IEEE International Conference on Modeling, Simulation and Applied Optimization (ICMSAO), Marrakesh, Morocco, April 2023.

Uncertainty-Aware Classification of Tuberculosis Subtypes with Machine Learning Techniques and Probabilistic Calibration

Jayroop Ramesh,[1, a)] Zahra Solatidehkordi,[1, b)] Donthi Sankalpa,[1, c)] Amar Khamis,[2, d)] Assim Sagahyroon,[1, e)] and Fadi Aloul[1, f)]

[1]*Computer Science and Engineering American University of Sharjah Sharjah, United Arab Emirates.*
[2]*College of Medicine Mohammed Bin Rashid University Dubai, United Arab Emirates*

[a)]Corresponding author: b00057412@aus.edu
[b)]g00059068@aus.edu
[c)]dsankalpa@aus.edu
[d)]amar.hassan@mbru.ac.ae
[e)]asagahyroon@aus.edu
[f)]faloul@aus.edu

Abstract. Tuberculosis (TB) is a pulmonary infectious disease causing morbidity and mortality in developing countries. In 2021, an estimated 10.6 million individuals were affected with TB, and 10% of these cases proved fatal. Due to socioeconomic factors and the difficulty of diagnosis, this preventable and curable condition stays prevalent in many nations. With the advent of automated public health clinical support systems aided by machine learning (ML), there is potential to triage and prioritize at-risk strata of the population. In this work, we seek to assess the utility of algorithmic techniques applied to routinely available electronic health records (EHR) in classifying three types of TB: active in the lung, non-active in the lung, and extrapulmonary. The Light Gradient Boosting (CB) algorithm achieved the comparatively highest overall scores of accuracy: 77.4%, sensitivity: 56.1%, specificity: 74.1%, and F1-score: 55.2% for the multi-class scenario. In addition, post-hoc clinical explainability is introduced by means of Shapley values and permutation feature importance. We then employ Spline-based smoothing calibration to enable uncertainty quantification and confer confidence levels to probabilistic predictions. This work highlights the role of ML approaches in facilitating population-level screening for curbing the spread of TB in developing countries.

INTRODUCTION

According to the World Health Organization, approximately 1.6 million people died from Tuberculosis (TB) in 2021, making it the second leading infectious condition with high mortality after the novel coronavirus (COVID-19) and before Human Immunodeficiency Virus/ Acquired Immunodeficiency Syndrome (HIV/AIDS). Thus, bringing an end to the TB epidemic, primarily endemic in the developing world, by 2030 is among the healthcare targets of the United Nations Sustainable Development Goals [1].

Tuberculosis (TB) is a contagious pulmonary disease caused by the etiologic agent Mycobacterium tuberculosis. The deposition of this bacteria initiates TB in aerosol droplets onto lung alveolar surfaces. The severity of progression is determined by several factors such as the current state of the host immune system, age/gender, and genetics. Generally, it impairs the functioning of the lungs by moving through the lymphatic system/blood and leads to potential respiratory failure, liver cirrhosis, and eventual death [2].

The timely detection of TB and identification of potential patients is instrumental in effectively treating the condition to avoid adverse effects [3]. Moreover, earlier recognition can forewarn the respective health authorities to take action in order to prevent transmission. However, owing to the elusive nature of the condition and the lack of specific clinical symptoms, it is considerably challenging to diagnose tuberculosis [4]. Also, several socioeconomic factors in developing countries hinder large-scale screening due to limited infrastructure and trained staff [5].

Artificial intelligence and machine learning (ML) techniques can play an important role in screening and prognosis of TB, given their success in using public health information such as electronic health records (EHR) and questionnaires responses to accurately identify at-risk patients for sleep apnea [6], Chronic Obstructive Pulmonary Disease (COPD) [7] and COVID-19 [8] ahead of time of disease onset. Automated models can assist in the easing burden on clinical staff, and help curb the incidence rate by analyzing and prioritizing at-risk patients at a quicker rate. Indeed, several national strategies in many countries leverage automated clinical support models as the first phase in the overall diagnosis loop, in order to triage the vulnerable strata of society [9].

Motivated by previous work in this area, which explores different types of feature dimensions for the classification of TB risk, we retrospectively analyze a dataset consisting of questionnaire responses, symptomatology, and demographic factors using a variety of ML approaches. We summarize the findings of recent literature as follows.

The authors in [10] utilize 7 primary variables of gender, age, socioeconomic background, location, the status of HIV/AIDs, and status of antiretroviral treatment, to obtain scores of accuracy, sensitivity, and area under the curve (AUC) of 85%, 93%, and 82% respectively with an Artificial Neural Network (ANN) when detecting TB positive and TB negative. The work in [11] saw the combination of six different datasets across several nations and demonstrated the ability of the J48 classifier to predict treatment failure with an average accuracy of 78% given 22 variables belonging to demographics and clinical categories. As a consequence of privacy and security concerns, [12] assessed the utility of synthetic data for training ML algorithms based on 31 inflammatory antigen biomarker profiles. They found that the real and optimal synthetic datasets had accuracy, sensitivity, and specificity scores of 90%. 89%, 100%, and 91%, 93%, and 77% respectively in classifying the outcomes of TB positive or TB negative. In a similar vein, another study [9] sought to also evaluate the serological features (as multiple antibodies arising from variable immune responses can be informative) and discovered that 23 antigens were robust enough to detect TB cases among a mixed (sick vs healthy) population with a sensitivity of 90.5% and specificity of 100%. A more complex approach in [13] examined the resistance of genes for TB drugs to determine drug resistance-associated mutations with ML and obtained an average accuracy of 85% across all models. Another modality typically considered is also imaging, where chest X-Ray radiographs are used to develop Computer-Aided Diagnostic (CAD) systems with deep learning with promising results.

Aside from the first study mentioned, the rest involved the collection of blood, plasma, sputum, genetic samples, or body imaging scans, which may not be possible in developing countries. We aim to contribute to the existing body of work by studying the potential of routinely acquired data to infer the possibility of TB active in the lung, TB passive in the lung (non-infectious), and extra-pulmonary (TB in other organs). Thereby, the contributions of this work are as follows:

1. Evaluate machine learning methods using a routinely available questionnaire, symptoms, and demographic variable for classifying three TB states.

2. Assessing variable relevance to outcomes by virtue of feature importance and explainable approaches.

3. Calibrating multi-class uncertainty to confer interpretable confidence to probabilistic outcomes.

This paper is organized in the following manner: Section 2 outlines the methodology, Section 3 presents the results, Section 3 discusses the findings and Section 5 concludes the work.

Materials and Methods

The dataset acquisition procedure was carried out in a private hospital in Sudan. All persons with respiratory symptoms seen in the general health services were considered eligible for the study. A structured interview was conducted to collect information on age, sex, residence, medical history, education, occupation, ethnicity, income, and the number of people living in the same room.

According to National Tuberculosis Programme (NTP) policy [14], all patients presenting with cough for 3 weeks or more and/or other symptoms such as night sweats, fever, chest pain and haemoptysis, were designated suspects for TB and referred for sputum examination. Direct microscopy for Acid-Fast Bacilli (AFB0 was performed on three sputum specimens stained using the Ziehl-Neelsen method and graded according to standard classification. 6,23 Those who were found smear-positive for AFB in two or more samples were diagnosed as having smear-positive TB. If one sample only was found to be smear-positive, three new sputum samples were examined. For those in whom all specimens failed to demonstrate AFB (or only one out of six was smear-positive), a chest radiograph was performed [5]. If this demonstrated abnormalities possibly associated with tuberculosis a course of ordinary antibiotics was prescribed. If there was no response to the antibiotics, the patient was referred for evaluation by a medical officer. If the medical officer judged that the patient was suffering from TB, the patient was registered as a case of smear-negative TB and given treatment. Extra-pulmonary TB is a case proved by one culture-positive specimen from an extra-pulmonary site, histo-pathological evidence from a biopsy or based on strong clinical evidence consistent with active extra-pulmonary TB, followed by the decision by a physician to treat with a full course of antituberculosis chemotherapy.

The features included in the dataset and subsequently used for classification are sex (male or female), age, marital status, number of people in the room they reside in, presence of coughing, duration of cough, recent weight loss status, recent weight loss duration, presence of tiredness, duration of tiredness, presence of fever, duration of fever,

presence of night sweats, duration of night sweats, presence of chest pain, duration of chest pain, shortness of breath, duration of shortness of breath, loss of appetite, duration of loss of appetite, hemptosis, duration of hemptosis, difficulty in swallowing, duration of difficulty in swallowing, reduced vision, duration of reduced vision, presence of skin lesions, duration of skin lesions, presence of diarrhoea, duration of diarrhoea, presence of buccal leucoplakia, duration of buccal leucoplakia, fatigue, fatigue duration, headache, duration of headache, generalized pains, duration of generalized pains, other symptoms, duration of other symptoms, presence of HIV/AIDS, contact with TB patient, contact with coughing or sick family members, anonymous testing, number of injections, duration of first complaint, general appearance, weight (kg), height (cm) and presence of BCG vaccination scar. Weight and height were used to derive the composite measurement of body-mass index (BMI), with units kg/m^2.

In accordance with recent literature [6], the commonly used traditional and ensemble machine learning methods were considered. These are Support Vector classifier (SVC), Logistic Regression (LR) and K-Nearest Neighbors (KNN), Light Gradient Boosting (LGB), eXtreme Gradient Boosting (XGB), Categorical Boosting (CB), and Random Forest (RF).

For conferring a notion of interpretability to the outputs predicted by the various models, we apply both Shapley values [15] and post-hoc permutation feature importance [16]. The latter is a global view, dependent on the decrease in model performance, i.e., ranks features whose exclusion will cause in an increase in model prediction error, either by its relation with the target or other predictor variables. Shapley values are rooted in additivity and monotonicity, which means i) the sum of local feature attributions equals the difference between base values and average values of the features, and ii) ignores features whose contributions to the final classification outcome are negligible. A summary plot obtained with Shapley values combines feature importance with an instance-based view. This allows for some insight into the local behavior of each feature, and its magnitude of attribution to each instance across target classes [15]. A certain level of agreement between both approaches can provide value to clinicians in defining which features were determined to be important by the model for a single classification outcome. This reduces the black-box nature of the deployed model and ensures that desirable or plausible patterns are mined instead of picking up on spurious relations or hidden biases.

Uncertainty estimation is vital to identify erroneous samples during training and mitigate confounding due to out-of-distribution samples during inference [17]. Essentially, if a sample has a predicted probability p, then based on its observed frequency in the dataset, its likelihood of belonging to its predicted class must also be p. Calibration is a model-agnostic point-estimation approach that lets the predicted probabilities be interpreted with an inherent confidence level. In this work, we use the non-parametric Spline-based probability calibration method [18] which relies on smoothing cubic splines instead of piecewise constants or sigmoid functions. This method was shown to be a better fit for multi-class problems, and dampen overconfident prediction probabilities in imbalanced datasets. The total uncertainty in any TB prediction depends on both data (aleatoric) and model (epistemic) uncertainty. As EHR data is prone to incompleteness and irregular values, it is beneficial to quantify the epistemic uncertainty as a baseline first, keeping the quality of the data consistent.

RESULTS

After employing a five-fold cross-validation approach, accuracy, sensitivity, specificity, F1-score, and Matthew's Correlation Coefficient [19] [20] were the performance measures used to ascertain the models' objective quality. The first three metrics and standard diagnostic measures used by clinicians, the latter two are used to note the balance between the cause of type-1 errors and the cause of type-2 errors, and aggregate score across true positives, true negatives, false positives, and false negatives in proportion to the number of samples respectively. The One-Versus-Rest classification process for all models was followed to elicit multi-class outcomes [21].

From Table 1, it can be seen that the highest accuracy, sensitivity, specificity, and MCC are obtained by CB, and F1-score is obtained by XGB. The accuracy sensitivity, specificity, F1-score, and MCCranges for almost all models are above 70%, 55%, 70%, 50%, and 25% with the standard deviations obtained being lesser than $\sim \pm 5\%$.. If three classes are considered with equal weight, then the baseline scores would be roughly 33%, which is the act of random prediction with no patterns learned. As such, it appears that the models have internalized a satisfactory knowledge representation for this dataset. In the interest of keeping a good trade-off between false positives and false negatives, based on the F1 score, we select CB for further post-hoc analysis.

Metrics to evaluate calibration quality estimated the differences between true confidence and predicted confidence. For the multi-class case, the quantification metrics are Log-Loss, Expected Calibration Error (ECE), Adaptive Calibration Error (ACE), and Static Calibration Error (SCE) [22]. 10% of the data was held out for testing with the rest

TABLE 1. Quantitative performance metrics across five-fold cross-validation.

Model	*Accuracy*	*Sensitivity*	*Specificity*	$F1-Score$	*MCC*
LGB	75.0	54.6	73.7	55.5	25.3
RF	77.1	55.3	74.1	54.6	30.5
XGB	76.5	56.1	74.2	56.3	29.1
KNN	70.8	44.1	71.0	45.7	14.1
SVC	76.3	54.9	72.5	49.4	28.9
LR	74.3	54.9	72.5	54.8	23.0
CB	77.4	56.1	74.1	55.2	32.0

being saved for calibration.

In Table 2, the calibration performance of the models is reported (lower the better). Reduction in error between predicted and observed probability estimates is obtained. Thus, given a new data instance for classification, statements such as "CB predicted chance of TB active in lungs with 60% accuracy", can be alleged with reasonable confidence.

TABLE 2. Uncertainty quantification measures before and after Spline-based calibration for CB.

Metric	Before Calibration	After Calibration
Log-Loss	0.96	0.87
ECE	0.08	0.05
ACE	0.18	0.16
SCE	0.13	0.09

Correlating both Figure 1 and Figure 2 in this cohort reveals the presence of BMI (kg/m^2), coughing, duration of cough, age, contact with TB patient, headache, duration of headache, recent weight loss status, recent weight loss duration, presence of night sweats, duration of night sweats, and a number of people in the room they reside in to be pertinent factors for classifying each of the three classes. There are also trace amounts of value given to a few other variables, but they could be inconclusive and as such, we consider only the primary indicators.

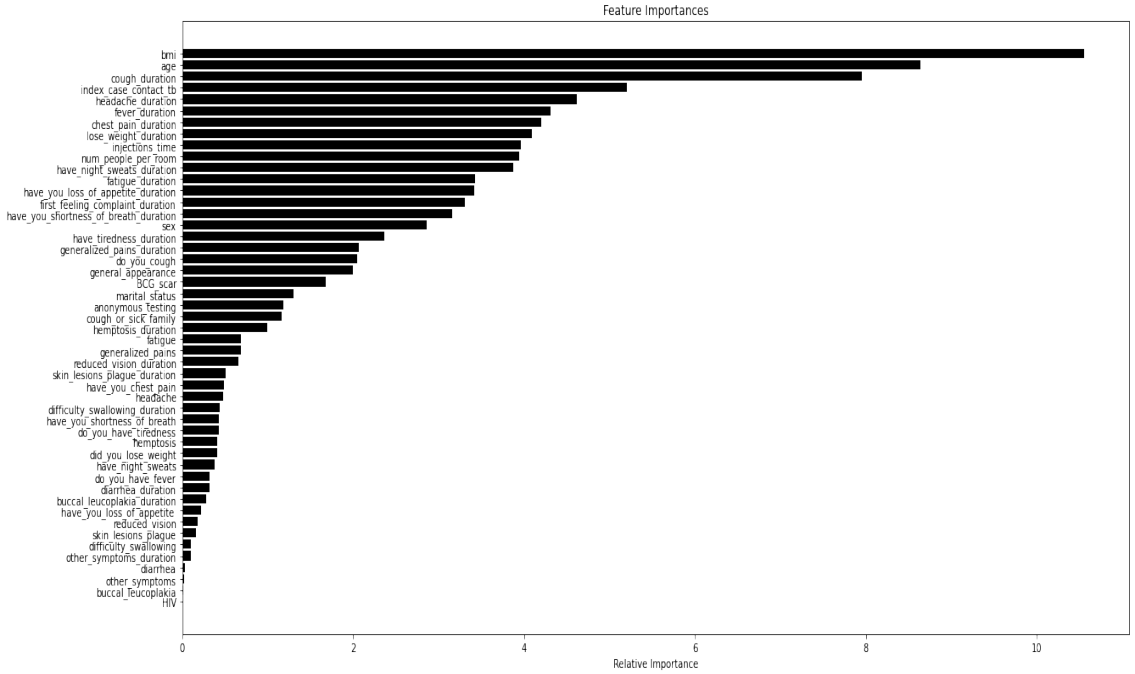

FIGURE 1. Permutation feature importance with CB.

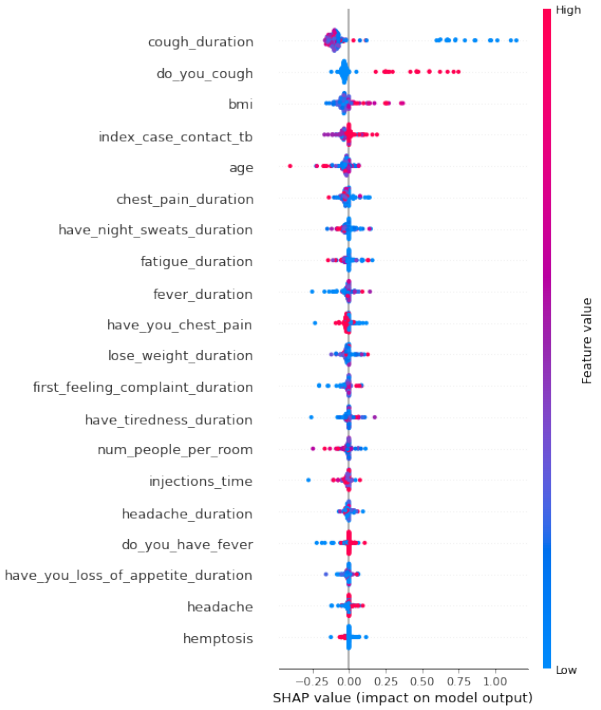

FIGURE 2. Shapley feature attributions with CB.

DISCUSSION

It is interesting to observe that despite the generality of the questions asked, the models are able to produce discernible classifications. Since there is a considerable overlap between the pathology of TB and other infections itself, differentiating between subtypes of TB with minimal physical examination samples like blood/sputum shows promise for public health monitoring. We hypothesize that ensemble models tended to fare better than traditional models because there appears to be no implicit linearity, and higher dimensional interactions are present among multiple predictor variables. Tree-based or iterative learning models handle perturbations better and are less affected by confounding effects of outliers [23]. Empirically, high duration of cough, younger ages, lower BMI, contact with confirmed patients, severe duration of headaches, long periods of night sweats, and a large number of people per room along with weight loss emerged as markers of active pulmonary distress owing to TB.

While coughing is not necessarily always an accompanying symptom of TB, the respiratory nature of the condition and the airborne transmission medium lend themselves to the prevalent general consensus that cough and prolonged contact with TB patients/dense crowds will increase the odds of infection [24] [25] [26]. Recently, [27] and [28] found the incidence of TB was higher in underweight individuals, which concurs with our findings (lower BMI). Lastly, both chronic headaches and increased night sweats are frequent dominant symptoms of TB, particularly when coupled with the previous symptoms [29] [30] [31]. However, other common TB factors such as fever, fatigue, and chest pain [32] were not deemed as vital as the aforementioned predictors. It is likely that this particular cohort did not experience them as strongly as otherwise recorded.

In terms of limitations, we mention that data collection procedures might not be standardized across developing countries, thereby limiting the application of the model. There is also the homogeneity of ethnicity, nation, and socioeconomic background which might hinder the generalizability of the model. Lastly, it is unknown if the participants were recovering or had medical conditions beyond what was reported, rendering them more susceptible to TB.

CONCLUSION

To the best of the authors' knowledge, this work is one of the first to explore the clinical utility of EHR containing routinely acquired questionnaires, symptoms, and demographic information to discriminate between multiple subtypes of TB through an interpretable view. The limitations in this work stem from the region-specific nature of the dataset which can minimize the generalizability. Future work can explore the harmonization of site-agnostic data and synthetic sampling to boost the utilization of the available records.

REFERENCES

1. "Tuberculosis (TB),".
2. I. Smith, "Mycobacterium tuberculosis Pathogenesis and Molecular Determinants of Virulence," **16**, 463–496, 12857778.
3. A. Bhargava and M. Bhargava, "Tuberculosis deaths are predictable and preventable: Comprehensive assessment and clinical care is the key," **19**, 100155, 32211519.
4. I. Barberis, N. Bragazzi, L. Galluzzo, and M. Martini, "The history of tuberculosis: From the first historical records to the isolation of Koch's bacillus," **58**, E9–E12, 28515626.
5. M. Pandiyan, O. El-Hassan, A. H. Khamis, and P. Rajasekaran, "Ontology with SVM Based Diagnosis of Tuberculosis and Statistical Analysis," **3**, 37–43.
6. J. Ramesh, N. Keeran, A. Sagahyroon, and F. Aloul, "Towards validating the effectiveness of obstructive sleep apnea classification from electronic health records using machine learning," in *Healthcare*, Vol. 9 (MDPI) p. 1450.
7. S. Muro, M. Ishida, Y. Horie, W. Takeuchi, S. Nakagawa, H. Ban, T. Nakagawa, and T. Kitamura, "Machine Learning Methods for the Diagnosis of Chronic Obstructive Pulmonary Disease in Healthy Subjects: Retrospective Observational Cohort Study," **9**, e24796, 34255684.
8. F. S. Heldt, M. P. Vizcaychipi, S. Peacock, M. Cinelli, L. McLachlan, F. Andreotti, S. Jovanović, R. Dürichen, N. Lipunova, R. A. Fletcher, A. Hancock, A. McCarthy, R. A. Pointon, A. Brown, J. Eaton, R. Liddi, L. Mackillop, L. Tarassenko, and R. T. Khan, "Early risk assessment for COVID-19 patients from emergency department data using machine learning," **11**, 4200, 33603086.
9. H. H. Rashidi, L. T. Dang, S. Albahra, R. Ravindran, and I. H. Khan, "Automated machine learning for endemic active tuberculosis prediction from multiplex serological data," **11**, 17900 ().
10. A. D. Orjuela-Cañón, A. L. Jutinico, C. Awad, E. Vergara, and A. Palencia, "Machine learning in the loop for tuberculosis diagnosis support," **10**, 876949, 35958865.
11. M. Asad, A. Mahmood, and M. Usman, "A machine learning-based framework for Predicting Treatment Failure in tuberculosis: A case study of six countries," **123**, 101944.
12. H. H. Rashidi, I. H. Khan, L. T. Dang, S. Albahra, U. Ratan, N. Chadderwala, W. To, P. Srinivas, J. Wajda, and N. K. Tran, "Prediction of Tuberculosis Using an Automated Machine Learning Platform for Models Trained on Synthetic Data," **13**, 10 (), 35136677.
13. S. Jamal, M. Khubaib, R. Gangwar, S. Grover, A. Grover, and S. E. Hasnain, "Artificial Intelligence and Machine learning based prediction of resistant and susceptible mutations in Mycobacterium tuberculosis," **10**, 5487.
14. S. A. Hassanain, J. K. Edwards, E. Venables, E. Ali, K. Adam, H. Hussien, and A. Elsony, "Conflict and tuberculosis in Sudan: A 10-year review of the National Tuberculosis Programme, 2004-2014," **12**, 18.
15. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, Vol. 30 (Curran Associates, Inc.).
16. C. Molnar, *Interpretable Machine Learning*.
17. C. F. Dietrich, *Uncertainty, Calibration and Probability: The Statistics of Scientific and Industrial Measurement* (Routledge).
18. B. Lucena, "Spline-Based Probability Calibration," arXiv:1809.07751 [cs, math, stat].
19. D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," **21**, 6.
20. S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, and S. Parasa, "On evaluation metrics for medical applications of artificial intelligence," **12**, 5979.
21. C. M. Bishop, *Pattern Recognition and Machine Learning*, Information Science and Statistics (Springer).
22. J. Nixon, M. Dusenberry, G. Jerfel, L. Zhang, and D. Tran, "Measuring Calibration in Deep Learning," .
23. M. Salama, H. Abdelkader, and A. Abdelwahab, "A novel ensemble approach for heterogeneous data with active learning," **14**, 18479790221082605.
24. R. Adigun and R. Singh, "Tuberculosis," in *StatPearls* (StatPearls Publishing) 28722945.
25. S. K. Field, P. Escalante, D. A. Fisher, B. Ireland, and R. S. Irwin, "Cough Due to TB and Other Chronic Infections," **153**, 467–497, 29196066.
26. B. Patterson and R. Wood, "Is cough really necessary for TB transmission?" **117**, 31–35, 31378265.
27. H.-H. Lin, C.-Y. Wu, C.-H. Wang, H. Fu, K. Lönnroth, Y.-C. Chang, and Y.-T. Huang, "Association of Obesity, Diabetes, and Risk of Tuberculosis: Two Population-Based Cohorts," **66**, 699–705, 29029077.
28. H. Choi, J. E. Yoo, K. Han, W. Choi, S. Y. Rhee, H. Lee, and D. W. Shin, "Body Mass Index, Diabetes, and Risk of Tuberculosis: A Retrospective Cohort Study," **8**, 739766, 34926543.
29. S. Kumar, R. Verma, R. K. Garg, H. S. Malhotra, and P. K. Sharma, "Prevalence and outcome of headache in tuberculous meningitis," **21**, 138–144, 27094524.
30. A. J. Viera, M. M. Bond, and S. W. Yates, "Diagnosing Night Sweats," **67**, 1019–1024.
31. E. W. Orenstein, "Tuberculosis: A comprehensive clinical reference," **10**, 80–81.
32. C. Lange and T. Mori, "Advances in the diagnosis of tuberculosis," **15**, 220–240.