European Journal of Radiology Open 11 (2023) 100518

Contents lists available at ScienceDirect

European Journal of Radiology Open

journal homepage: www.elsevier.com/locate/ejro

Predictive capabilities of baseline radiological findings for early and late disease outcomes within sensitive and multi-drug resistant tuberculosis cases

Gabriel Rosenfeld *, Andrei Gabrielian, Darrell Hurt, Alex Rosenthal

Office of Cyber Infrastructure and Computational Biology, National Institutes of Allergy and Infectious Diseases, 5601 Fishers Lane, Rockville, MD 20852, USA

ARTICLE INFO

Keywords:
X-ray
Radiographic features
Tuberculosis
Multi-drug resistant

ABSTRACT

Purpose: This study compares performance of Timika Score to standardized, detailed radiologist observations of Chest X rays (CXR) for predicting early infectiousness and subsequent treatment outcome in drug sensitive (DS) or multi-drug resistant (MDR) tuberculosis cases. It seeks improvement in prediction of these clinical events through these additional observations.

Method: This is a retrospective study analyzing cases from the NIH/NIAID supported TB Portals database, a large, trans-national, multi-site cohort of primarily drug-resistant tuberculosis patients. We analyzed patient records with sputum microscopy readings, radiologist annotated CXR, and treatment outcome including a matching step on important covariates of age, gender, HIV status, case definition, Body Mass Index (BMI), smoking, drug use, and Timika Score across resistance type for comparison.

Results: 2142 patients with tuberculosis infection (374 with poor outcome and 1768 with good treatment outcome) were retrospectively reviewed. Bayesian ANOVA demonstrates radiologist observations did not show greater predictive ability for baseline infectiousness (0.77 and 0.74 probability in DS and MDR respectively); however, the observations provided superior prediction of treatment outcome (0.84 and 0.63 probability in DS and MDR respectively). Estimated lung abnormal area and cavity were identified as important predictors underlying the Timika Score's performance.

Conclusions: Timika Score simplifies the usage of baseline CXR for prediction of early infectiousness of the case and shows comparable performance to using detailed, standardized radiologist observations. The score's utility diminishes for treatment outcome prediction and is exceeded by the usage of the detailed observations although prediction performance on treatment outcome decreases especially in MDR TB cases.

1. Introduction

The World Health Organization (WHO) estimates 10 million new cases of tuberculosis in 2020 with 1.5 million deaths, which is, unfortunately, the first increase in deaths since 2005 [1]. The COVID-19 pandemic has resulted in decreased reporting of cases and treatment for drug-resistant tuberculosis due to pandemic-related strains on healthcare systems globally [2]. The pandemic is viewed as one of the primary reasons for these recent setbacks with TB disease management. Recovery from the lingering effects of the pandemic necessitates decreasing the burden of poor outcomes from sensitive (DS) and drug resistant (DR) tuberculosis cases. Bacteriologically-confirmed pulmonary tuberculosis cases represent a significant risk of TB transmission to other persons in close proximity when the TB infected person generates airborne particles known as droplet nuclei created by common activities such as sneezing, coughing, singing, laughing, or even breathing [3].

Routine monitoring by microscopy or culture for the absence of detectable *M.tuberculosis* in sputum, sputum conversion, is performed throughout TB treatment when mycobacteria is detected at the start of treatment. Nevertheless, culture-negative *M. tuberculosis* has been reported in up to 15–25% of TB cases [4] representing a gap that could be filled by radiological monitoring for screening and detection of *M. tuberculosis*. Moreover, monitoring efforts via microscopy or culture are challenged by the difficulty of obtaining sputum from certain patients (children and those individuals with HIV infection) [5,6]. Genetic molecular diagnostic testing is filling the gap without necessarily

* Corresponding author.
E-mail address: gabriel.rosenfeld@nih.gov (G. Rosenfeld).

https://doi.org/10.1016/j.ejro.2023.100518
Received 22 June 2023; Received in revised form 2 August 2023; Accepted 10 August 2023

providing a quantitative assessment of the level of bacterial burden that microscopy or culture can provide. Without confirming the absence of mycobacteria during treatment, bacteriologically-confirmed TB patients remain infectious and pose a risk of transmission as well as opportunity for the spread of drug resistance. Radiology imaging is an alternative method that can help fill the gaps identified above by assessing the pathophysiology of the lung in response to the TB infection. The pathophysiological responses can help to diagnose TB and understand a patient's initial disease burden and transmission risk to others [7].

Clinicians have a need for radiologically-derived clinical scores that could serve as a useful adjunct to identify high-risk patients having a greater disease severity, transmission risk potential, as well as longer-term potential for detrimental treatment outcome. This is especially important in the growing number of DR cases endemic in certain parts of the world where early diagnosis, risk management, and treatment are key to containing the spread. The Timika Score was first reported in 2010 and has showed strong association with sputum smear grade and early disease burden [8–10]; however, its reported performance for treatment outcome has not been as strong [11]. Since earlier studies often focused on DS, it is imperative to reassess Timika Score especially in the context of endemic DR TB in many areas of the world. The question is whether improvements in Timika Score's reported performance for early disease burden and treatment outcome are possible and in what context, DS only, DR only, or both DS and DR?

The ability to study the relationship between radiologically-derived clinical scores and treatment outcome over time in both DS and DR TB cases requires a detailed dataset spanning these distinct domains of data and containing a sufficient number of cases. We leverage the TB Portals database [12–14], a publicly-accessible, patient-centric resource on both DS and DR TB that capture treatment outcomes, radiological findings (both by professional radiologists and artificial intelligence), bacteriology, and other important aspects of the patient's case. Using this unique combination of data, we assess the utility of baseline radiologist findings of CXRs (including Timika Score) to predict early disease burden as well as longer-term treatment outcome in DS and DR TB cases. The study is important because it investigates whether improvements in Timika Score are possible using a standardized, detailed set of radiologist observations beyond those that were originally used to derive the score. To our knowledge, this is the first study to assess baseline radiological findings by TB resistance profile to compare predictive ability for early and late clinical outcomes of interest during a TB patient's routine clinical care.

2. Materials and methods

This study adheres to the transparent reporting of a multivariate prediction model for individual prognosis or diagnosis (TRIPOD) guidelines [15].

2.1. Data source and access

2.1.1. Data source

All data was obtained from TB Portals, a trans-national collaboration led by the NIAID, covering forty sites from sixteen countries in Eastern Europe, Asia, and sub-Saharan African. The data warehouse currently contains over 8000 patients registered between the years of 2008–2022. Only publicly-available, de-identified data was used. Each participating clinical research institution (https://tbportals.niaid.nih.gov/where-do-our-cases-come-from) receives approval from the participating institution's IRB for public-sharing and follows strict adherence to ethics rules requirements of CRDF Global and the International Science and Technology Center who are the grant-issuing institutions [12].

2.1.2. Data access

The data was accessed using an available API service (https://data-sharing.tbportals.niaid.nih.gov/#/data-api) through an R package wrapper (https://niaid.github.io/tbportals.depot.api/).

2.2. Study Population

2.2.1. Inclusion and exclusion criteria

Registered patients diagnosed with DS or multidrug-resistant (MDR) TB having an outcome of treatment success ("completed" or "cured") or treatment failure ("died" or "failure"), an age of onset greater than 18 years old, available sputum microscopy testing before treatment initiation, a radiologist-annotated CXR within 90 days of treatment start and 30 days of sputum microscopy testing were included.

2.2.2. Matched population

To mitigate potential impacts of certain demographic or radiological factors that could bias comparison of predictive accuracy between resistance status, matching was performed on similar patient characteristics in the sensitive and MDR subgroups. Age, gender, HIV status, case definition, Timika Score, BMI, smoking use, and drug use were matched using the coarsened exact match algorithm [16].

2.3. Clinical outcomes

Two distinct clinical outcomes at the beginning and end of the patient's clinical history were considered. Sputum positivity prior to treatment was considered as the early outcome and treatment outcome was considered as the later outcome. Sputum positivity was defined as positive, if results were one of the following, "1–9 in 100 (1–9/100)", "10–99 in 100 (1 +)", "1–9 in 1 (2 +)", "10–99 in 1 (3 +)", "More than 99 in 1 (4 +)"; and negative (if results were "Negative") from the microscopy examination counting the amount of *M. tuberculosis* bacteria present in sputum. In situations where multiple microscopy results were available, the last available result was used. In situations where multiple microscopy results were available on the same day, the result with largest count of *M. tuberculosis* was used. Treatment response of the patient was defined as either treatment success ("completed" or "cured") or failure ("died" or "failure").

2.4. Predictors

2.4.1. Matching predictors

For the matching procedure, several attributes were used. A binary indicator of reported HIV status (1 for reported HIV, 0 for no reported HIV or missing) was derived from reported comorbidity. A binary indicator for smoking or drug use respectively (1 for reported use, 0 for no reported use or missing) was derived from the social risk factors attribute. BMI was categorized into "underweight" (<18.5), "healthy" (between 18.5 and 25), "overweight/obese" (> 25), or missing. Age, gender, case definition were used as described by the TB portals data model (https://datasharing.tbportals.niaid.nih.gov/#/about-the-data). Timika score was derived from the estimated overall volume of the lungs with any reported abnormality (0–100%) by adding 40 to this value in presence of a reported cavity [7,8].

2.4.2. Predictive modeling predictors

Predictors for modeling were derived from the radiologist findings of CXR. The TB portals describes the data model used for the manually annotated CXR (https://datasharing.tbportals.niaid.nih.gov/#/about-the-data). Some attributes are described at the whole lungs level (overall percent of abnormal volume, percent of hemithorax involvement, etc.) whereas other attributes are described at the sextant level (presence and size of cavity, presence and size of nodules, etc.). Timika score was derived as described in the matching procedure. Lung location was generated from location of reported sextant level findings ("upper" if occurring in either left or right upper sextant, "middle" if occurring in either left or right middle sextant, and "lower" if occurring in either left or right lower sextant). A binary indicator of cavity or nodule was

derived using the sextant level reporting for cavity or nodules respectively.

2.5. Missing data

2.5.1. Intention to treat analysis

The TB Portals database contains patient's treatment history including the final outcome of treatment. For the purpose of this study, only treatment success or failure were modeled excluding other available treatment outcomes (e.g. loss to follow up, Still on treatment, etc.). There is a risk of bias in the generalization of the predictive performance of the resulting models due to exclusion of patients with these alternative outcome. An intention to treat analysis was performed to assess the likelihood that bias would impact the conclusions from the modeling. The 262 patients removed during treatment outcome selection step were included and outcomes imputed as treatment success to assess impact on model predictive performance.

2.5.2. Chest X ray findings

An expert radiologist reviews the CXR image and provides findings in a standardized format required by the TB Portals program. Radiological findings were imputed as 0 for numerical attributes or "No" for binary "Yes/No" attributes if abnormalities were not identified by the radiologist. This is important since certain sextants of the lungs may not have any identifiable findings (e.g. upper left or lower right).

2.6. Modeling

2.6.1. Calculation of model performance within each patient subgroup

A nested cross validation (CV) stratified by outcome was performed within each subgroup (resistance type, matching, or intention to treat) to assess model predictive performance on clinical outcome of interest. Initially, the data is split into 10 equal sets termed outer folds, which are then split into a train and test set comprising 90% and 10% of the outer folder respectively. A model workflow was trained on the inner fold 90% training data and then tested on the 10% testing data using the tidymodels R package [17]. The workflow incorporated feature selection and class balancing via upsampling to ensure that the ROC metric accounted for these upstream processing steps during training; therefore, data leakage or unfair estimation were mitigated. The seed to select each fold was kept constant when modeling each event to identify identical groups of patients for fold1, fold2, etc. to facilitate subsequent use of the fold ROC data for Bayesian ANOVA analysis.

We trained a logistic regression model and random forest model using distinct workflows that leveraged all available features excluding Timika Score, only the Timika Score, or a domain-expert informed set of features [e.g., Timika Score, lung location (upper, middle, lower), and presence of nodules]. In the logistic model workflows, we included a near zero variance exclusion step (https://recipes.tidymodels.org/reference/step_nzv.html) to remove sparse variables and a Minimum Redundancy Maximum Relevancy (https://stevenpawley.github.io/recipeselectors/reference/step_select_mrmr.html) step to include only the features showing the most correlation with outcome but least correlation with each other. For the Random Forest workflows, we include all variables as the algorithm selects the most important variables during modeling. All analyses were done as a targets workflow [18] for enhanced reproducibility using R version 4.2.1. The code for the analysis can be found at https://github.com/niaid/tbportals.timika.comparison.

2.6.2. Bayesian ANOVA

With the tidyposterior R package [19], we calculate the probability density of the model ROC by Bayesian analysis (https://tidyposterior.tidymodels.org/reference/perf_mod.html) using the observed performance for the best models. In doing so, we could posit questions such as "what is the probability that Timika Score can more reliably predict early versus late outcomes within DS cases" or "what is the probability that additional radiologist findings can more reliably predict treatment outcome versus Timika Score in MDR TB cases". To address these probabilistic questions, we defined a region of practical equivalence (ROPE) as 0.02 ROC meaning that any resulting probability describes boundaries of this threshold in the posterior distributions (e.g., 0.02 lower, 0.03 equivalent, and 0.95 higher probabilities corresponding to a comparison of Model A versus Model B indicates that the most probable result is that Model A's ROC exceeds Model B's by at least 0.02 ROC at 0.95 probability and we can reject null hypothesis of both falling within the ROPE).

3. Results

3.1. Study population and outcomes

2142 patients were included in the study population to model early and late clinical outcomes. Table 1 shows the loss of patients at each step of the inclusion criteria with 1325 matched patients identified for the matched analyses. After matching for age, gender, HIV, case definition, Timika Score, BMI, smoking, and drug use, matched covariates show a negligible absolute standardized mean difference between the sensitive and MDR cases of less than 0.1 (Supplementary Figure 1). The patient characteristics of the unmatched (Table 2) and matched (Table 3) study populations demonstrate observed relationships between the events of interest by resistance type. These observations are consistent with the matching procedure balancing the matched covariates between the DS and MDR TB subgroups.

The unmatched MDR subgroup in Table 2 shows a lower percentage of sputum negative microscopy results (29.8% of MDR versus 36.8% of DS) and treatment success (76.3% of MDR versus 89% of DS). Moreover, there is a greater percentage of relapse cases or other case definitions suggesting previous history of TB infection (33.7% of MDR versus 12% of DS). Certain comorbidities and risk factors such as reported HIV infection, smoking, alcohol and drug use were higher in the unmatched MDR population. In the matched population, these differences between MDR and DS patients in Table 3 are reduced. For example, reported HIV was reduced from 15.3% of MDR versus 3.3% of DS in the unmatched to 5.1% of MDR versus 1.8% of DS in the matched. Case definition suggesting previous history of TB was reduced from 33.7% of MDR versus 12% of DS in the unmatched to 13.1% of MDR versus 6.4% of DS after matching and mean BMI was equivalent at 20.5. Demonstrating successful matching is important: if predictions are consistent in both unmatched and matched populations, this supports that findings are generalizable to new cohorts of DS and MDR TB patients.

Table 1
Inclusion criteria.

Case Criteria	Number of Cases Meeting Criteria	Number of Excluded Cases
All public TB Portals cases	8779	0
Age greater than 18 years	8643	-136
Has sputum microscopy result prior to treatment	6981	-1662
Has annotated CXR within 90 days of treatment start	4089	-2892
Has annotated CXR within 30 days prior to sputum microscopy	3109	-980
Clinically reported type of resistance is DS or MDR	2404	-705
Treatment outcome is one of Completed, Cured, Died, or Failure	2142	-262
Matched clinical resistance on key covariates such as age, gender, etc.	1325	-817

Overview of stepwise cohort selection strategy used to select the final study population. The inclusion criteria is shown in the case criteria column along with the identified number of cases included and excluded at each step which occurs consecutively.

Table 2
Unmatched study population.

covariate	MDR (N = 1067)	Sensitive (N = 1075)	Total (N = 2142)
* *microscopyresults* *			
Negative	318 (29.8%)	396 (36.8%)	714 (33.3%)
1–9 in 100 (1–9/100)	157 (14.7%)	96 (8.9%)	253 (11.8%)
10–99 in 100 (1 +)	251 (23.5%)	276 (25.7%)	527 (24.6%)
1–9 in 1 (2 +)	136 (12.7%)	146 (13.6%)	282 (13.2%)
10–99 in 1 (3 +)	199 (18.7%)	130 (12.1%)	329 (15.4%)
More than 99 in 1 (4 +)	6 (0.6%)	31 (2.9%)	37 (1.7%)
* *outcome* *			
Completed	137 (12.8%)	75 (7.0%)	212 (9.9%)
Cured	677 (63.4%)	879 (81.8%)	1556 (72.6%)
Died	190 (17.8%)	35 (3.3%)	225 (10.5%)
Failure	63 (5.9%)	86 (8.0%)	149 (7.0%)
* *timika_score* *			
Mean (SD)	46.492 (35.805)	49.222 (31.406)	47.862 (33.689)
Range	0.000–140.000	0.000–140.000	0.000–140.000
* *upper* *			
FALSE	99 (9.3%)	127 (11.8%)	226 (10.6%)
TRUE	968 (90.7%)	948 (88.2%)	1916 (89.4%)
* *middle* *			
FALSE	396 (37.1%)	337 (31.3%)	733 (34.2%)
TRUE	671 (62.9%)	738 (68.7%)	1409 (65.8%)
* *lower* *			
FALSE	684 (64.1%)	636 (59.2%)	1320 (61.6%)
TRUE	383 (35.9%)	439 (40.8%)	822 (38.4%)
* *cavity* *			
FALSE	570 (53.4%)	567 (52.7%)	1137 (53.1%)
TRUE	497 (46.6%)	508 (47.3%)	1005 (46.9%)
* *nodule* *			
FALSE	154 (14.4%)	297 (27.6%)	451 (21.1%)
TRUE	913 (85.6%)	778 (72.4%)	1691 (78.9%)
* *registration_date* *			
Mean (SD)	2019.870 (1.600)	2019.638 (1.518)	2019.754 (1.564)
Range	2011.000–2022.000	2010.000–2022.000	2010.000–2022.000
* *age_of_onset* *			
Mean (SD)	43.420 (12.367)	44.477 (14.989)	43.951 (13.752)
Range	18.000–85.000	18.000–88.000	18.000–88.000
* *gender* *			
Female	254 (23.8%)	285 (26.5%)	539 (25.2%)
Male	813 (76.2%)	790 (73.5%)	1603 (74.8%)
* *country* *			
Azerbaijan	1 (0.1%)	0 (0.0%)	1 (0.0%)
Belarus	56 (5.2%)	15 (1.4%)	71 (3.3%)
Georgia	139 (13.0%)	621 (57.8%)	760 (35.5%)
Kazakhstan	82 (7.7%)	4 (0.4%)	86 (4.0%)
Moldova	75 (7.0%)	165 (15.3%)	240 (11.2%)
Romania	21 (2.0%)	8 (0.7%)	29 (1.4%)
Ukraine	693 (64.9%)	262 (24.4%)	955 (44.6%)
* *education* *			
Basic school (incl. primary)	280 (26.2%)	149 (13.9%)	429 (20.0%)
College (bachelor)	161 (15.1%)	51 (4.7%)	212 (9.9%)
Complete school (a-level, gymnasium)	255 (23.9%)	219 (20.4%)	474 (22.1%)
Higher (university)	60 (5.6%)	12 (1.1%)	72 (3.4%)
No education	7 (0.7%)	1 (0.1%)	8 (0.4%)
Not Reported	304 (28.5%)	643 (59.8%)	947 (44.2%)
* *employment* *			
Disabled	55 (5.2%)	18 (1.7%)	73 (3.4%)
Employed	222 (20.8%)	195 (18.1%)	417 (19.5%)
Homemaker	8 (0.7%)	3 (0.3%)	11 (0.5%)
Not Reported	15 (1.4%)	11 (1.0%)	26 (1.2%)
Retired	76 (7.1%)	71 (6.6%)	147 (6.9%)
Self-employed	9 (0.8%)	5 (0.5%)	14 (0.7%)
Student	9 (0.8%)	5 (0.5%)	14 (0.7%)
Unemployed	654 (61.3%)	766 (71.3%)	1420 (66.3%)
Unofficially employed	19 (1.8%)	1 (0.1%)	20 (0.9%)
* *case_definition* *			
Chronic TB	25 (2.3%)	4 (0.4%)	29 (1.4%)
Failure	40 (3.7%)	7 (0.7%)	47 (2.2%)
Lost to follow up	24 (2.2%)	13 (1.2%)	37 (1.7%)
New	707 (66.3%)	946 (88.0%)	1653 (77.2%)
Other	16 (1.5%)	11 (1.0%)	27 (1.3%)
Relapse	253 (23.7%)	93 (8.7%)	346 (16.2%)
Unknown	2 (0.2%)	1 (0.1%)	3 (0.1%)
* *bmi* *			
N-Miss	66	627	693

(*continued on next page*)

Table 2 (*continued*)

covariate	MDR (N = 1067)	Sensitive (N = 1075)	Total (N = 2142)
Mean (SD)	20.310 (3.567)	20.517 (3.789)	20.374 (3.637)
Range	11.100–44.900	12.300–37.400	11.100–44.900
* *hiv* *			
FALSE	904 (84.7%)	1040 (96.7%)	1944 (90.8%)
TRUE	163 (15.3%)	35 (3.3%)	198 (9.2%)
* *risk_smoker* *			
Mean (SD)	0.601 (0.490)	0.158 (0.365)	0.379 (0.485)
Range	0.000–1.000	0.000–1.000	0.000–1.000
* *risk_alcohol* *			
Mean (SD)	0.278 (0.448)	0.097 (0.296)	0.187 (0.390)
Range	0.000–1.000	0.000–1.000	0.000–1.000
* *risk_drug* *			
Mean (SD)	0.059 (0.236)	0.003 (0.053)	0.031 (0.173)
Range	0.000–1.000	0.000–1.000	0.000–1.000

Overview of the unmatched study population prior to final matching step. Values indicate the count of patients with parenthesis indicating percentage unless otherwise indicated by a Mean (SD) or Range for numerical data.

3.2. Modeling

3.2.1. Identifying the best performing models

We use the median ROC from the CV results to identify the best performing models across DS and MDR TB patients. Identifying these models permitted us to formally test them later on by Bayesian analysis. CV results quickly established general trends in prediction performance across different models and prediction tasks. Interestingly, we noted better performance for predicting baseline sputum positivity in MDR TB compared to DS while the opposite was observed for treatment outcome. We compared the CV results from the unmatched populations since these correspond to the real-world characteristics of the selected cases in TB Portals. Most models within each prediction task and TB patient cohort perform similarly as evidenced by the range of ROC values. ROC between 0.6 and 0.63 and 0.66–0.71 were observed for baseline sputum positivity prediction in DS and MDR cases respectively. For treatment outcome, observed ROC values fell between the ranges of 0.58–0.69 and 0.58–0.64 in DS and MDR cases respectively. Model predictive performance for the two clinical outcomes is shown in Supplementary Table 1. Besides median, we report a number of other statistics on the performance of the CV ROC and also show performance on matched data. Model performance results demonstrate that matching had a modest effect if any at all on model performance.

3.2.1.1. Identifying the best models for DS. The regression model including Timika Score, nodule, and lung location performed the best when predicting treatment outcome with a median ROC of 0.69. For predicting baseline sputum positivity, the logistic regression model using Timika Score performed the best with a median ROC of 0.63. Most models performed equivalently with matching having minimal impact. The best performing model using matched data had 0.72 and 0.66 ROC for treatment outcome and baseline sputum positivity respectively.

3.2.1.2. Identifying the best models for MDR TB. Median ROC was lower for prediction of treatment outcome in general in the MDR TB cohort. Logistic regression using detailed radiologist findings performed the best with an ROC of 0.64. Matching had a modest impact on model performance for the logistic regression models as the median ROC decreased by ~0.05. Logistic regression using Timika Score demonstrated the best performance for predicting baseline sputum positivity with a ROC of 0.71. Predictive performance for baseline positivity was generally better than for sensitive TB patients. Matching showed a modest impact: ROC improved in the matched cohort for certain algorithms. The random forest model using detailed radiologist observations improved from 0.67 to 0.72 ROC for instance. This impact was smaller for other algorithms such as the logistic regression model using Timika Score which improved from 0.71 to 0.72 ROC.

3.2.2. Comparing the best performing models

We next identified whether the best performing models using Timika Score exceeded the performance of those including detailed radiologist observations. Comparisons of the best performing models with probabilities of each conclusion is shown in Supplementary Table 2. These comparisons identified with high likelihood that Timika Score performs equivalently or better than using detailed radiologist findings in TB Portals in both DS (0.77 probability) and MDR (0.74 probability) patients for baseline sputum positivity. We observed that this was not the case for treatment outcome. The detailed radiologist findings could improve prediction performance over Timika Score with 0.84 and 0.63 probability for DS and MDR respectively. Moreover, we noted a decrease in performance of Timika Score when predicting treatment outcome compared to its stronger performance for baseline positivity especially in MDR TB (0.96 probability). To make the above comparisons, we calculated the probability that the observed CV model performance exceeded a ROPE of 0.02 by Bayesian ANOVA analysis. A representative example of the Bayesian ANOVA results is shown in Fig. 1. By comparing these densities using the ROPE threshold, conclusions on the above differences could be made with associated probabilities for each.

3.2.2.1. Findings from comparison of the best models in DS. When predicting baseline positivity, Timika Score is equivalent to or exceeds modeling using detailed radiologist observations with a probability of 0.77. For predicting treatment outcome, including detailed radiologist findings exceeded Timika Score with a probability of 0.84. Timika Score's predictive performance for baseline positivity compared to treatment outcome did not differ significantly with 0.47 probability. There is a higher degree of uncertainty in the previous assessment. Matching did not have a substantial effect on the conclusions. As an example, we assessed if Timika Score is equivalent to detailed radiologist observations for predicting baseline positivity. We calculated a probability of 0.54 versus 0.46 in unmatched and matched DS patients respectively noting that the probabilities are similar. Importantly, the intention to treat (IIT) analysis indicated that conclusions are not significantly impacted by drop-out prior to treatment completion (Supplementary Table 3). For instance, we analyzed the probability of Timika score performing equivalently to detailed radiologist observations for baseline positivity in the IIT analysis. We noted a probability of 0.54 in unmatched analysis as compared to 0.53 in the IIT results.

3.2.2.2. Findings from comparison of the best models in MDR TB. Timika Score demonstrated equivalent or better prediction of baseline positivity compared to detailed radiologist observations (0.74 probability) as shown in supplementary table 2. For treatment outcome, detailed radiologist findings outperformed Timika Score with 0.63 probability. We also compared predictive performance of Timika Score for baseline

Table 3
Matched study population.

covariate	MDR (N = 604)	Sensitive (N = 721)	Total (N = 1325)
* *microscopyresults* *			
Negative	176 (29.1%)	265 (36.8%)	441 (33.3%)
1–9 in 100 (1–9/100)	90 (14.9%)	66 (9.2%)	156 (11.8%)
10–99 in 100 (1 +)	157 (26.0%)	183 (25.4%)	340 (25.7%)
1–9 in 1 (2 +)	78 (12.9%)	99 (13.7%)	177 (13.4%)
10–99 in 1 (3 +)	99 (16.4%)	93 (12.9%)	192 (14.5%)
More than 99 in 1 (4 +)	4 (0.7%)	15 (2.1%)	19 (1.4%)
* *outcome* *			
Completed	81 (13.4%)	60 (8.3%)	141 (10.6%)
Cured	420 (69.5%)	583 (80.9%)	1003 (75.7%)
Died	82 (13.6%)	17 (2.4%)	99 (7.5%)
Failure	21 (3.5%)	61 (8.5%)	82 (6.2%)
* *timika_score* *			
Mean (SD)	41.518 (32.867)	44.345 (31.363)	43.057 (32.076)
Range	0.000–130.000	0.000–140.000	0.000–140.000
* *upper* *			
FALSE	59 (9.8%)	103 (14.3%)	162 (12.2%)
TRUE	545 (90.2%)	618 (85.7%)	1163 (87.8%)
* *middle* *			
FALSE	245 (40.6%)	244 (33.8%)	489 (36.9%)
TRUE	359 (59.4%)	477 (66.2%)	836 (63.1%)
* *lower* *			
FALSE	419 (69.4%)	431 (59.8%)	850 (64.2%)
TRUE	185 (30.6%)	290 (40.2%)	475 (35.8%)
* *cavity* *			
FALSE	345 (57.1%)	416 (57.7%)	761 (57.4%)
TRUE	259 (42.9%)	305 (42.3%)	564 (42.6%)
* *nodule* *			
FALSE	95 (15.7%)	197 (27.3%)	292 (22.0%)
TRUE	509 (84.3%)	524 (72.7%)	1033 (78.0%)
* *registration_date* *			
Mean (SD)	2019.821 (1.629)	2019.671 (1.535)	2019.740 (1.580)
Range	2011.000–2022.000	2010.000–2022.000	2010.000–2022.000
* *age_of_onset* *			
Mean (SD)	43.043 (11.994)	41.480 (13.776)	42.192 (13.013)
Range	18.000–83.000	18.000–86.000	18.000–86.000
* *gender* *			
Female	134 (22.2%)	180 (25.0%)	314 (23.7%)
Male	470 (77.8%)	541 (75.0%)	1011 (76.3%)
* *country* *			
Belarus	35 (5.8%)	13 (1.8%)	48 (3.6%)
Georgia	85 (14.1%)	357 (49.5%)	442 (33.4%)
Kazakhstan	43 (7.1%)	3 (0.4%)	46 (3.5%)
Moldova	50 (8.3%)	126 (17.5%)	176 (13.3%)
Romania	11 (1.8%)	8 (1.1%)	19 (1.4%)
Ukraine	380 (62.9%)	214 (29.7%)	594 (44.8%)
* *education* *			
Basic school (incl. primary)	148 (24.5%)	118 (16.4%)	266 (20.1%)
College (bachelor)	98 (16.2%)	41 (5.7%)	139 (10.5%)
Complete school (a-level, gymnasium)	146 (24.2%)	170 (23.6%)	316 (23.8%)
Higher (university)	39 (6.5%)	11 (1.5%)	50 (3.8%)
No education	2 (0.3%)	1 (0.1%)	3 (0.2%)
Not Reported	171 (28.3%)	380 (52.7%)	551 (41.6%)
* *employment* *			
Disabled	25 (4.1%)	12 (1.7%)	37 (2.8%)
Employed	144 (23.8%)	143 (19.8%)	287 (21.7%)
Homemaker	4 (0.7%)	3 (0.4%)	7 (0.5%)
Not Reported	10 (1.7%)	10 (1.4%)	20 (1.5%)
Retired	35 (5.8%)	41 (5.7%)	76 (5.7%)
Self-employed	8 (1.3%)	3 (0.4%)	11 (0.8%)
Student	6 (1.0%)	4 (0.6%)	10 (0.8%)
Unemployed	369 (61.1%)	504 (69.9%)	873 (65.9%)
Unofficially employed	3 (0.5%)	1 (0.1%)	4 (0.3%)
* *case_definition* *			
Chronic TB	4 (0.7%)	3 (0.4%)	7 (0.5%)
Lost to follow up	1 (0.2%)	1 (0.1%)	2 (0.2%)
New	525 (86.9%)	675 (93.6%)	1200 (90.6%)
Relapse	74 (12.3%)	42 (5.8%)	116 (8.8%)
* *bmi* *			
N-Miss	31	359	390
Mean (SD)	20.508 (3.367)	20.530 (3.580)	20.516 (3.449)
Range	11.100–44.900	12.900–37.400	11.100–44.900
* *hiv* *			
FALSE	573 (94.9%)	708 (98.2%)	1281 (96.7%)

(*continued on next page*)

Table 3 (*continued*)

covariate	MDR (N = 604)	Sensitive (N = 721)	Total (N = 1325)
TRUE	31 (5.1%)	13 (1.8%)	44 (3.3%)
* *risk_smoker* *			
Mean (SD)	0.550 (0.498)	0.205 (0.404)	0.362 (0.481)
Range	0.000–1.000	0.000–1.000	0.000–1.000
* *risk_alcohol* *			
Mean (SD)	0.260 (0.439)	0.128 (0.334)	0.188 (0.391)
Range	0.000–1.000	0.000–1.000	0.000–1.000
* *risk_drug* *			
Mean (SD)	0.012 (0.107)	0.003 (0.053)	0.007 (0.082)
Range	0.000–1.000	0.000–1.000	0.000–1.000

Overview of the matched study population after final matching step. Values indicate the count of patients with parenthesis indicating percentage unless otherwise indicated by a Mean (SD) or Range for numerical data.

positivity versus treatment outcome. Timika Score could more reliably predict baseline positivity with 0.96 probability. Matching did not have a substantial effect on predictions except for a modest effect on treatment outcome. It enhanced the likelihood that Timika Score performance was equivalent or exceeded detailed radiologist findings. We noted this increase from a calculation of 0.37 versus 0.71 probability in unmatched and matched MDR TB patients respectively. The IIT analysis revealed that conclusions regarding the predictive comparisons were not influenced by drop-out prior to treatment completion (Supplementary Table 3). Like in DS, we calculated the probability of Timika score performing equivalently to detailed radiologist observations for baseline positivity by IIT analysis. We calculated a probability of 0.55 in unmatched analysis as compared to 0.8 in the IIT analysis. In both comparisons, these were the most likely conclusions. The results from both DS and MDR modeling demonstrate the value of Timika Score for predicting baseline positivity and the potential for additional radiologist findings to enhance prediction of treatment outcome. The additional efforts to explore these comparisons with matching and IIT analyses indicate that the results are robust and likely to generalize well.

3.2.3. *Interpretable machine learning*

After observing comparable performance of Timika Score with detailed radiologist observations for baseline positivity but not for treatment outcome, we endeavored to understand why through interpretable machine learning (IML) approaches of predictor importance and partial dependence plots. The most important predictor across DS (Figs. 2A and 3A) and MDR (Figs. 2C and 3C) groups was estimated overall percent of abnormal volume of the lung. Presence of cavity was also identified as important for prediction of baseline sputum positivity; however, it was not found among the top predictors of treatment outcome. We also support this finding by comparing its effect on prediction using partial dependence plots. Presence of cavity strongly influences the model to predict baseline positivity across different estimates of overall abnormal lung volume (Fig. 2B and D); however, its influence wanes for prediction of treatment outcome (Fig. 3B and D). Presence of nodules of various sizes and opacity were among the top predictors for baseline positivity and treatment outcome across both DS and MDR patients. Interestingly, percent of pleural effusion of hemithorax was identified as a top predictor for treatment outcome in MDR patients (Fig. 3C).

These observations are consistent with the difference in Timika Score's predictive performance for baseline sputum microscopy positivity and treatment outcome that we observed by Bayesian ANOVA. They suggest that other baseline radiological factors provide additional predictive capability for treatment outcome. One example previously mentioned in MDR patients is the percent of pleural effusion of hemithorax whereas another are the nodules of various size and opacity. Nonetheless, Timika Score demonstrated some predictive capacity towards treatment outcome since overall abnormal lung volume is the most important factor identified in both predictive tasks. Moreover, the above results suggest that for treatment outcome prediction, extent of severity (e.g., greater than 50% lung abnormal) is important especially in MDR patients. This is consistent with other studies showing extent of infection or delay to treatment having an important impact on treatment outcome [13,20].

3.2.3.1. Findings in baseline sputum positivity. We first examined predictions of machine learning for baseline sputum microscopy positivity. The most important features for the random forest model workflow using all radiological findings are shown in Fig. 2. Given the comparable performance of Timika Score with all other predictors for this prediction task, we hypothesized that overall abnormal lung volume and presence of cavity would be among the top features. The feature importance analysis in Fig. 2A and C confirmed this hypothesis demonstrating that overall abnormal lung volume was the top predictive feature and presence of cavity was also among the top 10 important features. To explore how overall abnormal lung volume and cavity interaction impacted the model's predictions, we generated partial dependence plots for these two predictors in Fig. 2B and D. Presence of cavity increased the marginal probability of the model to predict an outcome of baseline sputum microscopy positive by ~0.15–0.2 across the range of overall abnormal volume reported indicating that presence of a cavity strongly influenced the model to predict a positive baseline sputum microscopy. Moreover, overall abnormal lung volume increase led to a higher marginal probability of predicting a positive baseline microscopy up to ~ 25% abnormal volume of the lungs where it plateaued; therefore, prediction of positive baseline sputum microscopy increases sharply for the first reported 25% of overall lung abnormal area and further reported increases beyond this area does not have much of an added effect.

3.2.3.2. Findings in treatment outcome. We next examined the machine learning predictions for the subsequent treatment outcome. The most important features for the random forest model workflow using all radiological findings are shown in Fig. 3. Timika Score did not show comparable performance for this task so we hypothesized that other features might be identified among the most important features as shown in Fig. 3A and C. Interestingly, overall abnormal lung volume was still identified as the most important feature but presence of cavity was no longer among the top 10 important features. Presence of cavity still increased the marginal probability of treatment failure in the Partial Dependence Plots; however, the increases across the range of reported overall abnormal volume values were less than what were observed when predicting baseline microscopy positivity and varied by reported abnormal lung volume (from a difference of ~0.02 to ~0.05 increase in probability). Moreover, the reported overall abnormal volume of the lung showed a more modest increase in marginal probability of poor outcome in Sensitive (increase of ~0.02 probability from 0% to 100% reported overall abnormal lung volume). In MDR, the probability of predicting poor outcome increased linearly by ~0.1 probability after the reported overall abnormal lung volume exceeded 50%.

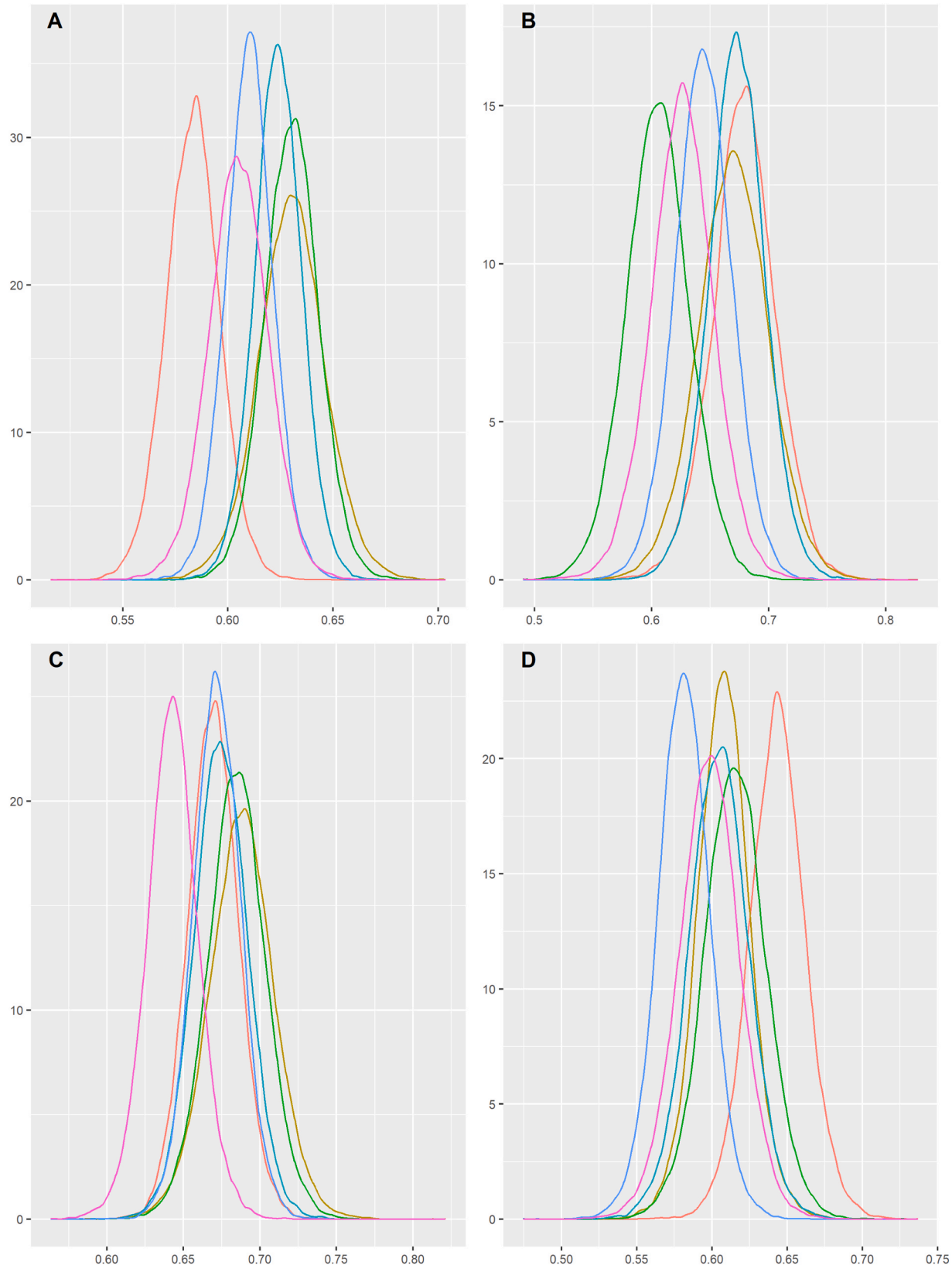

Fig. 1. Bayesian ANOVA estimation of model performance in unmatched cohort. The posterior distributions of the ROC metric calculated from the Bayesian ANOVA modeling of the different modeling workflows. Above example is shown for the unmatched subgroups and is representative of what can be generated for the other subgroups of interest. Different colors represent the posterior distributions of distinct workflows. Red: Logistic Regression using all variables, Yellow: Random Forest using all variables, Green: Logistic Regression using only Timika Score, Light Blue: Logistic Regression using Timika plus lung location and nodule present (Timika Plus), Dark Blue: Random Forest using Timika Plus, Pink: Random Forest using Timika Score. A) Prediction of the baseline sputum microscopy positivity in Sensitive TB patients. B) Prediction of the treatment outcome in Sensitive TB patients. C) Prediction of the baseline sputum microscopy positivity in MDR TB patients. D) Prediction of the treatment outcome in MDR TB patients.

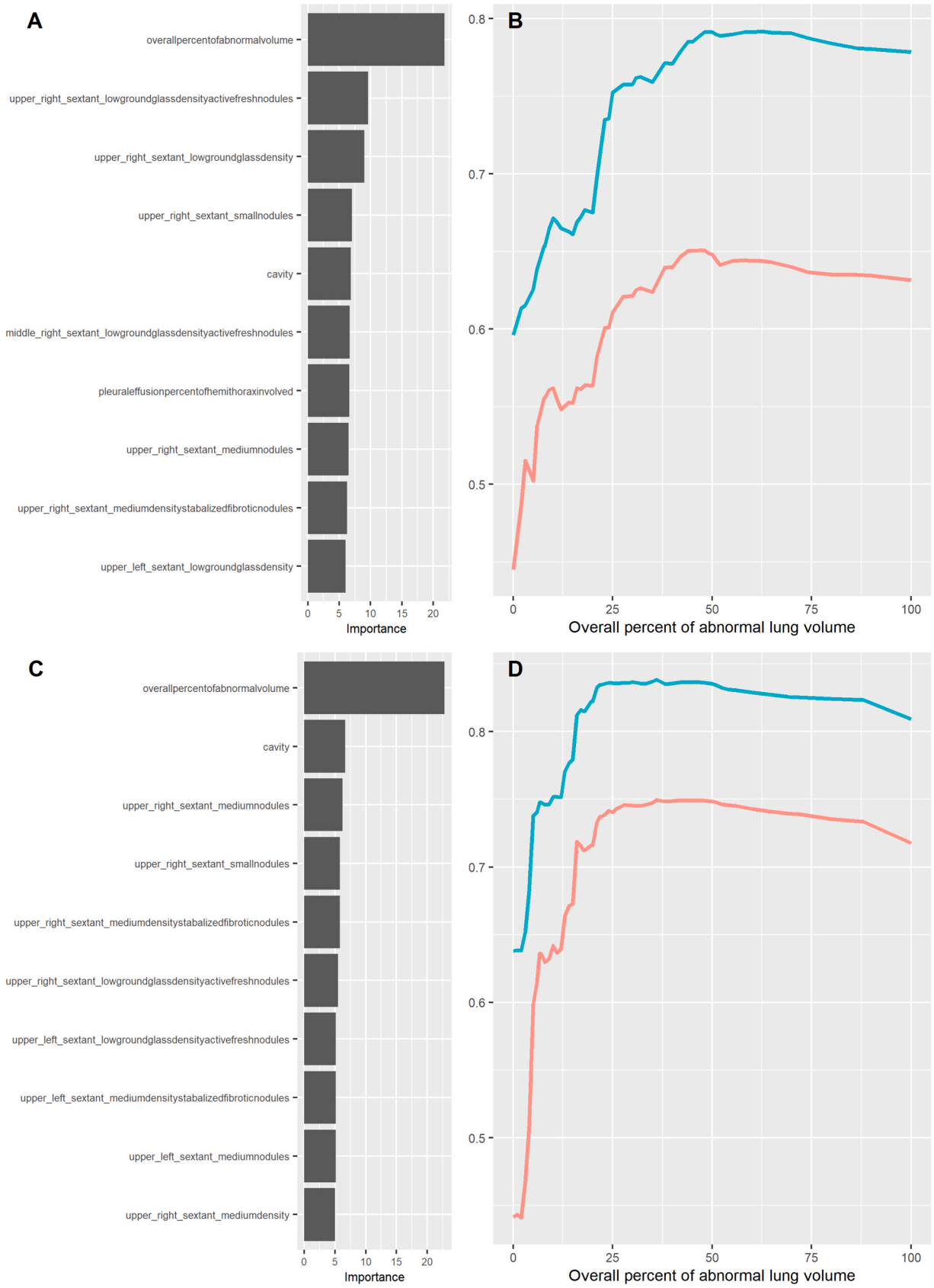

Fig. 2. Important radiological predictors from the Random Forest model to predict baseline microscopy positivity in matched Sensitive TB and MDR TB patients. The most important features from the random forest workflow trained on the matched population of Sensitive and MDR TB patients is shown for A) Sensitive TB Patients and C) MDR TB Patients. Partial Dependence Plot showing the predictive relationship between overall abnormal volume of lung and cavity status is shown for the same random forest workflow for B) Sensitive TB Patients and D) MDR TB Patients. Presence of cavity (blue) modify the model's predictive probability of baseline microscopy positivity by increasing this probability compared to absence of cavity (red).

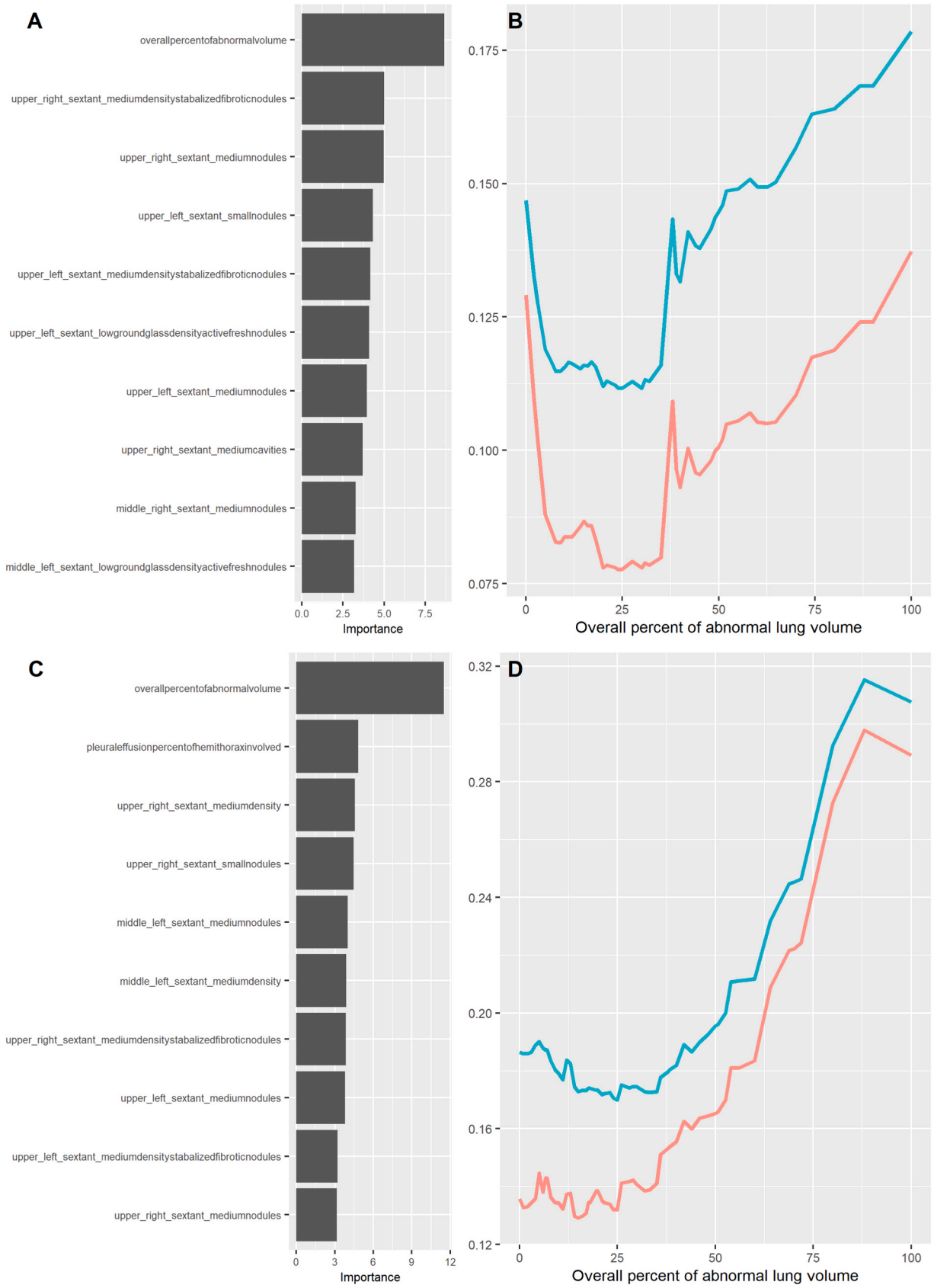

Fig. 3. Important radiological predictors from the Random Forest model to predict treatment success in matched Sensitive TB and MDR TB patients. The most important features from the random forest workflow trained on the matched population of Sensitive and MDR TB patients is shown for A) Sensitive TB Patients and C) MDR TB Patients. Partial Dependence Plot showing the predictive relationship between overall abnormal volume of lung and cavity status is shown for the same random forest workflow for B) Sensitive TB Patients and D) MDR TB Patients. Presence of cavity (blue) modify the model's predictive probability of baseline microscopy positivity by increasing this probability compared to absence of cavity (red).

4. Discussion

This study is unique in that it assesses detailed, baseline radiological findings' and Timika Score's predictive performance across two distinct clinical outcomes at the beginning and end of DS and MDR TB clinical care. To the best of our knowledge, this is the first study to use machine learning to compare the predictive performance of Timika Score with detailed radiologist findings within DS and MDR TB patients from a large, trans-national, multi-site cohort containing a wide variety of routinely collected radiological features and distinct clinical outcomes over time. Previous studies examined the performance of the Timika Score in one clinical site or prediction of a single clinical event [8,21,10,9,22–24,7]. Moreover, this study utilized IML approaches to explain how the models arrive at each prediction. Estimated overall abnormal percent of lung volume was the most important predictor identified across clinical outcomes for both DS and MDR whereas presence of cavity was only identified during baseline positivity prediction. Presence of nodules in the upper and middle sextants of the lungs were also consistently among the top predictors for each outcome. These observations are consistent with a previous study analyzing CT images that showed association of disease extent and presence of cavity with sputum smear grade [25]. Nevertheless, other features were unique to prediction of a specific outcome. For example, percent of pleural effusion of hemithorax was only identified in the top predictors for treatment outcome in MDR. Interestingly, pleural effusion has been suggested to be an indicator of TB reactivation in adults [26] with potential of reactivation in previously healed TB cases [27] and its remission has been associated with treatment response by CT data [28]. The IML results demonstrate why Timika Score performs equivalently or even better than detailed radiologist observations for baseline positivity but not for treatment outcome. The additional radiologist findings improved prediction with high likelihood.

To address the most frequent concerns related to real world studies of tuberculosis such as biases arising from cohort selection or missing treatment outcome, we performed a variety of additional, careful tests. Certain patient characteristics such as HIV or age may influence treatment outcome [29,30] or baseline positivity [31]. Furthermore, measuring predictive performance across multiple clinical outcomes necessitating co-occurring information (e.g., baseline radiology and sputum microscopy) during specific periods of time within patient clinical care may result in selection bias. To compare model performance we include an analysis on matched DS and MDR patients on several covariates that had significant differences such as age, reported HIV, smoking, drug use, etc. We demonstrate that these covariates did not significantly impact the generalization of the findings in the matched analysis. The matched analysis showed consistent results with the unmatched analysis. Our IIT analysis also demonstrates that censoring or drop-out due to exclusion of certain treatment outcomes does not impact our findings. By verifying that the results are consistent across these additional cross-checks, we provide stronger evidence in support of the main findings.

Previous machine learning studies may not consider how biases in study population or clinical events (such as patient drop-out prior to treatment completion) might impact model performance. Moreover, such studies may not consider model performance in a way to assess the clinical relevance. We employed state of the art Bayesian methods to compare model performance from the best model workflows [19]. If machine learning studies compare only CV results, comparisons can results in statistically significant differences that are not clinically meaningful (e.g. a negligible difference in AUC). The Bayesian ANOVA approach assesses model predictive performance in a probabilistic manner to overcome these limitations [32]. For example, we clearly demonstrate with a defined ROPE the likelihood of improvements of models incorporating standardized, detailed radiologist findings compared to Timika Score. These comparisons show where improvements with clinical impact are likely versus unlikely to be achieved.

In our study, Timika Score performed equivalently or even exceeded using all other radiological features for prediction of baseline sputum microscopy. This is a useful finding for radiologists as it is an easy to implement score and valuable clinical tool. Our results for prediction of sputum microscopy are consistent with prior findings showing the utility of Timika Score for baseline disease severity and infectiousness [9,7,8]. With regards to treatment outcome, our results fall into the range of previously reported ROC metrics showing a modest predictive ability for Timika Score between 0.6 and 0.7 ROC [11,21]. We show that Timika Score performs equivalently in DS patients for baseline sputum positivity and treatment outcome but not in MDR patients where it has less reliability to predict treatment outcome. The difference we observed in performance for early and late clinical outcomes may be due to the differing treatment lengths of DS and MDR patients[33,34]. The extended time associated with MDR treatment may diminish the capability of baseline radiology to reliably predict treatment outcomes. The comparisons showing that use of detailed radiologist findings outperforms Timika Score suggest opportunity for baseline, radiology-informed clinical score development for treatment outcome especially in DS patients with shorter treatments. While CXR serves as a useful tool for TB diagnosis and monitoring, they are two dimensional making some of the clinical parameter estimates difficult. As CT technology becomes more accessible, future research could examine Timika Score and other radiology-informed scores in three dimensional CT images permitting greater confidence in the estimates of salient radiological findings.

5. Conclusion

Timika Score performed equivalently to using all other radiological findings for prediction of baseline sputum microscopy positivity in both sensitive and MDR patients; however, prediction of treatment outcome using Timika Score was less reliable compared to using all other radiologist findings especially in sensitive patients. We also demonstrate that Timika Score predictive performance for baseline microscopy sputum positivity exceeds that of its performance for treatment outcome in MDR patients but not sensitive patients. These findings may aid in the appropriate usage of Timika Score for the identification of high-risk patients at the beginning and end of TB clinical care, research prioritization on automated features for deep learning, and opportunities for the development of novel radiology-informed clinical scores. Timika Score makes it easier to use baseline chest X-rays for predicting early infectiousness of the case and exhibits performance that is comparable to that of using more comprehensive, standardized information from the radiologist. Although the score's usefulness for predicting treatment outcomes declines, it is still surpassed by the use of the detailed radiologist observations, particularly in cases of MDR TB.

Ethics statement

Only publicly-available, de-identified data was used. Each participating clinical research institution (https://tbportals.niaid.nih.gov/where-do-our-cases-come-from) receives approval from the participating institution's IRB for public-sharing and follows strict adherence to ethics rules requirements of CRDF Global and the International Science and Technology Center who are the grant-issuing institutions.

Funding statement

This research was supported in part by the Office of Science Management and Operations of NIAID at the NIH.

CRediT authorship contribution statement

Hurt Darrell: Writing – review & editing, Supervision, Resources, Project administration. **Rosenthal Alex:** Writing – review & editing,

Supervision, Resources, Project administration. **Rosenfeld Gabriel:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Gabrielian Andrei:** Writing – review & editing, Writing – original draft, Validation, Supervision, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank Michael Tartakovsky and the program leadership of TB Portals for all their support and contributions.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ejro.2023.100518.

References

[1] C. Jeremiah, E. Petersen, R. Nantanda, B.N. Mungai, G.B. Migliori, F. Amanullah, P. Lungu, F. Ntoumi, N. Kumarasamy, M. Maeurer, A. Zumla, The WHO global tuberculosis 2021 report not so good news and turning the tide back to End TB, Int. J. Infect. Dis. (2022), https://doi.org/10.1016/j.ijid.2022.03.011.

[2] M. Pai, T. Kasaeva, S. Swaminathan, Covid-19's devastating effect on tuberculosis care A path to recovery, N. Engl. J. Med. 386 (2022) 1490–1493, https://doi.org/10.1056/nejmp2118145.

[3] E.A. Nardell, Transmission and institutional infection control of tuberculosis, Cold Spring Harb. Perspect. Med. 6 (2015), a018192, https://doi.org/10.1101/cshperspect.a018192.

[4] M. Nguyen, N. Levy, S. Ahuja, L. Trieu, D. Proops, J. Achkar, Factors associated with sputum culture-negative vs culture-positive diagnosis of pulmonary tuberculosis, JAMA Netw. (2019), https://doi.org/10.1001/jamanetworkopen.2018.7617.

[5] A. Harries, A. Kumar, Challenges and progress with diagnosing pulmonary tuberculosis in low- and middle-income countries, Diagnostics 8 (2018) 78, https://doi.org/10.3390/diagnostics8040078.

[6] N.F. Sabur, A. Esmail, M.S. Brar, K. Dheda, Diagnosing tuberculosis in hospitalized HIV-infected individuals who cannot produce sputum: is urine lipoarabinomannan testing the answer? BMC Infect. Dis. 17 (2017) https://doi.org/10.1186/s12879-017-2914-7.

[7] G. Rosenfeld, A. Gabrielian, A. Meyer, A. Rosenthal, Radiologist observations of chest x-rays (CXR) predict sputum smear microscopy status in TB portals, a real-world database of tuberculosis (TB) cases, (2022). https://doi.org/10.1101/2022.04.21.22273975.

[8] A.P. Ralph, M. Ardian, A. Wiguna, G.P. Maguire, N.G. Becker, G. Drogumuller, M. J. Wilks, G. Waramori, E. Tjitra, Sandjaja, E. Kenagalem, G.J. Pontororing, N. M. Anstey, P.M. Kelly, A simple, valid, numerical score for grading chest x-ray severity in adult smear-positive pulmonary tuberculosis, Thorax 65 (2010) 863–869, https://doi.org/10.1136/thx.2010.136242.

[9] A. Chakraborthy, A.J. Shivananjaiah, S. Ramaswamy, F. van der Kuyp, N. Chikkavenkatappa, Chest X ray score (Timika score): an useful adjunct to predict treatment outcome in tuberculosis, Adv. Respir. Med. 86 (2018), https://doi.org/10.5603/ARM.2018.0032.

[10] B.A. Thiel, C.M. Bark, J.G. Nakibali, F. van der Kuyp, J.L. Johnson, Reader variability and validation of the Timika X-ray score during treatment of pulmonary tuberculosis, Int. J. Tuberc. Lung Dis. 20 (2016) 1358–1363, https://doi.org/10.5588/ijtld.16.0186.

[11] Y. Krishnamoorthy, S. Knudsen, S. Sarkar, et al., Accuracy of Timika X-ray scoring system to predict the treatment outcomes among tuberculosis patients in India, Indian J. Tube (2022), https://doi.org/10.1016/j.ijtb.2021.08.004.

[12] A. Rosenthal, A. Gabrielian, E. Engle, D.E. Hurt, S. Alexandru, V. Crudu, E. Sergueev, V. Kirichenko, V. Lapitskii, E. Snezhko, V. Kovalev, A. Astrovko, A. Skrahina, J. Taaffe, M. Harris, A. Long, K. Wollenberg, I. Akhundova, S. Ismayilova, A. Skrahin, E. Mammadbayov, H. Gadirova, R. Abuzarov, M. Seyfaddinova, Z. Avaliani, I. Strambu, D. Zaharia, A. Muntean, E. Ghita, M. Bogdan, R. Mindru, V. Spinu, A. Sora, C. Ene, S. Vashakidze, N. Shubladze, U. Nanava, A. Tuzikov, M. Tartakovsky, The TB portals: an open-access, web-based platform for global drug-resistant-tuberculosis data sharing and analysis, J. Clin. Microbiol. 55 (2017) 3267–3282, https://doi.org/10.1128/jcm.01013-17.

[13] G. Rosenfeld, A. Gabrielian, Q. Wang, J. Gu, D.E. Hurt, A. Long, A. Rosenthal, Radiologist observations of computed tomography (CT) images predict treatment outcome in TB Portals, a real-world database of tuberculosis (TB) cases, PLOS ONE 16 (2021), e0247906, https://doi.org/10.1371/journal.pone.0247906.

[14] A. Gabrielian, E. Engle, M. Harris, K. Wollenberg, O. Juarez-Espinosa, A. Glogowski, A. Long, L. Patti, D.E. Hurt, A. Rosenthal, M. Tartakovsky, TB DEPOT (Data Exploration Portal): a multi-domain tuberculosis data analysis resource, PLOS ONE 14 (2019), e0217410, https://doi.org/10.1371/journal.pone.0217410.

[15] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G.M. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement, g7594–g7594, BMJ 350 (2015), https://doi.org/10.1136/bmj.g7594.

[16] D.E. Ho, K. Imai, G. King, E.A. Stuart, MatchIt: Nonparametric preprocessing for parametric causal inference, 42 (2011). https://doi.org/10.18637/jss.v042.i08.

[17] M. Kuhn, H. Wickham, Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles., (2020). ⟨https://www.tidymodels.org⟩.

[18] W.M. Landau, The targets r package: A dynamic make-like function-oriented pipeline toolkit for reproducibility and high-performance computing, 6 , 2021 2959. https://doi.org/10.21105/joss.02959.

[19] M. Kuhn, Tidyposterior: Bayesian analysis to compare models using resampling statistics, 2022. ⟨https://CRAN.R-project.org/package=tidyposterior⟩.

[20] A. Asres, D. Jerene, W. Deressa, Delays to treatment initiation is associated with tuberculosis treatment outcomes among patients on directly observed treatment short course in Southwest Ethiopia: a follow-up study, BMC Pulm. Med (2018), https://doi.org/10.1186/s12890-018-0628-2.

[21] M. Kriel, J.W. Lotz, M. Kidd, G. Walzl, Evaluation of a radiological severity score to predict treatment outcome in adults with pulmonary tuberculosis, Int. J. Tuberc. Lung Dis. 19 (2015) 1354–1360, https://doi.org/10.5588/ijtld.15.0098.

[22] C. Riou, D.B. E, S. Ruzive, R.T. Goliath, C.S. Lindestam Arlehamn, Disease extent and anti-tubercular treatment response correlates with Mycobacterium tuberculosis-specific CD4 T-cell phenotype regardless of HIV-1 status, Clin. Transl. Immunol. (2020), https://doi.org/10.1002/cti2.1176.

[23] F. Di Gennaro, G. G, F. Palmieri, et al., Increase in tuberculosis diagnostic delay during first wave of the COVID-19 Pandemic: Data from an Italian infectious disease referral hospital, Antibiotics (2021), https://doi.org/10.3390/antibiotics10030272.

[24] E. Du Bruyn, S. Ruzive, C.S. Lindestam Arlehamn, A. Sette, A. Sher, D.L. Barber, R. J. Wilkinson, C. Riou, Mycobacterium tuberculosis-specific CD4 T cells expressing CD153 inversely associate with bacterial load and disease severity in human tuberculosis, Mucosal Immunol. 14 (2020) 491–499, https://doi.org/10.1038/s41385-020-0322-6.

[25] J. Ko, H. Park, H. Kim, S. S, The relation between CT findings and sputum microbiology studies in active pulmonary tuberculosis, Eur. J. Radio. (2015), https://doi.org/10.1016/j.ejrad.2015.07.032.

[26] H. Kim, H. Lee, J. Yim, et al., The prevalence of pulmonary parenchymal tuberculosis in patients with tuberculous pleuritis, Chest (2006), https://doi.org/10.1378/chest.129.5.1253.

[27] C. Horsburgh, J. Goldberg, T.E.S. Consortium, et al., Latent TB infection treatment acceptance and completion in the United States and Canada, Chest (2010), https://doi.org/10.1016/10.1378/chest.09-0394.

[28] J. Lee, P. Chong, C. L, et al., High resolution chest CT in patients with pulmonary tuberculosis: characteristic findings before and after antituberculous therapy, Eur. J. Radio. (2008), https://doi.org/10.1016/j.ejrad.2007.07.009.

[29] M. Alemu, A. Yesuf, A. Tesfahun, et al., Impact of HIV-AIDS on tuberculosis treatment outcome in Southern Ethiopia - a retrospective cohort study, J. Clin. Tube Other Mycobact. Dis. (2021), https://doi.org/10.1016/j.jctube.2021.100279.

[30] M. Araia, F. Kibreab, A. Mesfin, et al., Determinants of unsuccessful tuberculosis treatment outcome in Northern Red Sea region, Eritrea, PLOS One (2022), https://doi.org/10.1371/journal.pone.0273069.

[31] P. Quinco, S. Buhrer-Skeula, C.-S. Marcelo, et al., Increased sensitivity in diagnosis of tuberculosis in hiv-positive patients through the small-membrane-filter method of microscopy, J. Clin. Microbiol (2013), https://doi.org/10.1128/JCM.00683-13.

[32] A. Benavoli, G. Corani, J. Demšar, M. Zaffalon, Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis, J. Mach. Learn. Res. 18 (2017) 1–36. ⟨http://jmlr.org/papers/v18/16-305.html⟩.

[33] C.A. Winston, K. Mitruka, Treatment duration for patients with drug-resistant tuberculosis, united states, Emerg. Infect. Dis. 18 (2012), https://doi.org/10.3201/eid1807.120261.

[34] J.G. Jang, J.H. Chung, Diagnosis and treatment of multidrug-resistant tuberculosis, Yeungnam Univ. J. Med. 37 (2020) 277–285, https://doi.org/10.12701/yujm.2020.00626.