https://doi.org/10.1590/1980-549720210035

ORIGINAL ARTICLE / *ARTIGO ORIGINAL*

Classification and regression trees for predicting the risk of a negative test result for tuberculosis infection in Brazilian healthcare workers: a cross-sectional study

Árvore de regressão e classificação na predição do risco de resultados negativos para infecção por tuberculose em profissionais de saúde no Brasil: um estudo transversal

Fernanda Mattos Souza[I], Thiago Nascimento do Prado[II], Guilherme Loureiro Werneck[I,III], Ronir Raggio Luiz[III], Ethel Leonor Noia Maciel[II], Eduardo Faerstein[I], Anete Trajman[III,IV]

ABSTRACT: ***Objectives:*** Healthcare workers (HCWs) have a high risk of acquiring tuberculosis infection (TBI). However, annual testing is resource-consuming. We aimed to develop a predictive model to identify HCWs best targeted for TBI screening. ***Methodology:*** We conducted a secondary analysis of previously published results of 708 HCWs working in primary care services in five Brazilian State capitals who underwent two TBI tests: tuberculin skin test and Quantiferon®-TB Gold in-tube. We used a classification and regression tree (CART) model to predict HCWs with negative results for both tests. The performance of the model was evaluated using the receiver operating characteristics (ROC) curve and the area under the curve (AUC), cross-validated using the same dataset. ***Results:*** Among the 708 HCWs, 247 (34.9%) had negative results for both tests. CART identified that physician or a community health agent were twice more likely to be uninfected (probability = 0.60) than registered or aid nurse (probability = 0.28) when working less than 5.5 years in the primary care setting. In cross validation, the predictive accuracy was 68% [95% confidence interval (95%CI): 65 – 71], AUC was 62% (95%CI 58 – 66), specificity was 78% (95%CI 74 – 81), and sensitivity was 44% (95%CI 38 – 50). ***Conclusion:*** Despite the low predictive power of this model, CART allowed to identify subgroups with higher probability of having both tests negative. The inclusion of new information related to TBI risk may contribute to the construction of a model with greater predictive power using the same CART technique.

Keywords: Latent tuberculosis. Occupational risks. Machine learning. Decision trees.

[I]Universidade do Estado do Rio de Janeiro – Rio de Janeiro (RJ), Brazil.
[II]Universidade Federal do Espírito Santo – Vitória (ES), Brazil.
[III]McGill University – Montreal (QC), Canada.
[IV]Universidade Federal do Rio de Janeiro – Rio de Janeiro (RJ), Brazil.
Corresponding author: Anete Trajman. Rua Macedo Sobrinho, 74, ap. 203, Humaitá, CEP: 22271-080, Rio de Janeiro, RJ, Brazil. E-mail: *atrajman@gmail.com*

Conflict of interests: nothing to declare – **Financial support:** none.

RESUMO: ***Objetivos*:** Desenvolver um modelo preditivo para identificar profissionais de saúde com maior probabilidade de resultado negativo para dois testes de diagnóstico da infecção latente por *Mycobacterium tuberculosis* (ILTB). ***Métodos*:** Foi realizada uma análise secundária dos resultados publicados anteriormente de 708 profissionais de saúde da atenção primária, de cinco capitais brasileiras, submetidos à prova tuberculínica e ao Quantiferon®-TB Gold in-tube. Um modelo preditivo com árvore de classificação e regressão (CART, *Classification and regression tree*) foi construído. A avaliação do desempenho foi realizada por meio da análise *receiver operating characteristics* (ROC) e *area under the curve* (AUC). Utilizamos o mesmo banco de dados para validação cruzada do modelo. ***Resultados*:** Entre os 708 profissionais de saúde, 247 (34,9%) apresentaram resultado negativo para os testes. A CART identificou que os médicos e agentes comunitários de saúde apresentaram duas vezes mais chances de não estarem infectados (probabilidade = 0,60) que os enfermeiros e técnicos/auxiliares de enfermagem (probabilidade = 0,28) nos casos com menos de 5,5 anos de atuação na atenção primária. Na validação cruzada, a acurácia do modelo preditivo foi de 68% [intervalo de confiança de 95% (IC95%) 65 – 71)], AUC de 62% (IC95% 58 – 66), especificidade de 78% (IC95% 74 – 81) e sensibilidade de 44% (IC95% 38 – 50). ***Conclusão*:** Apesar do baixo poder preditivo do modelo, a CART permitiu identificar subgrupos com maior probabilidade de terem ambos os testes negativos. A inclusão de novas informações relacionadas ao risco de ILTB pode contribuir para a construção de um modelo com maior poder preditivo utilizando a mesma técnica.

***Palavras-chave*:** Tuberculose latente. Riscos ocupacionais. Aprendizado de máquina. Árvores de decisões.

INTRODUCTION

Tuberculosis (TB) is a global epidemics that caused an estimated 1.2 million deaths in 2019[1]. However, this disease is not only curable but also preventable through treatment of TB infection (TBI). Healthcare workers (HCWs) are at high risk of TBI because of occupational exposure[2,3], and recent TBI is one of the risk factors for progression to active disease. Thus, HCWs without evidence of previous TBI should be annually tested for conversion of one of the available tests, such as the tuberculin skin test (TST) or the interferon-gamma release assays (IGRA), and eventually treated[4]. Those with positive results should be carefully followed up, but no re-testing or treatment is recommended.

Nonetheless, tuberculin and consumables for IGRA tests are costly. Additionally, TST needs trained personnel, and IGRA tests need laboratory infrastructure, not widely available in many poor-resource settings. While awaiting for effective and affordable predictors of progression to active TB, identifying the targeted HCWs population that would most benefit from testing, i.e., those without evidence of previous TBI, could enable more efficient use of available resources. We built a predictive model to identify HCWs with negative results for these two tests using a machine learning technique, the classification and regression trees (CART). Since both tests may provide false positive and negative results and there is no evidence that one is superior to the other, we considered that HCWs with negative results of both TST and IGRA were free of TBI[4].

CART allows development of prediction models using binary splits and offers an intuitive method for obtaining predictions of outcome using processes familiar to HCWs (e.g., "high" versus "low" values of a predictor)[5]. Models are easily read and interpreted using a flow chart diagram[6].

We hypothesized that CART model for predicting negative result for TST and IGRA would have good overall performance in terms of accuracy, sensitivity, specificity and area under the curve (AUC).

METHODS

STUDY DESIGN, SETTING AND SOURCE OF DATA

We analysed factors potentially associated with the risk of having negative results for both tests for TBI in a database from a previously conducted cross-sectional observational multicenter study[7,8]. The database contains information on sociodemographic characteristics, health facility, and work conditions collected from face-to-face interviews conducted between June 2011 and September 2013[7,8], as well as BCG status (through observation of scar) and Quantiferon® TB Gold in-tube (QFT) – and TST (PPD RT23 - Tuberculin PPD Evans 2 TU) results from HCWs working in primary care services in five Brazilian State capitals, those with the highest TB incidence rates at the moment of data collection: Cuiabá (TB incidence of 52/100,000), Manaus (71/100,000), Salvador (60/100,000), Porto Alegre *(87/100,000)* and Vitória (40/100,000)[9].

Since the 1990s, primary care in Brazil has been progressively shifted to the Family Health Strategy model, in which residents of the adjacent area are actively taken in charge by family health teams composed by one medical doctor (MD), one registered nurse (RN), one to two technical nurse assistants (TNA) and six to ten community health agents (CHA) who pay regular home visits, regardless of the demand care[10].

Some cities have rapidly adopted the Family Health Strategy model, with decentralized TB care, as others still use a mixed approach with TB services still centralized in specialized services. For the original study, primary care units were selected by simple random sampling. The selected units were classified into three categories: "traditional" primary health care units, Family Health Strategy units, and "traditional" primary health care units with CHA program[11]. The number of health units was defined based on the number of professionals in each unit and stratified by type of service organization in Brazil.

All HCWs from the selected units were invited to participate in the study, and those who signed the informed consent were eligible. The description of the sampling procedure is available in the article by Prado et al.[7]. Exclusion criteria were known human immunodeficiency virus (HIV) infection, a positive rapid HIV test, past or current active TB, any positive test for TBI in the past, and pregnancy. HCW who did not return for

TST reading were also excluded from this analysis. Inclusion criteria were to be MD, RN, TNA or CHA.

The original study was approved by the ethics committee of the Federal University of Espírito Santo on March 3, 2010 (number: 007/10). The currently analysed database is anonymous.

DATABASE VARIABLES

Outcome

Because there is no reference standard for TBI diagnosis, we considered any positive test as evidence of TBI[4]. Thus, our outcome was set as having negative results for both TST and QFT, as a surrogate for absence of TBI. For TST, 5 mm cut-off value was considered, according to Brazilian guidelines[12].

Predictive variables

After review of the literature[7,13-20], the following predictive variables were selected from the database: TB incidence in the city – intermediate (TB incidence < 50/100,000) versus high (TB incidence ≥ 50/100,000); type of TB care provided in the city (centralized or decentralized); type of health clinics (traditional, traditional with CHA program or family health clinics); working in primary care unit with specialized services including TB services; working in specialized TB services; air flow in the unit (no open door or window versus at least one open door or window); professional category (MD, CHA, TNA and RN); years served in a primary health care unit; work in highly TB-exposed setting (necropsy room, radiotherapy and respiratory disease wards); home visits to nursing home, asylum or prison; assistance of patients with active TB; use of N95 masks (always versus not always/never); TB training; household contact of active pulmonary TB; morbidities (diabetes mellitus, chronic cardiovascular disease, rheumatic disease, respiratory disease, chronic kidney disease or chronic liver disease); tobacco use (smokers versus non- or ex-smokers); age (years) and alcohol use (no or yes).

CLASSIFICATION AND REGRESSION TREES ANALYSES

Supervised learning methods can be used as strategy for the prediction of test results. In supervised learning, input variables (X) and an output variable (Y) enter an algorithm to learn the mapping function from the input to the output Y = f(X). The goal is to approximate the mapping function so well that new input data (X) can predict the output variables (Y) for that data[21]. Supervised machine learning algorithms include CART.

CART models were developed using an algorithm first introduced by Breiman et al.[22]. These models offer clear interpretation by relating continuous or categorical predictive variables to the outcome of interest based on optimal splitting criteria from an automated algorithm. CART is a non-parametric method that builds a binary classification system (tree) through recursive partitioning, so that the data set is successfully split into increasingly homogenous subgroups[6]. Firstly, a variable that optimally separates outcome groups is selected, and a binary split is made. Then, from both subgroups, subsequent variables are selected, and second levels of binary splits are made. Variables can be used more than once within a model. Variable splits are made recursively until stopping criteria are reached, and a terminal node is defined with a prediction for the specific subset of data in this node[5,6].

The trees should be read from top to bottom in order to obtain a prediction for a specified outcome. Starting at the top of a tree, branches corresponding to observed features are followed until a terminal node has been reached and the fraction of patients contained in each outcome group is displayed. These tables may be used to assess the probability that a patient falls within each outcome category[5].

CART models were constructed using all HCWs, followed by cross validation in 10 subsets the same complete dataset. The R package *rpart* was used to develop the CART models[23]. For cross-validation, the minimum number of observations in terminal nodes was 15.

DATA ANALYSES

Information was encoded and stored anonymously in an Excel for Windows® database; data analyses were performed using RStudio software[24] and Stata 13 software[25]. The variance inflation factor (VIF) was used to evaluate the presence of collinearity between the predictive variables (R package *usdm*). The variable has been classified as not correlated VIF is less than 10. The association of the predictive variables with the outcome was calculated in bivariate analyses as the crude odds ratio (OR) with their 95% confidence intervals (95%CI) and p.

The performance of the model was evaluated using the receiver operating characteristics (ROC) curve and the AUC, cross-validated using 10 subsets the same dataset. The accuracy, sensitivity, specificity, negative predictive value (NPV) and positive predictive value (PPV) were also calculated. For this, we generated a discrete variable in which each terminal node of CART received an increasing score according to the probability of occurrence of the outcome of interest. This score was equal for nodes 7 and 10. The cut-off point used to calculate discriminatory capacity was selected in order to maximize the sensitivity and specificity (probability of outcome threshold = 0.35).

RESULTS

Out of 740 enrolled HCWs, 708 were included. The reasons for exclusion were: 22 (3%) HCWs did not return for TST reading, 7 (1%) had active TB or were under TB treatment, 1 (0.1%) was HIV positive (positive rapid test), and 2 (0.3%) excluded because refused to have blood drawn (Supplementary Figure 1). A BCG scar was observed in 87.6%. Then mean age was 41.4 [standard deviation (SD) = 9.9] years, and 633 (89.4%) were female (Supplementary Figure 1); mean time of work was 9.5 (SD = 6) years. Among the total HCWs included, 247 (34.9%) were negative for both tests and 461 (65.1%) were at least one of the positive tests, with 57.3% presented TST positive and QFT negative. We did not identify collinearity between the predictive variables.

Among the 708 HCWs, 247 (34.9%) had negative results for both TST and QFT. Not use alcohol (OR = 1.42, 95%CI 1.03 – 1.95); do not smoke (OR = 2.05, 95%CI 1.13 – 3.71); less than 10 years of work in primary care (OR = 1.47, 95%CI 1.08 – 2.01); city where TB service is decentralized (OR = 1.59, 95%CI 1.16 – 2.18), and city with intermediary TB incidence (OR = 1.66, 95%CI 1.19 – 2.32) were all associated with negative results for both TST and QFT (Supplementary Table 1). On the other hand, RN (OR = 0.41, 95%CI 0.19 – 0.85) or TNA (OR = 0.45, 95%CI 0.24 – 0.86) professional category (Supplementary Table 1) and be older (OR = 0.97, 95%CI 0.95 – 0.98) were associated with the lowest chance of presenting negative results for both TST and QFT.

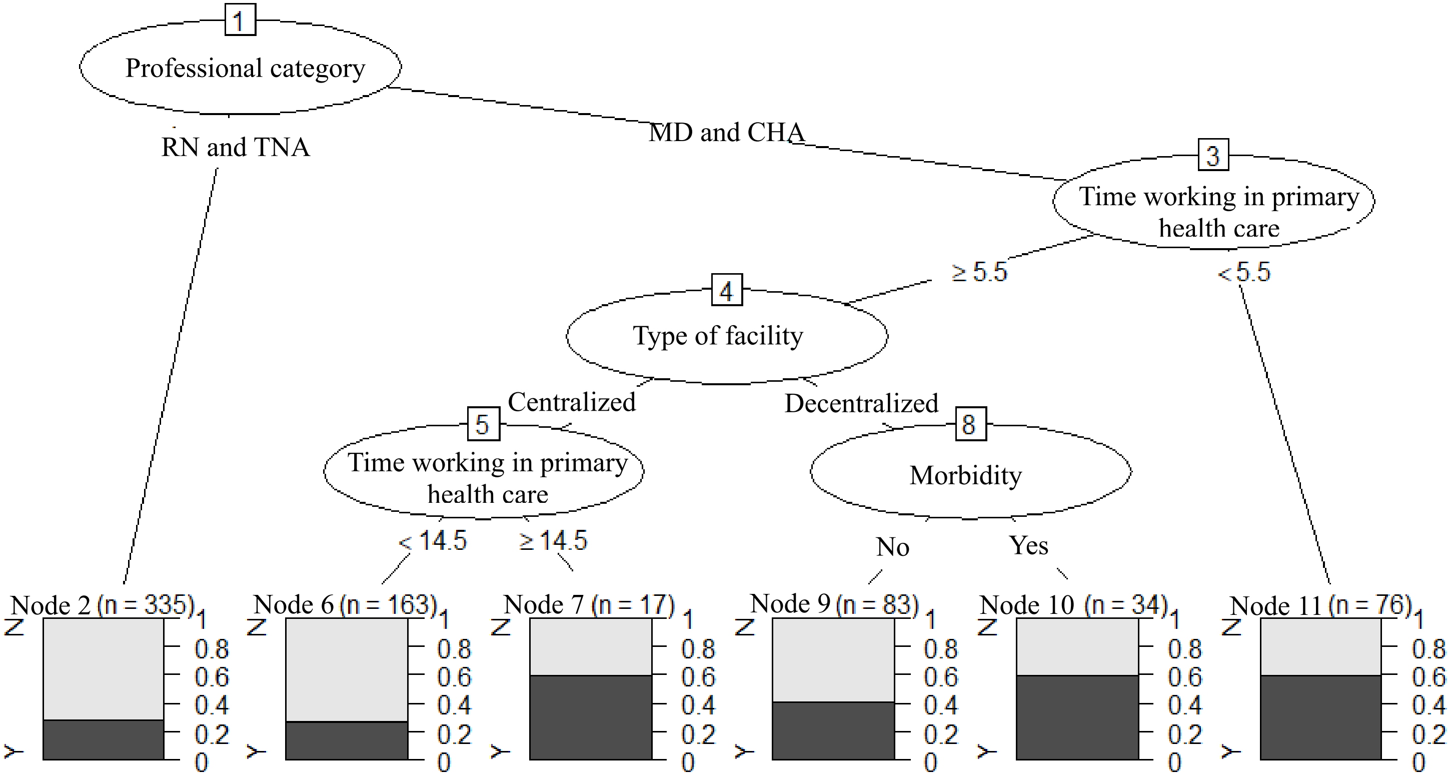
The CART model (Figure 1) also identified the professional category as the most important predictor of negative test results. The following set of features were associated with a higher probability of having negative results of both tests:

- MD or CHA working for less than 5.5 years in primary care (node 11, probability = 0.60);
- MD or CHA working for more than 5.5 years in primary care in a city with decentralized assistance to patients with TB and with any morbidity (node 10, probability = 0.60);
- MD or CHA working for more than 5.5 years in primary care in a city with centralized assistance to patients with TB for more than 14.5 years in primary care (node 7, probability = 0.59).

Conversely, the following features were associated with a lower probability of having the outcome (both tests negative):

- MD or CHA working for more than 5.5 years in primary care in a city with centralized assistance to patients with TB for less than 14.5 years in primary care (node 6, probability = 0.27);
- TNA or RN (node 2, probability = 0.28).

Performance measures and the associated confidence intervals for the CART model are presented in Table 1 for cut-off equal 0.35. The sensitivity was 44%, with a predictive



CHA: community health agents; MD: medical doctors; QFT Quantiferon® TB Gold in-tube test; RN: registered nurse; TNA: technical nurse assistants; TST: tuberculin skin test. Positive criterion: for TST ≥ 5 mm and for QFT ≥0.35 UI.

Figure 1. Classification and regression tree for predicting model of risk of both negative tests: TST and QFT. The decision has four predictors: professional category; time working in primary health care (years); type of facility (centralized – city where tuberculosis treatment occurs only in reference units in primary care; or decentralized – this treatment is available in all primary health care units of the city) and morbidities. Terminal nodes containing predictions for new observations include 2, 6 and 9 (predict the risk for at least one positive test) and 7, 10 and 11 predict the risk for negative tests. To obtain a prediction, one starts at the top of the tree and follows the arrow corresponding to data for the new observation until a terminal node is reached.

Table 1. Predictive performance of classification and regression tree model for risk of having negative results of both Quantiferon® TB Gold in-tube test and tuberculin skin test in healthcare workers of primary care (threshold = 0.35).

	Both negative tests (QFT and TST)	
	Result	95%CI
AUC	0.62	0.58 – 0.67
Accuracy	0.68	0.65 – 0.71
Specificity	0.78	0.74 – 0.81
Sensibility	0.44	0.38 – 0.50
PPV	0.52	0.45 – 0.59
NPV	0.73	0.68 – 0.76

AUC: area under the curve; 95%CI: 95% confidence interval; QFT: Quantiferon® TB Gold in-tube test; NPV: negative predictive value; PPV: positive predictive value; TST: tuberculin skin test.

accuracy of 68%, a specificity of 78% and PPV of 52%. The AUC was 62%. When reducing the cut-off point to 0.27, the specificity was 26%; the sensitivity, 82%; and the PPV, 37% (Supplementary Table 2). The age variable was not included in the proposed model as it increases complexity without gaining accuracy (0.69, 95%CI 0.65 – 0.72) (Supplementary Figure 2 and Supplementary Table 3).

DISCUSSION

In this study, we found a high prevalence of TBI among HCWs, with only 34.9% being negative for both TST and QFT. This finding indicates that the strategy of predicting those without evidence of TBI, at risk for conversion, and focusing efforts to test them may contribute to better allocate human and supply resources. The use of CART could be an alternative to this end. However, the current model — that used the available variables — had a low predictive power. Nevertheless, the CART was still useful for selecting subgroups that are most likely to have negative results of both tests, thus at-risk for conversion and worthy testing. Most importantly, the current exercise pointed out a method that could be useful if further variables were available.

CART identified that MDs and CHAs were twice more likely to be uninfected than TNA or RN when working less than 5.5 years in the primary care setting with high specificity (78%) despite a PPV of 52% because of the low prevalence of this condition in a high--TB burden country. Among those with more than 5.5 years of work, existing morbidities and work in a city with decentralized assistance to TB patients also had high specificity.

Reducing the cut-off point to 0.27 resulted in a higher sensitivity (82%), but on expense of lower specificity (26%), and decreased PPV (37%) (Supplementary Table 2).

Interestingly, CART identified mostly occupational characteristics for TBI, as opposed to the bivariate analysis that also identified individual characteristics. In a systematic review of 85 studies from low and middle-income countries published from 2005 to 2017, occupational categories and years of work had already been reported to be an independent risk factor for both prevalent and incident TBI[26]. However, in our study, CART identified subgroups with distinct characteristics at higher risk. RN and TNA had a less probability for negative tests results regardless of any other variable, while among MD and CHA, years of work in primary health care influenced this probability, which could double in those having worked for less than 5.5 years (node 11). Thus, unlike bi or multivariate analyses, CART points out to a set of characteristics of subgroups that can be then identified, allowing a specific strategy to be proposed to these different groups.

Time of work was also clearly relevant in other subgroups (nodes 6, 7 and 11). One difficult-to-explain finding was the high probability of negative results among HCWs in which TB care is centralized (node 7). This finding should be interpreted with caution since the node contains only 17 observations.

Our study has a few limitations. First, the cross-sectional design does not allow to predict the risk for conversion (incidence of TBI), which would be more informative than the risk for absence of prevalent TBI[27]. Second, there is no reference test for detecting TBI, thus the estimated TBI prevalence might be impacted by the TST and QFT performance. Persons with one negative (discordant) test might be uninfected. The cross validation was performed with the same set used to build CART, but the external data should be used to further validate the CART model[28]. Third, the overall accuracy of the CART model was low because the three nodes with the highest pretest probability (nodes 7, 10 and 11) have only 127 observations. Finally, given the possibility of cross-reaction between TST and BCG (31), the result of this test may have been affected by previous BCG vaccination since 87.6% of HCWs had the BCG scar.

Despite these limitations, we here demonstrate the possibility of the use CART for the development of a simple and intuitive predictive model for absence of TBI in HCWs considering a strict criterion, i.e., both QFT and TST results. CART identified specific subgroups that should be prioritized for targeted TBI testing, such as those with less time of work or those with existing morbidities working outside specialised TB services. The inclusion of new information related to TBI risk among HCWs at this level of attention may contribute to the construction of a model with greater predictive power.

REFERENCES

1. World Health Organization. Global tuberculosis report 2020. Geneva: World Health Organization; 2020.
2. Uden L, Barber E, Ford N, Cooke GS. Risk of Tuberculosis Infection and Disease for Health Care Workers: An Updated Meta-Analysis. Open Forum Infect Dis [Internet] 2017; 4(3): 1-7. https://doi.org/10.1093/ofid/ofx137
3. Schmidt BM, Engel ME, Abdullahi L, Ehrlich R. Effectiveness of control measures to prevent occupational tuberculosis infection in health care

workers: A systematic review. BMC Public Health 2018; 18(1): 661. https://doi.org/10.1186/s12889-018-5518-2

4. World Health Organization. Latent Tuberculosis Infection. Updated and consolidated guidelines for programmatic management. Geneva: WHO Library; 2018.
5. Speiser JL, Karvellas CJ, Shumilak G, Sligl WI, Mirzanejad Y, Gurka D, et al. Predicting in-hospital mortality in pneumonia-associated septic shock patients using a classification and regression tree: a nested cohort study. J Intensive Care 2018; 6: 66. https://doi.org/10.1186/s40560-018-0335-3
6. Yohannes Y, Hoddinott J. Classification and regression trees: An Introduction. Tech report, Int Food Policy Res Inst [Internet]. 1999 [cited 2018 Feb 11]. Available from: http://pdf.usaid.gov/pdf_docs/Pnach725.pdf
7. Prado TN, Riley LW, Sanchez M, Fregona G, Nóbrega RLP, Possuelo LG, et al. Prevalence and risk factors for latent tuberculosis infection among primary health care workers in Brazil. Cad Saude Publica [Internet]. 2017 [cited 2018 Feb 23]; 33(12): e00154916. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-311X2017001205008&lng=en&tlng=en. https://doi.org/10.1590/0102-311x00154916
8. Souza FM, Prado TN, Pinheiro JS, Peres RL, Lacerda TC, Loureiro RB, et al. Comparison of Interferon-c Release Assay to Two Cut-Off Points of Tuberculin Skin Test to Detect Latent Mycobacterium tuberculosis Infection in Primary Health Care Workers. PLoS One [Internet] 2014 [cited 2018 Feb 3]; 9(8): e102773. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4138087/pdf/pone.0102773.pdf. https://doi.org/10.1371/journal.pone. 0102773
9. Brasil. Ministério da Saúde. DATASUS. Indicadores de morbidade [Internet]. Brasil: Ministério da Saúde; 2009 [cited 2014 Feb 4]. Available from: http://tabnet.datasus.gov.br/cgi/tabcgi.exe?idb2010/d0202.def
10. Paim J, Travassos C, Almeida C, Bahia L, Macinko J. The Brazilian health system: history, advances, and challenges. Lancet 2011; 377(9779): 1778-97. https://doi.org/10.1016/S0140-6736(11)60054-8
11. Brasil. Ministério da Saúde. Política Nacional de Atenção Básica. Brasil: Ministério da Saúde; 2012.
12. Brasil. Ministério da Saúde. Manual de Recomendações para o Controle da Tuberculose no Brasil. 2ª ed. Brasilia: Ministério da Saúde; 2018. 364 p.
13. Rogerio WP, Prado TN, Souza FM, Pinheiro JS, Rodrigues PM, Sant'anna APN, et al. Prevalência e fatores associados à infecção pelo Mycobacterium tuberculosis entre agentes comunitários de saúde no Brasil, usando-se a prova tuberculínica. Cad Saúde Pública 2015 [cited 2018 Dec 17]; 31(10). Available from: https://doi.org/10.1590/0102-311X00152414
14. Rodrigues PM, Moreira TR, Moraes AKL, Araújo Vieira RC, Dietze R, Lima RCD, et al. Mycobacterium tuberculosis infection among community health workers involved in TB control. J Bras Pneumol [Internet]. 2009 [cited 2018 Dec 17]; 35(4): 351-8. Available from: http://www.scielo.br/pdf/jbpneu/v35n4/en_v35n4a09.pdf
15. Moreira TR, Zandonade E, Maciel ELN. Risco de infecção tuberculosa em agentes comunitários de saúde. Rev Saúde Pública [Internet]. 2010 [cited 2018 Dec 17]; 44(2): 332-40. Available from: https://doi.org/10.1590/S0034-89102010000200014
16. Menzies D, Joshi R, Pai M. Risk of tuberculosis infection and disease associated with work in health care settings. Int J Tuberc Lung Dis [Internet]. 2007 [cited 2018 Feb 3]; 11(6): 593-605. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17519089
17. Zhang H, Xin H, Li X, Li H, Li M, Lu W, et al. A dose-response relationship of smoking with tuberculosis infection: A cross-sectional study among 21008 rural residents in China. PLoS One 2017; 12(4): e0175183. https://doi.org/10.1371/journal.pone.0175183
18. Soto Cabezas MG, Munayco Escate CV, Chávez Herrera J, López Romero SL, Moore D. Prevalencia de infección tuberculosa latente en trabajadores de salud de establecimientos del primer nivel de atención. Lima, Perú. Rev Peru Med Exp Salud Publica [Internet] 2017 [cited 2018 Feb 23]; 34(4): 649. Available from: https://doi.org/10.17843/rpmesp.2017.344.3035
19. Milburn H, Ashman N, Davies P, Doffman S, Drobniewski F, Khoo S, et al. Guidelines for the prevention and management of Mycobacterium tuberculosis infection and disease in adult patients with chronic kidney disease. Thorax [Internet] 2010 [cited 2018 Feb 3]; 65: 559-70. Available from: http://thorax.bmj.com/content/thoraxjnl/65/6/559.full.pdf. https://doi.org/10.1136/thx.2009.133173
20. Jeon CY, Murray MB. Diabetes Mellitus Increases the Risk of Active Tuberculosis: A Systematic Review of 13 Observational Studies. PLoS Med [Internet] 2008 [cited 2018 Feb 3]; 5(7): e152. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2459204/pdf/pmed.0050152.pdf. https://doi.org/10.1371/journal.pmed.0050152
21. Kotsiantis SB, Zaharakis ID, Pintelas PE. Supervised Machine Learning: A Review of Classification Techniques. Emerg Artif Intell Appl Comput Eng 2007; 160: 249-68.
22. Breiman L, Friedman JH, Olshen RASC. Classification and regression trees. CA: Wadsworth Brooks; 1984.
23. Therneau TM, Atkinson EJ. An Introduction to Recursive Partitioning Using the RPART Routines [Internet]. 2019 [cited 2018 Jul 3]. Available from:

https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

24. Team RC. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2019.

25. StataCop. Stata Statistical Software: Release 13. Coll Station: StataCorp LP; 2013.

26. Apriani L, McAllister S, Sharples K, Alisjahbana B, Ruslami R, Hill PC, et al. Latent tuberculosis infection in health care workers in low and middle-income countries: an updated systematic review. Eur Respir J [Internet]. 2019; 53: 1801789. Available from: http://erj.ersjournals.com/lookup/doi/10.1183/13993003.01789-2018. https://doi.org/10.1183/13993003.01789-2018

27. Szklo M, Nieto FJ. Epidemiology: Beyond the Basics. 4ª ed. Burlington; 2019. 588 p.

28. Rokach L, Maimon O. Data mining with decision trees: theory and applications [Internet]. 2ª ed. World Scientific; 2014 [cited 2018 Feb 11]. Available from: https://doc.lagout.org/Others/Data Mining/Data Mining with Decision Trees_ Theory and Applications %282nd ed.%29 %5BRokach %26 Maimon 2014-10-23%5D.pdf

Received on: 11/05/2020
Revised on: 03/05/2021
Accepted on: 03/05/2021

Authors' contributions: FMS and AT designed the research, and RRL and TNP provided guidance in study design. FMS carried out the analysis. FMS, AT and EF wrote the first manuscript draft. RRL and GWL provided assistance with the data analysis and interpretation. TNP and ELNM reviewed drafts of the paper, provided edits and input to the final draft. All authors read, edited and approved the final manuscript.