Mashhad University of Medical Sciences (MUMS)

Reviews in Clinical Medicine

Clinical Research Development Center Ghaem Hospital

Prediction of Tuberculosis Using a Logistic Regression Model

Kiarash Ghazvini (Ph.D)[1]*, Shamsoddin Mansouri (MSc)[1], Mohammad-Taghi Shakeri (Ph.D)[2], Masoud Youssefi (Ph.D)[1], Mohammad Derakhshan (Ph.D)[1], Masoud Keikha (Ph.D)[1]

[1]*Department of Microbiology and Virology, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.*
[2]*Department of Biostatistics and Epidemiology, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.*

ARTICLE INFO

Article type
Original article

Article history
Received: 12 Jul 2019
Revised: 23 Jul 2019
Accepted: 21 Aug 2019

Keywords
Influential Factors
Logistic Regression
Tuberculosis

ABSTRACT

Introduction: Tuberculosis (TB) is a chronic bacterial disease and a leading cause of mortality among single-agent infectious diseases following the human immunodeficiency virus infection across the world. Logistic regression is a method of statistical analysis with predictive capability. This multivariate statistical method could be used to evaluate the correlations between independent variables (albeit confounding) and a dependent variable. The present study aimed to assess the influential factors in the incidence of TB based on the estimations of a logistic regression predictive model.

Methods: This cross-sectional study was conducted on two groups consisting of 189 TB patients and 189 controls. The influential factors in TB were compared between the groups, including age, gender, marital status, risk of acquired immunodeficiency syndrome (AIDS), smoking habits, history of asthma, organ transplantation, body mass index (BMI), vitamin D3 level, diabetes, and rate of hemoglobin and malignant diseases. In addition, the predictive potential of the logistic regression model was determined based on various indices, such as sensitivity, specificity, and receiver operating characteristic (ROC) curve.

Results: The sensitivity and specificity of the regression model were estimated at 78% and 68%, respectively, and the area under the ROC curve was calculated to be 0.821. Among the available influential factors in the dependent variable (i.e., TB), the variables of vitamin D3 and hemoglobin levels and BMI were considered significant.

Conclusion: According to the results, the logistic regression model is appropriate for the prediction of TB considering the accuracy and predictive power of its criteria, as well as the area under the ROC curve (0.821), which could provide the test accuracy for the diagnosis TB.

Please cite this paper as:
Ghazvini K, Mansouri S, Shakeri MT, Youssefi M, Derakhshan M, Keikha M. Prediction of Tuberculosis Using a Logistic Regression Model. Rev Clin Med. 2019;6(3):108-112.

Introduction

Tuberculosis (TB) is a long-standing infectious disease and a leading cause of mortality following human immunodeficiency virus (HIV) infection. TB is commonly referred to as the 'white plague' (1). According to the World Health Organization (WHO), the mortality rate of TB was approximately 1.7 million only in 2017 (1).

Iran is a developing country located in the central Asia, which has a critical status in terms of the prevalence of TB due to its neighboring countries, such as Azerbaijan and Armenia (countries with high rate of multidrug-resistant tuberculosis [MDR-TB]), as well as Afghanistan and Pakistan (countries with high prevalence of TB). Consid-

***Corresponding author:** Kiarash Ghazvini.*
Department of Microbiology and Virology, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.
***E-mail:** Ghazvinik@mums.ac.ir*
***Tel:** 989151248938*

ering the endemic nature of TB in Iran, regional programs for the screening and identification of the patients are essential (2). According to recent reports, the prevalence of pulmonary TB in Iran is approximately 22 cases per 100,000, while the associated mortality rate has been estimated at 3.5 cases per 100,000 (3).

Statistics show that almost one-third of the world's population is infected with Mycobacterium TB, and the symptoms of TB have progressed gradually in 5-10% of the patient population, exposing it to the high risk of active TB infection. Therefore, identification of accurate mechanisms and developing models based on the prediction of TB play a pivotal role in TB control programs, as well as the provision of prophylaxis treatment for the predisposed individuals and preventing the spread of the disease in the community (4,5).

In general, smear staining and sputum culture are considered to be the initial steps in the examination of the patients suspected of pulmonary TB. However, not all patients are able to provide proper samples for sputum examination, and in some cases, the amount of the excreted bacillus in the sputum of the patient is insufficient for observation in the sputum smear (6). In addition, Mycobacterium is a slow-growing organism, the culture of which in the Lowenstein-Jensen medium ('gold standard' diagnostic method) requires 3-8 weeks to be performed, while the use of modern culture techniques for clinical specimens in the BACTEC medium reduces the time to 9-16 days (7). Moreover, the increased prevalence of HIV infection, widespread use of steroids, and growing number of organ transplant recipients have contributed to the increased cases of pulmonary, smear-negative, and extrapulmonary TB in recent years (8).

Use of simple, rapid method without the need for specific or invasive sampling methods, as well as the necessity of acceptable accurate diagnosis, seems essential considering the need for the rapid isolation of TB patients, achieving definitive diagnosis within the shortest time, the initiation of prophylaxis depending on TB cases in order to obtain samples for smear and culture involving invasive procedures (pulmonary smear-negative patients), conditions with reduced possibility to access diagnostic samples (e.g., diffuse TB), and cases in which the clinical conditions of the patients does not allow the implementation of invasive methods (e.g., reduced consciousness) (9). In this regard, the emergence and development of molecular diagnostic methods have improved the diagnosis of TB. These techniques are able to diagnose TB within less than one day. Furthermore, they are highly sensitive and could detect the smallest levels of microorganisms in suspected samples (10 microorganisms per one milliliter of the sputum sample). However, one of the most important limitations of these methods is reduced sensitivity and specificity in the diagnosis of negative-smear samples. In addition, these methods cannot be used for the prediction of disease progress and treatment monitoring (patient follow-up) (10). With the advancement of computer methods in the modern era, these methods are considered to be a new approach to the diagnosis and prediction of TB. Although the sensitivity and specificity of these techniques are not limited to culture and polymerase chain reaction (PCR), the cost-efficient computer methods that provide simple and rapid analysis and easy access are among the optimal candidates for the early diagnosis of TB, as well as the identification of susceptible individuals and primary screening of the patients (5,11).

In the mentioned methods, regression correlations or mathematical formulation are inferred from the influential factors or risk factors of TB, such as age, smoking habits, organ transplantation, renal and dialysis disorders, HIV infection, hemoglobin levels, vitamin D, and weight loss. Based on these variables, the methods could predict the risk of TB in the community members with a history of TB. For instance, Aguiars et al. proposed the tree regression model to predict pulmonary TB in the patients based on the risk factors of smoking habits, weight loss, AIDS, alcohol consumption, and pulmonary radiographic findings, which could predict TB in a population exposed to these risk factors (12).

The present study aimed to develop and propose an appropriate statistical method for the diagnosis of TB.

Methods

This cross-sectional study was conducted on 189 patients with active TB and 189 healthy individuals (controls) referring to the central laboratories for TB diagnosis in Mashhad and Tehran, Iran. The subjects were randomly selected and enrolled in the study during 2016-2018.

Data of the patients and controls were collected, including age, gender, marital status, residence status, body mass index (BMI), diabetes, AIDS and other immunodeficiency diseases, smoking habits, cancer, organ transplantation, renal and dialysis disorders, history of asthma, and levels of hemoglobin and serum vitamin D. The obtained data were analyzed in order to determine the correlations between the independent variables using Chi-square and Fisher's exact test.

At the next stage, a logistic regression model was proposed based on the variables that were significantly correlated with TB, explaining the

correlations between the response variable (dependent variable) and a set of predictive variables (independent variables) in order to determine the effect of each predictive variable on TB. In addition to the modeling of the data, another advantage of the proposed logistic regression model was the prediction of the probability of belonging to each levels of the dependent variable, as well as the possibility of the direct calculation of the odds ratio based on the coefficients of the model.

Results

Among 189 patients with pulmonary TB, 53.4% were male (mean age: 51.9 years), and 46.6% were female (mean age: 48.6 years). In the control group, 51.8% of the subjects were male (mean age: 51.5 years), and 48.2% were female (mean age: 49.7 years). The frequency of each data was calculated in the patients and controls, and the frequency of the risk factors of smoking habits, HIV virus, transplant recipients, cancer, diabetes, history of asthma and dialysis in the patients was observed to be higher compared to the controls (Figure 1 & 2).

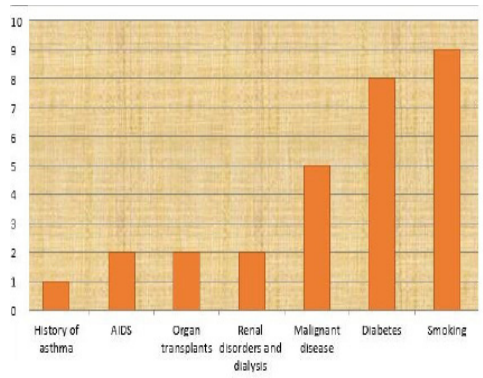

Figure 1. The characteristics of the effect of subject risk factors in tuberculosis patients.

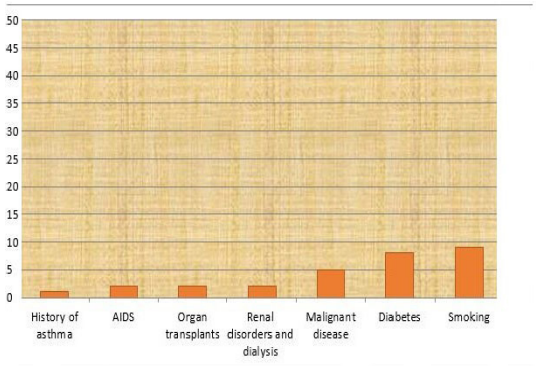

Figure 2: The characteristics of the effect of subject risk factors in control group

In a detailed analysis, the influential factors in TB were investigated using Chi-square in order to accurately determine their significance, and the three variables of hemoglobin level, BMI, and serum vitamin D were observed to be significantly correlated with TB. Although the other variables were considered to be important risk factors for TB, they were eliminated from further evaluation due to their insignificance (Table 1).

Table 1. Chi-square Results Regarding Risk Factors in TB Patients and Controls.

Variables	TB Patients (%) N	Controls (%) N	P-value
Smoking Habits	84 (44.4)	14 (7.4)	0.001
Diabetes	20 (10.6)	11 (5.8)	0.133
Malignancies	10 (5.3)	4 (2.1)	0.171
Renal and Dialysis Disorders	2 (1.1)	1 (0.5)	1
Organ Transplant	1 (0.5)	1 (0.5)	1
AIDS	1 (0.5)	0 (0.0)	1
History of Asthma	2 (1.1)	0 (0.0)	0.499
Body Mass Index (≤19 kg/m²)	132 (69.8)	45 (23.8)	0.001
Vitamin D Deficiency (≤25 nmol/l)	108 (57.1)	63 (33.3)	0.001
Hemoglobin Level (≥10 g/dl)	55 (29.1)	17 (9.0)	0.001

Comparison of the variables of BMI, serum vitamin D, and hemoglobin level between the TB patients and controls also confirmed this proposition (Figure 3).

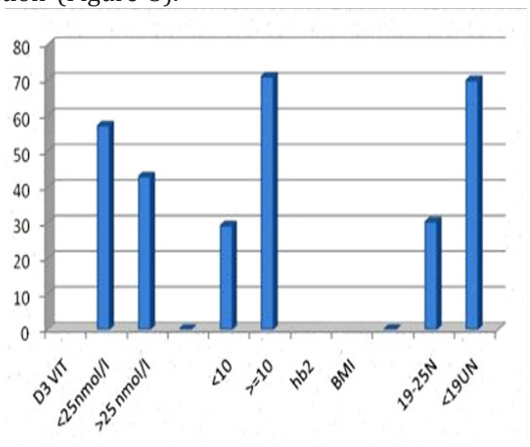

Figure 3. Comparison of frequency of vitamin D3, hemoglobin and body mass index variables in both TB-patient and control groups.

According to the obtained results, these variables alone had significant effects on TB. Furthermore, the outcome of these variables was considered significant in the proposed predictive model. Therefore, the prediction of TB was based on the proposed model (Figure 4).

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I... Lower
Step 1[a]	D3.vit	1.494	.417	12.809	1	.000	4.455	1.966
	BMI	-4.197	.460	83.068	1	.000	.015	.006
	hb2	3.080	.486	40.083	1	.000	21.758	8.385
	Constant	-1.438	.594	5.859	1	.015	.237	

Figure 4. The software calculation of the prediction model of tuberculosis

Logit (pi)=0.684 + 1.49 (Vitamin D) + 3.08 (Hemoglobin) – 4.197 (BMI)

The sensitivity and specificity of the proposed model were estimated at 78% and 68%, respectively, and the area under the receiver operating characteristic (ROC) curve was calculated to be 0.821 (Figure 5).

Considering the precision and predictive power criteria and area under the ROC curve, which provided overall accuracy of the test for the diagnosis of TB properly, it could be stated that the logistic regression model was appropriate for the prediction of TB.

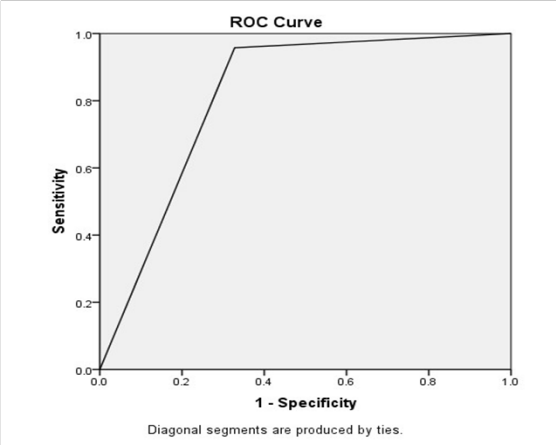

Figure 5. the calculation of sensitivity and specificity for the prediction model of tuberculosis in the ROC curve.

Discussion

Given the importance of TB and its high prevalence in various communities, the prediction of the disease is essential. For many years, direct smear staining of clinical specimens was the only rapid diagnostic tool for TB. This method has very high specificity (95%) in the HIV-negative patients with pulmonary TB, while its sensitivity has been reported to be 50-70% in several studies (13). Moreover, direct smear staining results may be false negative in some cases (14). Despite the acceptable specificity and sensitivity of direct smear staining and culture, these methods are time-consuming, and due to the slow growth of organisms, the isolation of Mycobacterium TB from the culture medium may take weeks. In addition, determining the species and its sensitivity to various antibiotics may require more time to achieve definitive diagnosis (15).

Although molecular methods are comparatively faster, they are not widely used in developing countries, especially Iran, due to their lower culture susceptibility and high costs (16). Currently, one-third of the world's population has Mycobacterium TB, with approximately 5-10% affected by active TB in specific conditions (e.g., poor immune function). Considering the contagious nature of TB, development of a model for the prediction of TB is considered to be an essential measure for the control of the disease (17).

In the present study, a combination of three risk factors for TB was proposed, based on which the input data of the model could be predicted on the basis of zero and one (i.e., healthy or diseased), and the sensitivity and specificity were calculated to be 78% and 68%, respectively. Moreover, the Kappa coefficient indicated the agreement rate to be average. The area under the ROC curve also confirmed the propriety of the proposed model. Therefore, it could be concluded that the logistic regression model had proper accuracy and predictive power, directly expressing the risk ratio of TB.

To date, few studies have been conducted in this regard. In Brazil, Fernanda Carvalho et al. (2006) assessed negative-smear patients with a history of alcohol and drug addiction and diabetes together with patients with TB and HIV infection as positive and negative TB, reporting the sensitivity of 64-71% and specificity of 58-76% based on variables such as chest X-ray, sputum, weight loss, and age using a tree regression model. The researchers concluded that in the identification of TB, the negative-smear patients could be identified with less costly methods than PCR and culture with positive or negative TB ratio. However, the evaluation was performed in a specific group with the considered variables (18).

In a similar study, Fabio S. Aguiar et al. (2012) incorporated factors such as smoking habits, weight loss, AIDS, alcohol consumption, and pulmonary radiography findings into a tree regression model for the prediction of pulmonary TB in hospitalized patients, denoting the correlations of these risk factors with TB. Therefore, the model was considered appropriate to facilitate the distinguishing of

TB patients by the attending physician for isolation, treatment, and prevention purposes in order to control the disease transfer (12).

In another research performed by K. Ladefoged et al. in Greenland, factors such as smoking habits, BMI, accompanied diseases, and alcohol consumption were considered to be the main risk factors for TB. In addition, the correlations of these factors with TB were confirmed through the comparison of TB patients with the control group (19). In another study, Salpeter et al. (1996) developed a mathematical model for the epidemiological studies of TB based on the delayed infection rate and recurrent infection, as well as the variables of latent infection, re-infection, and disease. Considering the increased population during 1930-1995 in the United States, it was concluded that the emergence rate of a new TB infection is 80,000 per year, while the incidence rate of latent infection was estimated at 5%, and the transfer rate of the infection in each individual with active infection was reported to be 5% (20).

Conclusion

Based on statistical analysis, a logistic regression model was developed in the present study based on factors such as serum vitamin D, serum hemoglobin, and BMI in order to predict the prevalence of TB with acceptable specificity and sensitivity. According to the findings, the risk factors of smoking habits, asthma, renal disorders, and some chronic diseases contributed to the progress of latent TB infection to active TB. However, this issue was not further investigated by the model due to the lack significance.

Acknowledgements

None.

Conflict of Interest

The authors declare no conflict of interest.

References

1. World Health Organization. Global tuberculosis report 2018. World Health Organization; 2018.
2. Hoffner S, Hadadi M, Rajaei E, et al. Geographic characterization of the tuberculosis epidemiology in iran using a geographical information system. Biomed Biotechnol Res J 2018;2:213-219.
3. Riyahi Zaniani F, Moghim S, Mirhendi H, et al. Genetic Lineages of Mycobacterium Tuberculosis Isolates in Isfahan, Iran. Curr Microbiol. 2017;74:14-21.
4. World Health Organization. Latent tuberculosis infection: updated and consolidated guidelines for programmatic management. World Health Organization; 2018.
5. Jurcev-Savicevic A, Mulic R, Ban B, et al. Risk factors for pulmonary tuberculosis in Croatia: a matched case–control study. BMC Public Health. 2013;13:991.
6. Roy M, Muyindike W, Vijayan T, et al. Implementation and Operational Research: Use of Symptom Screening and Sputum Microscopy Testing for Active Tuberculosis Case Detection Among HIV-Infected Patients in Real-World Clinical Practice in Uganda. J Acquir Immune Defic Syndr. 2016;72:e86-91.
7. Trivedi MK, Patil S, Shettigar H, et al. An Impact of Biofield Treatment: Antimycobacterial Susceptibility Potential Using BACTEC 460/MGIT-TB System. Mycobact Dis. 2015;5: 189.
8. Sharma S, Hanif M, Chopra KK, et al. Detection of multidrug resistance and extensively drug resistance among smear-negative extrapulmonary tuberculosis cases in a reference laboratory. Biomed Biotechnol Res J. 2018;2:132-135.
9. Getahun H, Matteelli A, Abubakar I, et al. Management of latent Mycobacterium tuberculosis infection: WHO guidelines for low tuberculosis burden countries. Eur Respir J. 2015;46:1563-1576.
10. Rahman SMM, Maliha UT, Ahmed S, et al. Evaluation of Xpert MTB/RIF assay for detection of Mycobacterium tuberculosis in stool samples of adults with pulmonary tuberculosis. PLoS One. 2018;13:e0203063.
11. Marino S, Gideon HP, Gong C, et al. Computational and empirical studies predict Mycobacterium tuberculosis-specific T cells as a biomarker for infection outcome. PLoS Comput Biol. 2016; 12: e1004804.
12. Aguiar FS, Almeida LL, Ruffino-Netto A, et al. Classification and regression tree (CART) model to predict pulmonary tuberculosis in hospitalized patients. BMC Pulm Med. 2012;12:40.
13. Nathavitharana RR, Cudahy PG, Schumacher SG, et al. Accuracy of line probe assays for the diagnosis of pulmonary and multidrug-resistant tuberculosis: a systematic review and meta-analysis. Eur Respir J. 2017;49. pii: 1601075.
14. Meaza A, Kebede A, Yaregal Z, et al. Evaluation of genotype MTBDR plus VER 2.0, line probe assay for the detection of MDR-TB in smear positive and negative sputum samples. BMC Infect Dis. 2017;17:280.
15. Yon Ju Ryu. Diagnosis of pulmonary tuberculosis: recent advances and diagnostic algorithms. Tuberc Respir Dis (Seoul). 2015; 78: 64–71.
16. Nurwidya F, Handayani D, Burhan E, et al. Molecular Diagnosis of Tuberculosis. Chonnam Med J. 2018; 54: 1–9.
17. Pollock KM, Tam H, Grass L, et al. Comparison of screening strategies to improve the diagnosis of latent tuberculosis infection in the HIV-positive population: a cohort study. BMJ Open. 2012;2:e000762.
18. Mello FC, Bastos LG, Soares SL, et al. Predicting smear negative pulmonary tuberculosis with classification trees and logistic regression: a cross-sectional study. BMC Public Health. 2006;6:43.
19. Ladefoged K, Rendal T, Skifte T, et al. Risk factors for tuberculosis in Greenland: case-control study. Int J Tuberc Lung Dis. 2011;15:44-49.
20. Salpeter EE, Salpeter SR. Mathematical model for the epidemiology of tuberculosis, with estimates of the reproductive number and infection-delay function. Am J Epidemiol. 1998;147:398-406.