International Journal of Computer Engineering and Data Science (ISSN: 2737-8543)
Volume 2– Issue 3, July – September 2022

Machine learning in epidemiology: Characterization of risk factors related to the occurrence of pulmonary and extra pulmonary tuberculosis in the province of Settat

Fatima Ezzahra SALAMATE [1], Mohamed EL AZHARI[1], Jamal ZAHI[1]

[1] University Hassan 1st, Faculty of Economics and Management, LM2CE, Settat, Morocco

E-mail : f.salamate@uhp.ac.ma , aelazharimohamed@gmail.com , zahi71@hotmail.com

Article history

Received July 04, 2022
Revised August 04, 2022
Accepted August 16, 2022
Published August 21, 2022

ABSTRACT

In this paper, we conduct a study based on machine learning tools to identify risk factors related to the occurrence of both forms of tuberculosis (noted "TB"). To do so, we use data collected from the registries of the Settat Center for Diagnosis of TB and Respiratory Diseases (CDTMR). As analysis method, we use the Probit logistic regression model. The results show that socio-demographic variables, such as patient age and gender, and clinical variables, such as registration group and duration of resistance, are risk factors that determine each of the forms of TB in patients in the province of Settat.

Keywords: ***Tuberculosis, Logistic regression, Risk factors, Probit, machine learning.***

I. INTRODUCTION

Despite the availability of a vaccine (BCG) and effective treatments, TB continues to claim lives around the world. It is among the top 10 infectious diseases causing fatalities. In 2019, the disease killed 1.4 million people [1]. It is an infectious disease caused by a mycobacterium, called bacillus Koch (BK), which most often attacks the lungs (pulmonary TB). However, it can affect other organs causing extra-pulmonary TB. It is a disease that is spread in most cases by airborne route [2]. The objective of this paper is to identify, through an econometric analysis, the risk factors related to the occurrence of these two forms of TB. We estimate a Probit logistic regression model to find out the most decisive variables that affect the occurrence of each form. The paper will be divided into two parts. In the first part, a bibliographic research, about the variables that affect the occurrence of each of these forms, will be carried out. In the second part, we conduct an empirical study using the Probit model of logistic regression on a sample of 1266 patients from the Settat region.

II. BACKGROUND

A. Epidemiological situation in Morocco

In Morocco, TB is one of the most common infectious diseases. Figures presented by the National Association for TB Awareness and Prevention (ANSPT) indicate that 3,000 people lose their lives to this disease [3]. In addition, the bulletin of the World Health Organization, published in 2019, estimates that the number of people affected is 35,000, with a specific mortality rate of 8.1/100,000 inhabitants and a reported incidence of 80 /100,000 inhabitants [4]. On the other hand, the statistics provided by the Ministry of Health stipulate that the Casablanca-Settat region is the region most affected by this disease. The province of Settat alone has nearly 1266 cases in 2020, according to figures provided by the ANSPT. The following figure traces the evolution of TB prevalence during the period 2015-2019 in the province of Settat.

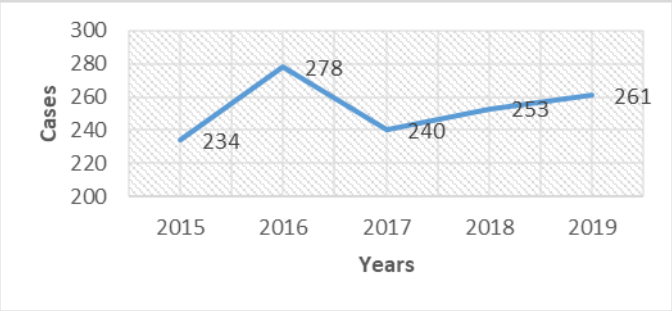

Figure 1. Number of TB cases in the province of Settat

The results in Figure 1 show that the number of TB cases increased significantly in 2016. Thereafter, it decreased by 38 cases in 2017, a rate of 3%, before resuming its increasing trend in 2019.

B. Review of the literature

In recent years, several attempts have been made to link various demographic, socioeconomic, and epidemiological factors to the development of TB, which could eventually lead to good decision making to control this disease. The above literature review shows for both forms of TB which variables are related to the occurrence of each form, as well as the

predicted influence (positive or negative) of each relationship.

TABLE I. SUMMARY OF THE LITERATURE

Paper	Review	Risk factors
L.Aazri et al [5]	Risk factors and diagnosis of TB	Standard of living, environment, lack of vaccination, recent tuberculosis count, smoking, type of diabetes, HTA and HIV status
Adil Sbayi et al [6]	Epidemiological characteristics and some risk factors of extra pulmonary TB in Larache, Morocco	Gender, age, setting and location of infection
Ossalé Abacka et al [7]	Extrapulmonary versus pulmonary TB: epidemiological, diagnostic and evolutionary aspects	Factors: sociodemographic (age, gender and residence), topographic diagnosis of TB
Muluye et al [8]	Prevalence of tuberculous lymphadenitis in Gondar University Hospital, Northwest Ethiopia	Gender, age, site of lymph node infection.
Hamzaoui, G et al [9]	Lymph node TB: epidemiology, diagnostic and therapeutic aspects	Gender, age, location of infection, environment
T. Khemis et al[10]	Sociodemographic characteristics of patients with diagnostic delay in pulmonary TB	the affiliates of the security funds, socio-demographic characteristics: place of residence, age and level of education.

III. MATERIALS AND METHODS

The data collection is carried out from the registers of tuberculosis patients registered in CDTMR of Settat. In order to meet the requirements of the research a database is created

using Excel software. We targeted all patients with TB in any form during the period 2015-2019, i.e. 1266 TB patients. The database was collected from CDTMR of Settat. Estimates were made using the RStudio software.

A. Description of variables

Two groups of variables are considered: a socio-demographic group and an epidemiological group. The following table summarizes the variables that will be tested by our model.

TABLE II. DESCRIPTION OF VARIABLES

Variable	Description	Pulmonary	Extra-pulmonary	Khi2/ T
Gender (male) (%)	Female :0 Male :1	62.71	37.29	89.47***
Age (average) (years)	Quantitative	35.32	33.55	1.86*
Residence (Rural) (%)	Urban: 0 Rural: 1	51.13	48.87	0.03
Patient Group (Relapse) (%)	New :0 Relapse:1	64.28	35.72	5.21**
Duration of resistance (+6 months) (%)	6 months :0 +6 months :1	32.11	67.89	22.62***
HIV Status (Negative) (%)	Unknown :0 Negative:1	50.7	49. 3	4.37**
Codes : '***' 1% '**' 5% '*'10%. Khi2 : Chi-square statistic. T : Student's t-statistic				

The results of the above table show that the most dominant form in men is pulmonary TB (62.71%). On the other hand, extra pulmonary TB is the most frequent in women (65%). This male predominance has been reported by other authors [7-8]. The mean age of patients with pulmonary TB is 35.32 years. These patients are older than those affected by extra pulmonary TB who have an average age of 33.5 years. In addition, we note that pulmonary TB is frequent in relapsed patients (64.28%). Also, the statistics indicate that 67.89% of the cases affected by extra pulmonary TB have a duration of resistance of the disease that exceeds 6 months. Finally, the results show that half of the TB patients are HIV negative.

B. the model

The Probit model belongs to the family of non-linear probabilistic models. It is a model based on a normal distribution [11]. The basic expression of the model is written asfollows:

$$p_i = Prob(y_i = 1|x_i) = F(x_i\beta) \forall_i = 1, \dots, N \quad (1)$$

with

$$\begin{cases} y_i = 0 \text{ if patient is affected by pulmonary TB} \\ y_i = 1 \text{ if the patient is affected by Extrapulmonary TB} \end{cases}$$

when $F(.)$ denotes a distribution function. The choice of the distribution function

$F(.)$ Is a priori unconstrained. In our case, this function corresponds to that of the normal law centered reduced, it of the following form [12] $\forall \omega \in R$

$$F(\omega) = \int_{-\infty}^{\omega} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz = \Phi(\omega) \quad (2)$$

By definition, the Probit model defines the probability associated with the occurrence of extra pulmonary TB $(y_i = 1)$ as the value of the distribution function of the normal distribution $N(0,1)$ considered at the point $x_i\alpha$[12].

IV. DISCUSSION

The results obtained affirm that most of the variables used in the model have the expected signs. The Wald test that we performed to check the nullity of the coefficients rejected the null hypothesis.

The following table summarizes the output provided by the Probit model

TABLE III. ESTIMATION OF THE COEFFICIENTS BY THE PROBIT MODEL

	Estimation of the model		Calculation of marginal effects	
Variable	Coefficient	SE_1	AME	SE_2
Gender (male)	-0.677***	0.157	-0.2530	0.0272
Age	-0.004**	0.075	-0.0018	0.0008
Residence area (Rural)	0.065	0.002	0.0235	0.0286
Group (Relapse)	-1.412***	0.079	-0.3912	0.0361
Resistance term (+6 months)	1.356***	0.231	0.4132	0.0379
HIV Status (Negative)	0.204	0.197	0.0729	0.0473
Constant	0.285*	0.134		
Codes : '***' 1% '**' 5% '*'10%. SE1: the standard error related to the coefficients. SE2: the standard error related to the marginal effects. AME: Average Marginal Effect				

The variable " Gender " has a significant impact on the endogenous variable. Men are 25% more likely to have pulmonary TB than women. This result is consistent with the authors [7-9].

Also, the variable "Age" significantly affects the risk of occurrence of pulmonary TB. That is, as age increases by one year, the risk of occurrence of this form increases by 0.18%. This observation has been made by other authors, notably [8] in their studies in Ethiopia and Tanzania.

Furthermore, the model indicates that the risk of pulmonary TB occurrence is high in relapsed patients, with a probability of 39.12% compared to new patients. This result seems to be consistent with that reported in a study in Guinea Conakry [13]. Finally, our estimates indicated that the duration of resistance is a risk factor for the occurrence of TB. In fact, as the duration of resistance increases, there is a high probability (41.32%) that the TB is pulmonary, which confirms the contagious nature of this disease.

As for the discrimination power of the model, the sensitivity rate of our regression is 0.427 and the specificity rate is 0.275. The following figure shows the ROC curve resulting from our modeling.

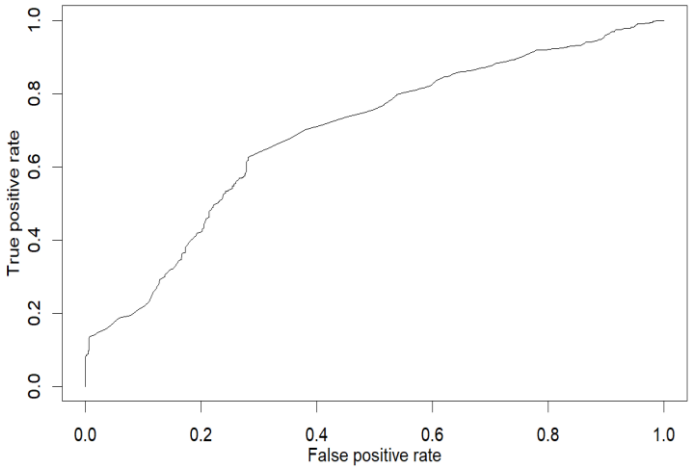

Figure 2. The ROC curve

V. CONCLUSION

In this article, we tried to identify the risk factors related to the occurrence of both forms of TB in patients from the province of Settat, which is considered the most affected in the Casablanca-Settat region. The objective was to contribute to efforts led by health professionals and policy makers. We were not able to include other sociodemographic variables that have a contribution to TB due to lack of data. This may prevent generalization of the overall prevalence trend. Studies that include the other variables may provide more generalized conclusions.

REFERENCES

[1] OMS, "Journée mondiale de lutte contre la tuberculose", 2019, https://www.who.int/fr/campaigns/world-tb-day/world-tb-

[2] "Ministère de la santé Française", 2022, https://solidarites-sante.gouv.fr/soins-et-maladies/maladies/maladies-infectieuses/article/la-tuberculose.

[3] H. bendad, "3.000 marocains meurent chaque année de la tuberculose", Maroc, 2022, 257-223, https://fr.le360.ma/societe/3000-marocains-meurent-chaque-annee-de-la-tuberculose.

[4] Ministère de la santé au Maroc "bulletin d'épidémiologie et de santé publique", 2020.

[5] L. Aazri, S. Aitbatahar, and L. Amro, "Facteurs de risques et diagnostic de la TB", 2020, Revue des MR Actualités,Volume 12, Issue1, Page 264, ISSN 1877-1203 , https://doi.org/10.1016/j.rmra.2019.11.598

[6] S. Adil , A. Amine , and J. Hasna, "Epidemiological characteristics and some risk factors of extrapulmonary tuberculosis in Larache", Pan African Medical Journal, Morocco, 2020, 36. 10.11604/pamj.2020.36.381.24870.

[7] O. Abacka, K. Koné ,A. Akoli , R. Bopaka, L.Siri ,and K.Horo ,"aspects épidémiologiques, diagnostiques et évolutifs",2018. Rev Pneumol Clin. 74(6), 452–457.

[8] M. Dagnachew, B. Belete, and G. Eden, "Prevalence of tuberculous lymphadenitis in Gondar University Hospital", BMCPublicHealth Northwest Ethiopia ,2013,13435.

[9] H. Lamyae, S. Hafsa ,TB ganglionnaire,"aspects épidémiologiques diagnostiques et thérapeutiques, à propos de 357cas", La revue médicale panafricaine , 2014 ,157, https://doi.org/10.11604/pamj,19.157.4916

[10] J. Ben Amar, T. Khemis, N. Ben Salah, “Délai diagnostique et de prise en charge de la tuberculose pulmonaire en Tunisie, Tunisie, 2015, Volume 4840, Issue100, Pages 1-296, ISSN : 0761-8425, http://dx.doi.org/10.1016/j.rmr.2015.10.269

[11] D. Bolduc, M. Kaci, " Estimation des Modeles Probit Polytomiques: Un Survol des Techniques ”, Recherche en Energie, Laval, 1991, Papers 9127.

[12] C. Hurlin, “Cours d’économétrie des Variables Qualitatives Chapitre 2: Modèles Logit Multinomiaux Ordonnées et non Ordonnés”. Université d’Orléans,2003.

[13] S. Bopaka, R. Diallo, M. Diallo, B. Diallo, and M. Sowoy, “Facteurs prédictifs de l'échec de traitement antituberculeux en Guinée Conakry”, La revue médicale panafricaine, Guinéeconakry, 2015, 146. https://doi.org/10.11604/pamj.2015.22.146.7216