Multiple Factors of Thyroid Nodule Occurrence in the Elderly-based on Logistic Regression Curves and Random Forest Models

Jiaqi Wang*

Tianjin Medical University, School of Public Health, Tianjin, 300000, China

*Corresponding author: wangjiaqi@tmu.edu.cn

Abstract. This study aims to ascertain the pivotal factors associated with thyroid nodules. To this end, an array of variables was meticulously examined, encompassing age, gender, lifestyle, diet, thyroid markers, and genetics, with a particular emphasis on the elderly population. Two distinct machine learning algorithms were employed for the analysis: logistic regression and a random forest model. The random forest model performs better in predicting thyroid nodule counts, with an accuracy of 0.97 and an AUC of 0.86, substantiating its efficacy in medical decision-making. The ordered multi-categorical regression model also demonstrates high accuracy in predicting nodule counts. The integration of these models is expected to improve diagnostic and therapeutic accuracy. The present study shows a high correlation between bilateral thyroid length, width, thickness, iodine concentration, and gender with the incidence of thyroid nodules. In terms of disease, hypertension and hyperlipidemia have also shown some correlation with thyroid nodules. This study offers a more reliable predictor of the probability of diagnosing thyroid nodules and provides clinicians with more personalized prevention strategies for older adults, thus helping clinicians to make effective diagnoses and treatments. Future studies should explore how to expand the sample size, train and integrate more models to further enhance the management of thyroid nodules, improve patients' quality of life, and prevent nodule enlargement and malignancy.

Keywords: Thyroid nodules; elderly; logistic regression; random forest model.

1. Introduction

The thyroid is crucial to the human body, playing a key role in endocrine, tissue metabolism, neurological effects, and growth and development. Thyroid nodules represent a prevalent disease of the thyroid gland, manifesting as an accumulation of tissue following the unregulated proliferation of thyroid cells. While the majority of thyroid nodules are benign, a small percentage (10-15%) are malignant, which undergo a transformation in nature and become thyroid cancer, which can have a significant impact on the patients' quality of life [1]. In China, the prevalence of thyroid nodules exceeds 20%, and studies have demonstrated a positive correlation between age and the incidence of thyroid nodules, with detection rates in middle-aged and elderly individuals often exceeding 50%.

The high prevalence suggests that the elderly population is particularly susceptible to thyroid diseases [2, 3]. In this demographic, the clinical characteristics of thyroid nodules are more complex due to the presence of a high proportion of thyroid micronodules and atypical symptoms, which hinders effective diagnosis [4]. Furthermore, the decline in physiological function limits treatment options. A high correlation has been identified between the age of thyroid nodule formation, gender, body fat percentage, alcohol consumption, social life factors, and family history. These factors interact with each other to influence the pathogenesis of thyroid nodules [3]. Therefore, complex analyses are essential for an in-depth understanding of the etiology of thyroid nodules, which can help to develop more precise diagnostic and therapeutic strategies.

Numerous studies have investigated to find out what causes thyroid nodules. Sex is important because the prevalence of thyroid nodules is significantly higher in females than in males. This might be because of the way the endocrine system works in females and changes in estrogen levels [5]. Age is also important because the risk of thyroid nodules gradually increases with age, which may be related to the natural aging of thyroid tissues as well as changes in the body's metabolic function [6].

Lifestyle factors also play an important role in the development of thyroid nodules. Smoking and alcohol consumption are thought to increase the risk of developing thyroid nodules by affecting the metabolism of thyroid hormones and disrupting the micro-environment of thyroid tissue [7]. The consumption of specific dietary components has been demonstrated to exert a considerable influence on the propensity to develop thyroid nodules. Iodine, an essential component of a balanced diet, plays a pivotal role in ensuring optimal function of the thyroid gland. Both iodine deficiency and excess can alter thyroid morphology and contribute to nodule development [8, 9].

The data for this study were collected from an iodine nutrition survey conducted in a hospital in North China. The objective of this study is to thoroughly analyze the relationship between various factors, including age, gender, lifestyle, dietary habits, thyroid-related indicators, genetic factors, and the development of thyroid nodules. This comprehensive analysis aims to identify the key determinants of thyroid nodule occurrence, particularly in the elderly population. To achieve this objective, two models will be employed: the logistic regression model and the random forest model. In previous studies, logistic regression models have been applied to analyze risk factors for a range of health problems. These include the identification of gut microbes and metabolites associated with colorectal cancer (CRC), as well as the analysis of risk factors for sports injuries in college students. This model effectively identifies independent variables associated with specific diseases and quantifies their impacts [10, 11]. The findings of this study facilitate a more profound comprehension of the etiology of thyroid nodules, thereby underpinning the development of more personalized prevention strategies for the elderly. They may also help clinicians to diagnose, assess, and treat thyroid nodules more effectively and improve prevention and treatment. Conversely, random forests are widely used in predictive analysis, including machine learning, meta-analysis, and systematic reviews. These models have historically demonstrated robust performance; however, it has been observed that the random forest model Area Under the Curve (AUC) and other performance metrics (e.g., sensitivity and specificity) have exhibited variability across different studies. Consequently, there is an opportunity to calibrate the model to achieve the desired performance balance, with the overarching objective being the suitability of its efficacy [12-14].

In this study, the former will be used to elucidate the impact of each factor on the development of thyroid nodules and to quantify their respective risks. The Random Forest model will facilitate a more nuanced understanding of the complex processes that lead to the development of thyroid nodules by taking advantage of integrated learning, leading to more precise diagnostic and therapeutic strategies. This study contributes to a deeper understanding of the etiology of thyroid nodules, thereby supporting the development of more personalized prevention strategies for older adults. The model may also assist clinicians in the more effective diagnosis, assessment, and treatment of thyroid nodules, thereby enhancing prevention and treatment.

2. Methods

2.1. Data Source and Description

The dataset used in this study encompasses a range of factors that have the potential to be associated with the occurrence of thyroid nodules in elderly individuals. The data were obtained from the Iodine Nutrition Survey, which was conducted in 2023 at a tertiary care hospital in Tianjin, China. The authors were directly involved in the survey to ensure the reliability and accuracy of the data.

A total of 515 samples were collected, and a total of 30 variables, covering demographic details such as age and gender, as well as lifestyle factors such as smoking and alcohol consumption, will be analyzed in this dataset. The dietary habits of the participants will also be analyzed, including the type and amount of water and salt intake. Additionally, health indicators such as sleep patterns, physical activity levels, incidence of thyroid disease, family history of thyroid disease, cohabitation history of thyroid disease, diabetes, and hyperthyroidism will be included. Before the analysis, the data set underwent rigorous processing to ensure its integrity and to identify and remove any anomalies or outliers.

2.2. Indicator Selection and Description

This data set encompasses 515 meticulously selected observations and 22 distinct variables (in Table 1). To facilitate a comprehensive study, a new variable was added to the dataset. This new variable is referred to as 'total number of nodes' and is defined by summing the number of nodules on the left and right thyroid lobes, thus providing a single indicator that summarizes the overall nodule burden for each patient. The incorporation of this new variable has the potential to streamline subsequent analyses, thereby facilitating a more comprehensive evaluation of the prevalence of thyroid nodules and their potential correlation with other variables. The variables were divided into five groups to facilitate subsequent analysis and summary of the experiment. The five groups are personal information, lifestyle habits, diseases, thyroid indicators, and iodine levels.

Table 1. Variable Explanation

Variable Name	Symbol	Value
sex	P1	0/1, Female=0,Male=1
age	P2	49-86(year)
smoking	L1	0/1, Nonsmoker=0,Smoker=1
drinking	L2	0/1,Non-alcoholic=0,Alcoholic=1
Amount of water	L3	100-5003(ml/d)
Amount of salt	L4	1-12(g/d)
sleeping	L5	1/2/3/4/5h/d
practising	L6	1/2/3/4, Not exercising=1, Weak (mean, no sweat)=2, Fair (appropriate heart rate, slight sweat)=3, More tired (sweating profusely)=4
Thyroid disease	D1	0/1, Don’t have disease=0, Have disease=1
Family thyroid disease	D2	0/1, Don’t have disease=0, Have disease=1
diabetic	D3	0/1, Don’t have disease=0, Have disease=1
hypertensive	D4	0/1, Don’t have disease=0, Have disease=1
hyperlipidemia	D5	0/1, Don’t have disease=0, Have disease=1
hyperuricaemia	D6	0/1, Don’t have disease=0, Have disease=1
Left thyroid1	T1	0-3.34cm
Left thyroid2	T2	0-4.7cm
Left thyroid3	T3	0-6.63cm
Left nodules	T4	0/1, Don’t have nodule=0, Have nodule=1
Left nodules amount	T5	0/1/2/3/4
Right thyroid1	T6	0-3.91cm
Right thyroid2	T7	0-6.64cm
Right thyroid3	T8	0-3.1cm
Right nodules	T9	0/1, Don’t have nodule=0, Have nodule=1
Right nodules amount	T10	0/1, Don’t have nodule=0, Have nodule=1
Thyroid volume	T11	0-24.98cm3
General iodine	I1	6.12-1637.50 μ g/l
Disassociate iodine	I2	3.81-778.86 μ g/l

2.3. Method Introduction

This study assembled a dataset encompassing multiple thyroid-related variables. Before data analysis, a comprehensive pre-processing procedure was carried out. First, missing values were detected and handled. For numerical variables, missing values were filled with their means, while categorical variables were replaced with the most frequent categories. Subsequently, data standardization was performed. This transformed data with varying scales into a unified scale, effectively mitigating the potential influence of scale differences on model performance.

A random forest model was employed to predict the number of nodules in patients. The dataset was partitioned into training and testing subsets. Specifically, 80% of the samples were allocated to the training set for model training, whereas the remaining 20% constituted the testing set for evaluating the performance of the random forest model. Model efficacy was evaluated via a Receiver Operating Characteristic (ROC) curve. The True Positive Rate (TPR) and False Positive Rate (FPR) were plotted across different thresholds to visualize performance. Concurrently, the Area Under the Curve (AUC) was calculated. A higher AUC, approaching 1, indicated enhanced model accuracy. To further explore the prediction accuracy of nodule numbers, an ordered multiple regression model was constructed. Using prediction accuracy as an assessment metric, the model predicted nodule numbers at five levels (0, 1, 2, 3, 4). Mathematically, its prediction process was defined as:

$$logit[P(X)] = a_j + \beta_1 \cdot X_1 + \cdots + \beta_P \cdot X_P \tag{1}$$

Finally, the confusion matrix function was used to assess the model's classification performance. Multidimensional indicators, such as accuracy, Kappa value, sensitivity, and specificity, were computed for comprehensive evaluation.

3. Results and Discussion

3.1. Correlation Analysis

The heatmap was created for each variable in the dataset. The heatmap had two objectives: first, to visualize the relationship between each variable and the occurrence of nodules, and second, to filter out variables with high correlation for subsequent modeling. Figure 1 shows each variable's correlation, it shows that there's a high correlation between I1 and I2, T4 and T5, T9 and T10, and so on.

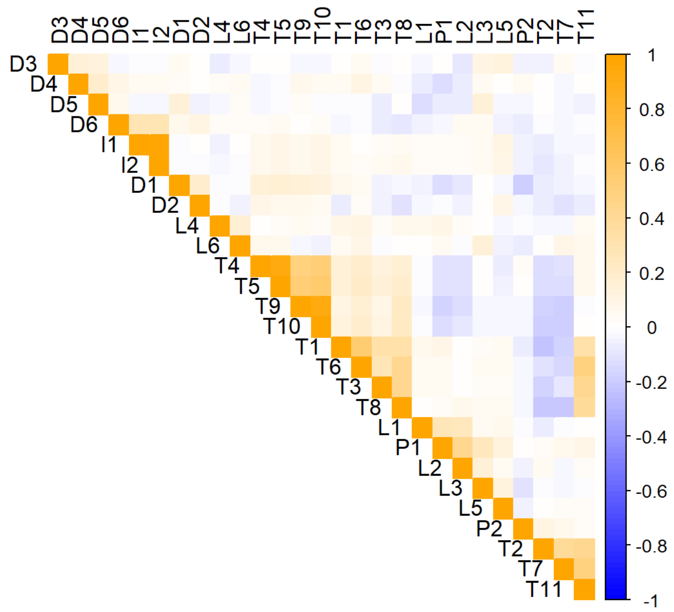

Fig. 1 Variables correlation heatmap

Bar charts were created to show comparisons of the number of men and women with different nodules. Figure 2 shows the general number of thyroid nodules and compares it with sex. The 0 represents female while 1 represents male, so this data set contains more female samples than male, and females have more nodules than males in this study.

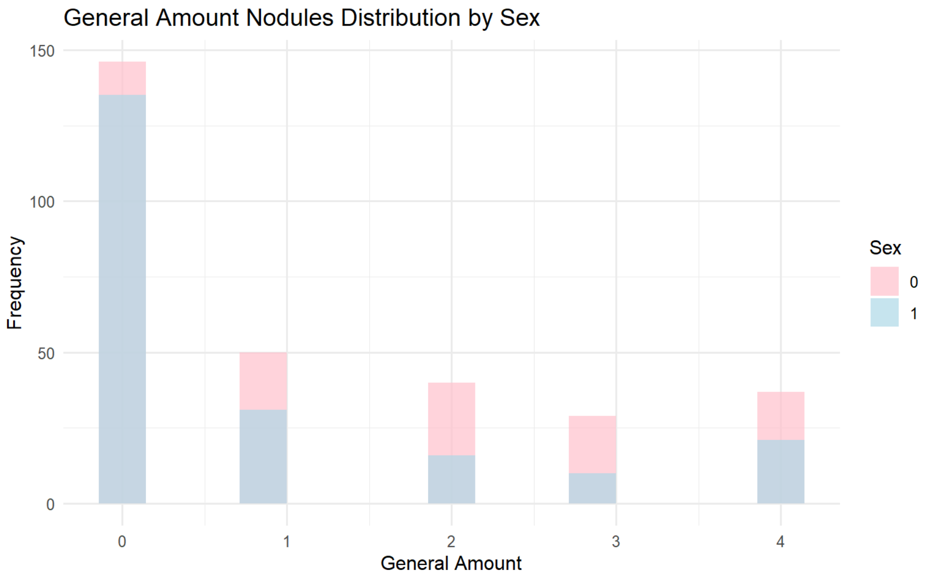

Fig. 2 Nodules analysis

3.2. Random Forest Analysis

The random tree model demonstrates that thyroid-related indexes and iodine content indexes serve as effective decision-making indexes for nodule screening. This model justifies the number of thyroid nodules by the process of Figure 3 and presents a superior performance in predicting thyroid nodule counts, with an accuracy of 0.97 and a P value less than 0.001.

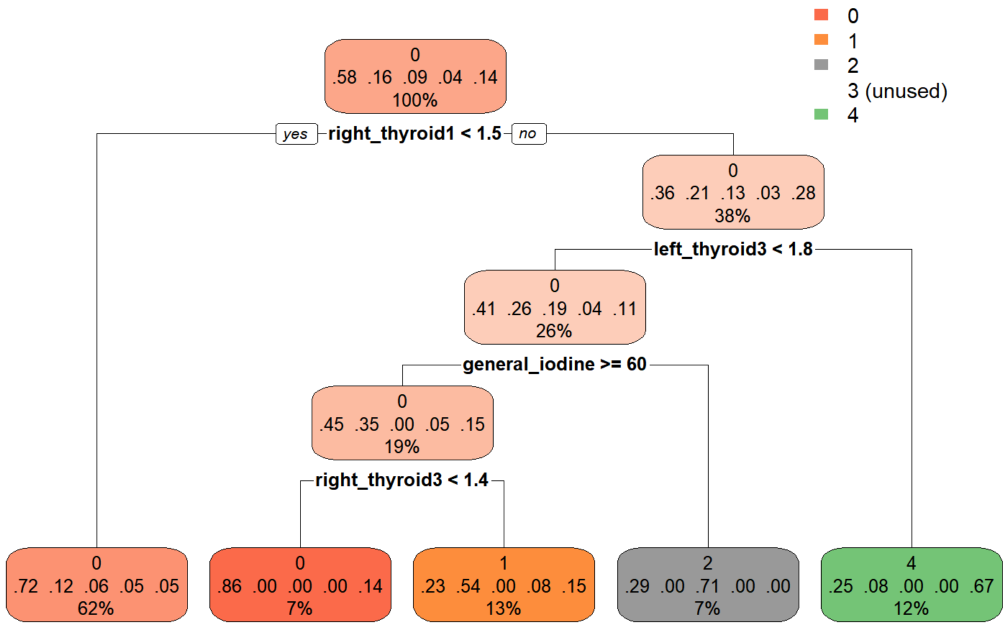

Fig. 3 Random forest model

With a well-structured design and informative node details, the model provides a reliable basis for accurate decision-making. Figure 4 shows that the area under the curve (AUC) of the ROC curve reaches 0.86, substantiating the model's efficacy in predicting thyroid nodules. The high AUC value, along with the clear visualization of the ROC curve (where the x-axis represents specificity and the y-axis represents sensitivity), indicates the model's high efficiency in predicting the number of thyroid nodules using dataset variables.

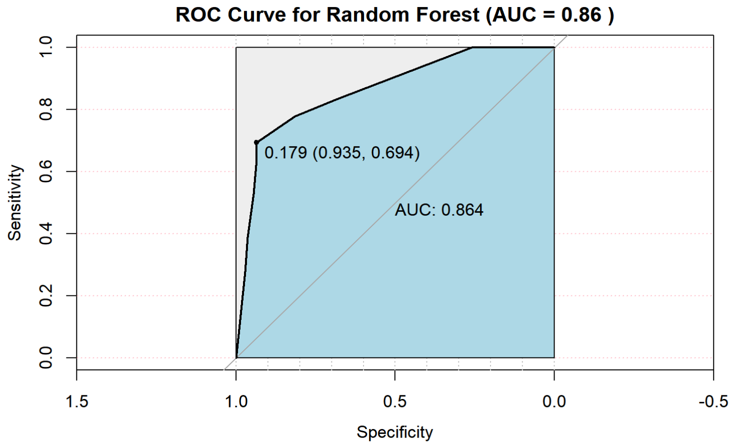

Fig. 4 Random forest ROC curve

Furthermore, as illustrated in Table 2, the model demonstrates significant judgmental sensitivity in each node, exhibiting a notable capacity to discriminate between positive and negative scenarios.

Table 2. Performance Evaluations of Random Forest

	Class 0	Class1	Class2	Class3	Class4
Sensitivity	1.00	1.00	1.00	0.69	1.00
Specificity	1.00	1.00	0.97	1.00	1.00
Positive Prediction Value	1.00	1.00	0.75	1.00	1.00
Negative Prediction Value	1.00	1.00	1.00	0.97	1.00
Prevalence	0.57	0.17	0.07	0.08	0.10
Detection Rate	0.57	0.17	0.07	0.05	0.10
Detection Prevalence	0.57	0.17	0.10	0.05	0.10
Balanced Accuracy	1.00	1.00	0.99	0.84	1.00

Finally, the correlation between the variables and the development of thyroid nodules in this experiment was analyzed through the random forest model, and Figure 5, a bar graph was made for comparison.

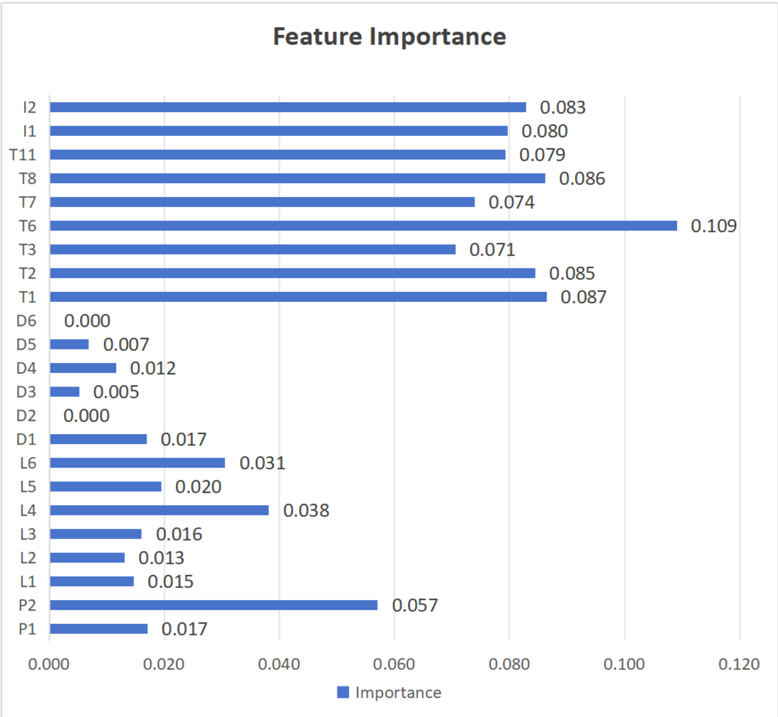

Fig. 5 Feature importance of variables

3.3. Ordinal Logistic Regression Analysis

When exploring multiple factors contributing to thyroid nodule occurrence in the elderly using logistic regression and random tree models. They achieve an 88.29% accuracy rate, correctly predicting thyroid nodule occurrences in the elderly. The Kappa value of 0.823 validates their reliability, demonstrating consistency between predictions and actual observations. The P - P-value, less than 0.001 when comparing accuracy against the null hypothesis (NIR), indicates the models' predictive capabilities far surpass random guessing. The evaluation data above validate their effectiveness in identifying factors related to thyroid nodule occurrence in elderly people (Table 3).

Table 3. Performance Evaluations of Ordinal Logistic Regression

	Class 0	Class1	Class2	Class3	Class4
Sensitivity	1.00	0.94	0.36	0.50	1.00
Specificity	1.00	0.93	0.98	0.98	0.97
Positive Prediction Value	1.00	0.76	0.67	0.73	0.80
Negative Prediction Value	1.00	0.99	0.93	0.96	1.00
Prevalence	0.52	0.18	0.11	0.08	0.12
Detection Rate	0.52	0.17	0.04	0.04	0.12
Detection Prevalence	0.52	0.22	0.06	0.05	0.15
Balanced Accuracy	1.00	0.94	0.67	0.74	0.98

The ordinal logistic regression model, when applied to the data, also demonstrated a high degree of efficacy in predicting the number of thyroid nodules present. When predicting the number of 0 nodules, the model demonstrated an accuracy of 1, indicating its reliability in cases where no nodules are present. When predicting the number of 1 nodule, the accuracy was 0.94, suggesting the model's capacity for accurate prediction in more prevalent cases. However, the accuracy reduced to 0.67 and 0.74 for 2 and 3 nodules, respectively, yet the model maintained a superior prediction ability. Despite the lower accuracy of 0.67 and 0.74 for two and three nodules, respectively, a favorable prediction trend was observed. Notably, the accuracy reached 0.99 for four nodules, signifying the model's robust predictive capacity, particularly for cases involving four nodules.

4. Conclusion

The study highlights the strengths of both the random forest and ordered multinomial regression models in predicting thyroid nodule counts. The random forest model demonstrates advanced feature learning and nonlinear combination techniques, showcasing strong predictive power, particularly in identifying nodules. This capability is crucial for early detection and intervention. In contrast, the ordered multinomial regression model excels in predicting the exact number of nodules, providing detailed insights into the relationship between various factors and nodule counts. This information is valuable for tailoring preventive strategies and clinical management. Furthermore, the findings are significant for developing personalized preventive strategies for thyroid nodules in the elderly. Identifying key risk factors aids clinicians in making informed decisions and enhancing diagnostic accuracy and patient outcomes. The models' ability to accurately predict nodule counts can lead to earlier interventions and improved management of thyroid conditions. The present study's correlation analysis of variables revealed that gender disparities are associated with the development of thyroid nodules, a finding that aligns with the outcomes of a prior study. However, the correlation between age and the presence of thyroid nodules was not particularly strong in this experiment, which may be attributed to the limited age range of the sample population in the dataset. In terms of lifestyle, salt intake and exercise intensity exhibited a higher correlation with the number of thyroid nodules and a lower correlation with the disease. The relationship between hypertension and thyroid nodules merits further attention. Conversely, familial thyroid disease and hyperuricemia demonstrated a stronger correlation with thyroid nodules in this study. The three-dimensional parameters of the thyroid gland, iodine levels, and the number of thyroid nodules showed a high degree of correlation. These findings establish a foundational principle for subsequent efforts aimed at the prevention and treatment of thyroid nodules.

However, the study has some limitations. The dataset's size and diversity may restrict the models' generalizability, necessitating larger and more diverse datasets for broader applications. Additionally, the models' complexity and interpretability require further refinement to enhance their practicality in clinical settings. Addressing these limitations will be crucial for improving the models' reliability and usability in real-world applications. Future research should focus on these areas to enhance the overall effectiveness of the models.

References

[1] Dean D S, Gharib H. Epidemiology of thyroid nodules. Best Pract Res Clin Endocrinol Metab, 2008, 22(6): 901-911.

[2] Zhang Wenxia, Yang Xiaoyan. Epidemiological Investigation of Thyroid Nodules in 1,956 Middle - aged and Elderly People in Physical Examinations. Chinese Journal of Geriatric Care, 2024, 22(02): 84-86.

[3] Zhang Lin, Wang Xiaoyan, Li Yingna, et al. Study on the Relationship between the Characteristics of Thyroid Nodules and Thyroid Function in the Elderly. China Preventive Medicine, 2011, 12(11): 977-978.

[4] Zhao Lingbai, Liu Yingzhe. Research Progress of Traditional Chinese Medicine in the Treatment of Thyroid Nodules. Advances in Clinical Medicine, 2024, 14.

[5] Li X, Chen Z, Wu L, Tu P, Mo Z, Xing M. Prevalence of thyroid nodules and its associations with physiological and psychosocial factors among adults in Zhejiang Province, China: A baseline survey of a cohort study. BMC Public Health, 2024, 24(1): 1854.

[6] Yeo Y, Shin D W, Han K, Kim D, Kim T H, Chun S, Jeong S M, Song Y M. Smoking, Alcohol Consumption, and the Risk of Thyroid Cancer: A Population-Based Korean Cohort Study of 10 Million People. Thyroid, 2022, 32(4): 440-448.

[7] Men Cheng, Wang Linqi, Zhao Di, et al. Iodine Nutrition and Detection of Thyroid Nodules and Influencing Factors among Residents in Health Examinations. Journal of Medical Forum, 2023, 44(23): 61-64+68.

[8] Kaloumenou I, Alevizaki M, Ladopoulos C, Antoniou A, Duntas LH, Mastorakos G, Chiotis D, Mengreli C, Livadas S, Xekouki P, Dacou-Voutetakis C. Thyroid volume and echostructure in schoolchildren living in an iodine-replete area: relation to age, pubertal stage, and body mass index. Thyroid, 2007, 17(9): 875-881.

[9] Yuan Pei, Qin Yuqin. Analysis of Influencing Factors of Low Birth Weight Infants Based on Decision Tree and Logistic Regression. Practical Preventive Medicine, 2025, 32(02): 192-196.

[10] Sinha R, Ahn J, Sampson J N, Shi J, Yu G, Xiong X, Hayes R B, Goedert J J. Fecal Microbiota, Fecal Metabolome, and Colorectal Cancer Interrelations. PLoS One, 2016, 11(3): 152126.

[11] Yuan Mengzhe, Bai Jiali, Guo Hui. Logistic Regression Analysis of Risk Factors for Sports Injuries Among College Students. Journal of Medical Frontiers, 2023, 13(1): 22-24.

[12] Mangold C, Zoretic S, Thallapureddy K, Moreira A, Chorath K, Moreira A. Machine Learning Models for Predicting Neonatal Mortality: A Systematic Review. Neonatology, 2021, 118(4): 394-405.

[13] Akbari S, Khodadadi B, Ahmadi S A Y, Abbaszadeh S, Shahsavar F. Association of vitamin D level and vitamin D deficiency with risk of preeclampsia: A systematic review and updated meta-analysis. Taiwan J Obstet Gynecol, 2018, 57(2): 241-247.

[14] Kent S T, Bromfield S G, Burkholder G A, Falzon L, Oparil S, Overton E T, Mugavero M J, Schwartz J E, Shimbo D, Muntner P. Ambulatory Blood Pressure Monitoring in Individuals with HIV: A Systematic Review and Meta-Analysis. PLoS One, 2016, 11(2): 148920.