Prediction of Tuberculosis from Clinical Data

Patrick Kaggwa

2024-02-24

# 1. Summary/Abstract

# 2. Introduction

Tuberculosis (TB) remains a significant global health challenge due to diagnostic limitations affecting both pediatric and adult populations, resulting in delayed diagnosis or misdiagnosis(**Xin2019?**). This delay impacts individual prognosis and community transmission. TB, caused by the bacterium Mycobacterium tuberculosis, primarily affects the lungs but can also involve other organs in its active form or remain latent without symptoms(**Bulled2020?**),. In 2020, TB ranked among the top 10 causes of death globally, with over 10 million new cases and 1.3 million deaths reported.

Africa bears a considerable burden, with an estimated quarter of new cases globally, resulting in approximately 417,000 deaths annually(**WHO2023?**). TB is a leading cause of death among HIV-infected patients in sub-Saharan Africa, compounded by the low sensitivity of the commonly used sputum smear microscopy, particularly in detecting TB among people living with HIV.

## 2.1 General Background Information

In Uganda, TB remains a significant public health concern, with an incidence of 330 cases per 100,000 people annually, including 136 new smear-positive cases per 100,000. Diagnostic delays persist due to health services often waiting for systematic symptoms before examining sputum smears, leading to missed opportunities for timely diagnosis.

The GeneXpert diagnostic tool offers superior sensitivity compared to sputum smear microscopy but is underutilized in Uganda due to its expense, resulting in over 41,000 undiagnosed cases annually. Even with existing diagnostic methods like blood tests or sputum tests, analysis times are prolonged, allowing culture-positive TB cases to go undetected.

## 2.2 Description of data and data source

The data comes from an going prospective cohort study conducted in Kampala, Uganda, involving 902 adults aged 15 to 65 years without evidence of previous infection with Mycobacterium tuberculosis (Mtb). The participants were recruited through community tuberculin skin test (TST) surveys and HIV/TB Care Clinics. Data collection included baseline evaluations, quarterly follow-ups for up to 18 months, questionnaires on TB exposure and alcohol consumption, and blood samples for Interferon Gamma Release Assay (IGRA) testing.

## 2.3 Questions/Hypotheses to be addressed

How can TB prediction be modeled using early clinical patient data?

What is the performance of the TB prediction models?

# 3. Methods

## 3.1 Study Design

The study was a retrospective study. This study was based on the secondary data, which was classified as clinical examination, patients’ history like alcohol use, smoking and diagnostics.

## 3.2 Study Population and Setting

The study population was the patients’ medical records between 2011 to 2018 who visited the hospital under an International Development Research Centre (IDRC) project. The project is currently running in China Friendship Hospital, Naguru.

## 3.3 Inclusion and Exclusion Criteria

The inclusion criteria was all patients who had cough for more than two weeks and the exclusion criteria was all adults living outside Kampala.

## 3.4 Data aquisition

The data was acquired from the TB clinic at Hospital Naguru patient records form patient forms and medical examination reports from the data records office. The TB dataset consisted of 2296 instances with 15 attributes.

## 3.5 Data import and cleaning

Our analysis began with thorough data preparation. Following a systematic approach, we ensured data integrity and reliability by sourcing and processing the dataset from an Excel file into R. To understand the data’s structure, we reviewed the codebook and conducted exploratory analyses, including summaries and visualizations. We then cleaned the data to enhance its quality for analysis. Character variables were converted to factors, and missing values were identified using visualizations and summaries and omitted. Finally, the data was strategically split into training and testing sets in an 70/30 ratio to prepare for model development and evaluation. These steps guaranteed a dataset ready for further analysis.

*Explain anything related to your statistical analyses.*

# 4. Results

## 4.1 Exploratory/Descriptive analysis

Data analysis was conducted using R Studio version RStudio 2023.09.1. Descriptive statistics (frequencies, mean, standard deviation (SD), median, interquartile range (IQR), and proportions) were used to present the baseline characteristics of participants. we used histograms, bar graphs, and box plots, to show how data was distributed.We did cross tabulation between the covariates and the outcome.A dataset of 2294 patients was analyzed. Amongst these patients, the gender distribution was 47.0% female and 53.0% male with mean age of 31 years.

table1 <- readRDS(here::here("results", "tables", "table1.rds"))  
table\_gt <- gtsummary::as\_gt(table1)  
table\_gt |> gt::as\_raw\_html()

| **Characteristic** | **N** | **N = 2,294**1 |
| --- | --- | --- |
| Gender | 2,294 |  |
| FEMALE |  | 1,089 (47%) |
| MALE |  | 1,205 (53%) |
| Asthma | 2,294 | 76 (3.3%) |
| Smoking Status | 2,294 | 605 (26%) |
| Alcohol | 2,294 | 1,495 (65%) |
| Fever Status | 2,294 | 1,802 (79%) |
| Weight Loss | 2,294 | 2,089 (91%) |
| Cough Status | 2,294 | 2,293 (100%) |
| Sputum Production | 2,294 | 2,276 (99%) |
| Blood in Sputum | 2,294 | 621 (27%) |
| CXR Conclusion | 2,294 |  |
| NEGATIVE |  | 1,193 (52%) |
| POSITIVE |  | 1,101 (48%) |
| Home Fuel | 2,294 | 1,427 (62%) |
| HIV Status | 2,294 |  |
| NEGATIVE |  | 833 (36%) |
| POSITIVE |  | 1,461 (64%) |
| TB | 2,294 | 951 (41%) |
| 1 n (%) | | |

Figure1. Age distribution by TB

#Figure1. Age distribution by TB  
knitr::include\_graphics(here("results","figures", "age\_boxplot.png"))

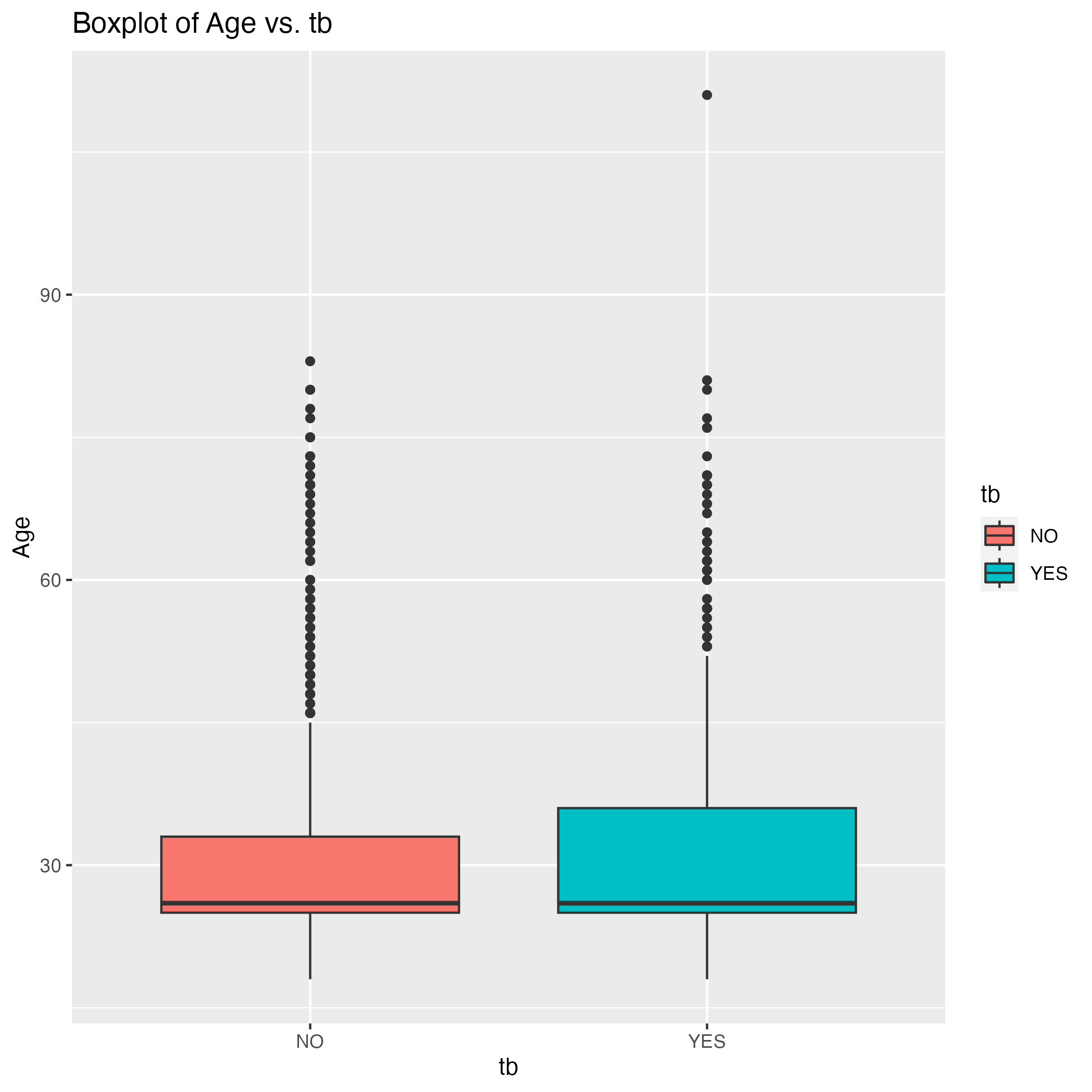


figure2. TB and Age

## 4.2 Basic statistical analysis

Model 1 which does not have cross validation has a very low ROC AUC value (12.21%), indicating poor discrimination between positive and negative instances. This suggests that Model 1’s performance in distinguishing between the two classes is not much better than random guessing. Model 2 which uses cross validation, on the other hand, exhibits a significantly higher ROC AUC value (86.78%), suggesting much better performance in distinguishing between the positive and negative instances. A value of 86.78% indicates strong discriminatory power, with higher true positive rates and lower false positive rates across various thresholds. talk about the model number of folds

#Figure1. Age distribution by TB knitr::include\_graphics(here(“results”,“figures”, “Roccurve.png”))

# 5. Discussion

## 5.1 Summary and Interpretation

## 5.2 Strengths and Limitations

## 5.3 Conclusions

# 6. References