxyvitamin D deficiency is associated with poor prognosis in patients with alcoholic liver disease. *J. Hepatol., 2013, vol. 59, no. 2, pp. 344–350. doi: 10.1016/j.jhep.2013.03.024*
29. Tseng T.-C., Kao J.-H. Clinical utility of quantitative HBsAg in natural history and nucleos(t)ide analogue treatment of chronic hepatitis B: new trick of old dog. *J. Gastroenterol., 2013, vol. 48, no. 1, pp. 13–21.*
30. Wang D., Wang Q., Shan F., Liu B., Lu C. Identification of the risk for liver fibrosis on CHB patients using an artificial neural network based on routine and serum markers. *BMC Infect. Dis., 2010, vol. 10, no. 1: 251. doi: 10.1186/1471-2334-10-251*
31. Wong G.L., Chan H.L., Chan H.Y., Tse C.H., Chim A.M., Lo A.O., Wong V.W. Adverse effects of vitamin D deficiency on outcomes of patients with chronic hepatitis B. *Clin. Gastroenterol. Hepatol., 2015, vol. 13, no. 4, pp. 783–790.e1. doi: 10.1016/j.cgh.2014.09.050*
32. Ziaei S., Norrozi M., Faghihzadeh S., Jafarbegloo E. A randomised placebo-controlled trial to determine the effect of iron supplementation on pregnancy outcome in pregnant women with haemoglobin ≥ 13.2 g/dl. *BJOG, 2007, vol. 114, no. 6, pp. 684–688. doi: 10.1111/j.1471-0528.2007.01325.x*

Автор:

Османи Ф., д.н., кафедра биостатистики и эпидемиологии, Исследовательский центр инфекционных заболеваний, Бирджандский университет медицинских наук, г. Бирджанд, Иран.

Author:

Osmani F., PhD, Department of Biostatistics and Epidemiology, Infectious Disease Research Center, Birjand University of Medical Sciences, Birjand, Iran.

Поступила в редакцию 23.11.2020
Принята к печати 10.04.2021

Received 23.11.2020
Accepted 10.04.2021