Evaluation of chest X-ray with automated interpretation algorithms for mass tuberculosis screening in prisons: a cross-sectional study

Thiego Ramon Soares,[a] Roberto Dias de Oliveira,[a,b] Yiran E. Liu,[c] Andrea da Silva Santos,[a] Paulo Cesar Pereira dos Santos,[a] Luma Ravena Soares Monte,[b] Lissandra Maia de Oliveira,[d] Chang Min Park,[e,f] Eui Jin Hwang,[e,f] Jason R. Andrews,[c,i] and Julio Croda[d,g,h,i,*]

[a]Faculty of Health Sciences of Federal University of Grande Dourados, Dourados, MS, Brazil
[b]Nursing School, State University of Mato Grosso do Sul, Dourados, MS, Brazil
[c]Division of Infectious Diseases and Geographic Medicine, Stanford University School of Medicine, Stanford, CA, United States of America
[d]Oswaldo Cruz Foundation, Campo Grande, MS, Brazil
[e]Department of Radiology, Seoul National University College of Medicine, Seoul, Korea
[f]Department of Radiology, Seoul National University Hospital, Seoul, Korea
[g]Department of Epidemiology of Microbial Diseases, Yale University School of Public Health, New Haven, CT, United States of America
[h]School of Medicine, Federal University of Mato Grosso do Sul, Campo Grande, MS, Brazil

The Lancet Regional Health - Americas 2023;17: 100388

Published Online 4 November 2022
https://doi.org/10.1016/j.lana.2022.100388

Summary

Background The World Health Organization (WHO) recommends systematic tuberculosis (TB) screening in prisons. Evidence is lacking for accurate and scalable screening approaches in this setting. We aimed to assess the accuracy of artificial intelligence-based chest x-ray interpretation algorithms for TB screening in prisons.

Methods We performed prospective TB screening in three male prisons in Brazil from October 2017 to December 2019. We administered a standardized questionnaire, performed a chest x-ray in a mobile unit, and collected sputum for confirmatory testing using Xpert MTB/RIF and culture. We evaluated x-ray images using three algorithms (CAD4TB version 6, Lunit version 3.1.0.0 and qXR version 3) and compared their accuracy. We utilized multivariable logistic regression to assess the effect of demographic and clinical characteristics on algorithm accuracy. Finally, we investigated the relationship between abnormality scores and Xpert semi-quantitative results.

Findings Among 2075 incarcerated individuals, 259 (12.5%) had confirmed TB. All three algorithms performed similarly overall with area under the receiver operating characteristic curve (AUC) of 0.88–0.91. At 90% sensitivity, only LunitTB and qXR met the WHO Target Product Profile requirements for a triage test, with specificity of 84% and 74%, respectively. All algorithms had variable performance by age, prior TB, smoking, and presence of TB symptoms. LunitTB was the most robust to this heterogeneity but nonetheless failed to meet the TPP for individuals with previous TB. Abnormality scores of all three algorithms were significantly correlated with sputum bacillary load.

Interpretation Automated x-ray interpretation algorithms can be an effective triage tool for TB screening in prisons. However, their specificity is insufficient in individuals with previous TB.

Funding This study was supported by the US National Institutes of Health (grant numbers R01 AI130058 and R01 AI149620) and the State Secretary of Health of Mato Grosso do Sul.

Keywords: Automated interpretation; Diagnostics; Prisons; Tuberculosis; X-ray

*Corresponding author. Oswaldo Cruz Foundation - Mato Grosso do Sul, Campo Grande, MS 79074-460, Brazil.
E-mail address: julio.croda@fiocruz.br (J. Croda).
[i]Authors contributed equally to the work.

Research in context

Evidence before this study

The World Health Organization (WHO) recommends systematic tuberculosis screening in prisons. We reviewed the evidence for using chest X-rays with automated interpretation for tuberculosis screening in prisons by searching the medline for articles published in English or Portuguese as of July 31, 2022, using the terms ("prison*" or "jail" or "correctional" or "detention") and ("computer-aided detection" or "artificial intelligence" or "automated interpretation" or "machine learning") and ("radiograph" or "X-ray"). Three studies evaluated the use of CAD4TB among incarcerated populations, finding that it was useful as a screening tool to guide further TB testing, though the largest of these studies included only 33 cases. None of the studies evaluated other automated interpretation algorithms or investigated demographic and clinical factors associated with algorithm performance.

Added value of this study

In a study conducted in three high tuberculosis-burden prisons in Brazil, we evaluated three automated chest x-ray interpretation algorithms (CAD4TBv6, LunitTB version 3.1.0.0 and qXR v3) as an initial screening tool for tuberculosis. Among 259 microbiologically-confirmed tuberculosis cases and 1816 tuberculosis-negative controls, we found that two of three systems met WHO target profile product benchmarks (90% sensitivity, 70% specificity). Accuracy was not substantially impacted by the presence or absence of symptoms, smoking, or drug use. The specificity of all interpretation algorithms declined modestly with age but was markedly diminished (<50%) for all three algorithms among individuals with prior tuberculosis. Among cases with medium or high sputum bacillary load by Xpert MTB/RIF G4, the sensitivity of all automated interpretation algorithms exceeded 95%.

Implications of all the available evidence

Globally, prisons are high-risk settings for tuberculosis, yet case detection remains poor, amid limited investments in active case finding. Tuberculosis screening by chest x-rays with automated interpretation algorithms, followed by confirmatory testing using molecular diagnostics, may be an efficient means to improve case detection among incarcerated populations. Our findings indicate that currently available automated interpretation algorithms perform with sufficient accuracy among incarcerated persons, particularly in identifying individuals with high sputum bacterial load, and are robust to individual clinical and demographic characteristics including symptoms, smoking and drug use. For individuals with prior tuberculosis, automated interpretation of chest x-rays falls below WHO accuracy benchmarks, and alternative means of tuberculosis screening are needed.

Introduction

Globally, tuberculosis (TB) incidence in prisons is more than ten times higher than the general population.[1] This disparity is especially alarming in South America, where TB cases in prisons have more than doubled since 2000 amid rising incarceration rates.[1,2] Several factors contribute to the elevated risk of TB in prisons, including overcrowding, poor ventilation, high rates of smoking, drug use, and limited access to medical care, leading to delays in TB diagnosis.[2]

Interventions to address this growing burden are urgently needed, including improvements in case detection. In 2021, the World Health Organization (WHO) released updated guidelines on screening for TB, upgrading to a strong recommendation that systematic screening be conducted in prisons and penitentiary institutions.[3] However, the recommendation is based on "very low certainty of evidence", and guidance on specific means for screening in this setting is lacking. Moreover, correctional health systems are often underfunded and poorly equipped, and few prison systems in low- and middle-income countries perform systematic screening for TB despite the widely acknowledged high burden. Therefore, effective, cost-efficient screening approaches are needed to bring case-finding to scale in these settings. An important part of such approaches is a point-of-care screening test that can substantially reduce the number of people who need further testing.

Chest radiography is among the oldest tools for pulmonary TB screening and historically played a major role in TB control programs in high burden settings.[4,5] However, by the 1970s, concerns were raised about the accuracy, logistics and personnel requirements for mass radiography, leading the WHO to conclude in its 9th expert committee report that "indiscriminate TB case finding by mobile mass radiography should be abandoned".[6]

Recently, there has been a resurgence in interest in the use of radiography as a screening tool for TB, leveraging recent advances in machine learning approaches to automate x-ray interpretation.[7,8] Clinic-based evaluations have demonstrated promising accuracy for several automated interpretation systems among individuals with TB symptoms, and the comparison between human and automated interpretations show comparable results.[9–12] As a result, the WHO guidelines provided a new, conditional recommendation that computer-aided detection may be used in place of human readers for screening and triage for tuberculosis.

However, there is a need to understand how well these algorithms will perform for the purpose of active case finding, irrespective of symptoms, in incarcerated populations with high prevalence of smoking, drug use, and history of TB. To address this knowledge gap, we evaluated the performance of three deep learning-based x-ray interpretation algorithms in the context of mass screening for TB in three high burden prisons in Brazil.

Methods

Study design

We performed a cross-sectional study, embedded within a larger prospective mass tuberculosis screening study, from October 2017 to December 2019 in three male prisons in Mato Grosso do Sul State, Brazil: Jair Ferreira de Carvalho Penitentiary (EPJFC), Campo Grande Penal Institute (IPCG), and Dourados State Penitentiary (PED). The prisons have a combined population of approximately 5500 individuals. All incarcerated individuals in each prison were invited to participate in TB screening. Those who agreed to participate in the study provided written informed consent. The study was approved by the institutional review boards (IRBs) of the Federal University of Grande Dourados (UFGD) (#3.483.377) and Stanford University (#40285).

Study procedures

We outfitted a Volkswagen Constellation 24–240 truck with a 9.8 × 2.5-m container, lead covering, an access ramp, an x-ray machine (Altus ST 543 HF, Sawae®), an x-ray scanner and digitizer (Agfa 15-X CR, Mortsel Belgium) and a separate room for sputum processing with two 4-module GeneXpert machines (Cepheid, Sunnyvale, USA). The mobile screening team consisted of a nurse, a laboratory technician, and an x-ray technician, with a physician available for consultation.

Study nurses administered a structured questionnaire to obtain demographic data, incarceration history, lifestyle factors, health history, and TB symptoms (cough, fever, night sweats, weight loss, loss of appetite, tiredness, and chest pain).[13] A spot sputum sample was collected from all participants who were able to produce sputum, with a target volume of at least 2 ml. After homogenizing, 1 ml of sputa was tested by Xpert MTB/RIF G4 (Cepheid, Sunnyvale, USA) and, if an additional 1 ml of sputum remained, it was used for culture on Ogawa-Kudoh media. *Mycobacterium tuberculosis* growth in cultures was confirmed by an immunochromatographic assay (TB Ag MPT64 Rapid Test, Standard Diagnostics, Seoul, South Korea).

A posterior-anterior chest x-ray was performed for all participants using a Sawae analog system and then scanned and digitized. To simulate a real-world screening program, all chest x-rays were included, irrespective of quality. The images (blinded, and without any metadata) were electronically transferred for automated analysis by the developers of Computer-Aided Detection for TB version 6 (CAD4TBv6) at Radboud University Medical Center (Netherlands); Lunit INSIGHT CXR2 version 3.1.0.0 (hereinafter LunitTB) developed by the South Korean medical software company Lunit; and qXR version 3, developed by Qure.ai in Mumbai, India. All information was recorded in Research Electronic Data Capture (REDCap®).[14,15]

Outcome definitions and analytic approach

We defined TB cases as individuals with a positive Xpert MTB/RIF or culture growing *M. tuberculosis*. We defined controls as individuals who had sputum testing for Xpert MTB/RIF and no positive result by Xpert or culture. As a secondary analysis, we included as controls all individuals who were screened for TB and did not have a positive test, regardless of whether they were able to provide sputum. Individuals already undergoing treatment for TB were excluded from all analyses.

For our primary analyses, we evaluated the performance of each algorithm with the WHO's Target Product Profile[16] (TPP) for a triage test by identifying the threshold that achieved 90% sensitivity and examining the corresponding specificity. As secondary analyses, we used pre-selected thresholds. For CAD4TBv6, which provides a score range of 0–100, we used a positivity threshold score of ≥60 through calibration with radiographic imaging data from a subset of participants with (n = 80) and without (n = 200) microbiologically confirmed TB. For LunitTB, which provides a score range of 0–1, we used a threshold of ≥0.72 as specified by the manufacturer and identified through prior calibration.[10,11] For qXR (score range 0–1), we used a threshold of ≥0.5 according to a previous study.[17]

For each algorithm we calculated the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the receiver operating characteristic (ROC) curve (AUC). We report PPV and NPV at a prevalence of 4%, which is the prevalence observed across three rounds of mass screening in the study prisons.[18] We calculated exact binomial confidence intervals (CIs) for sensitivity and specificity. We compared algorithm AUCs using DeLong's test. For demographic and clinical characteristics, continuous variables were compared using the Mann–Whitney U test and categorical variables using the chi-square test. To assess the influence of demographic and clinical characteristics on algorithm performance, we conducted multivariable logistic regression controlling for age, race, drug use, smoking, previous TB, and presence of any TB symptoms (cough, fever, night sweat, weight loss, loss of appetite, tiredness, and chest pain). We plotted binned residuals, evaluated Cook's distance, and tested for multiple collinearity by calculating variance inflation factors for

each variable. We report predicted marginal estimates of specificity for each characteristic at the WHO TPP threshold of 90% sensitivity. Finally, we investigated the relationship between Xpert semi-quantitative result and x-ray algorithm score among confirmed TB cases using Kendall's tau. The sample size for the study (anticipated number of confirmed cases, 158) was based on achieving precision of ±6% around the anticipated sensitivity (85%) of the primary study diagnostic (sputum pooling), which has been previously reported.[19] For the sub-study, the actual case numbers were found to be sufficient for characterizing x-ray sensitivity with equal or greater at the WHO TPP benchmark of 90%. Data were analyzed using SPSS version 25.0 and R version 4.0.3, including the pROC package (version 1.17.0.1).[20]

Role of the funding source

The funders did not have any role in study design, data collection, data analysis, interpretation, or writing of the report.

Results

Between October 2017 and December 2019, we enrolled 7081 participants across three male prisons in the Brazilian state of Mato Grosso do Sul. Sixty-six participants were excluded from further analyses as they were already under treatment for tuberculosis. Among the remainder, 2075 (29.3%) participants were able to produce valid sputum samples for Xpert and were included in the primary analysis (Fig. 1). Among these, 1084 (52.2%) additionally had sputum cultures performed.

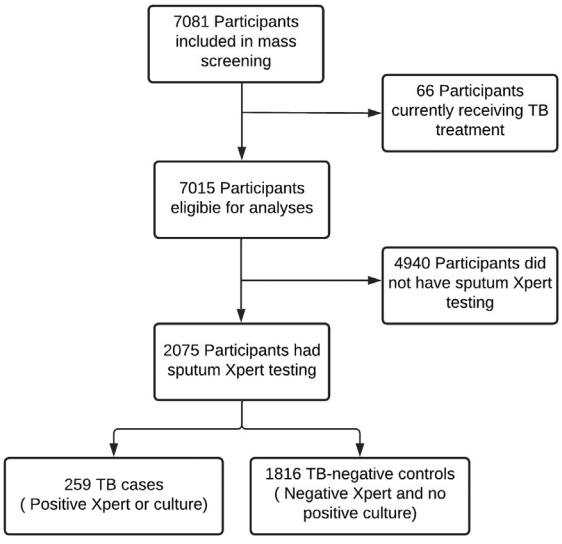

Fig. 1: Flow diagram of study participants in mass screening and inclusion of participants in x-ray evaluation.

Participants in the primary analysis had a median age of 33 years (IQR 28–40) (Table 1). Compared with participants who did not produce a valid sputum sample, those who did had a higher prevalence of TB symptoms (73% vs 18%, $p < 0.001$), smoking (73% vs 55%, $p < 0.001$), illicit drug use (70% vs 54% $p < 0.001$), and previous tuberculosis (14% vs 5%, $p < 0.001$) (Supplementary Table S1).

During the screening period, 259 (12.5%) participants were diagnosed with pulmonary TB, of which 113 (43.6%) were diagnosed by sputum Xpert alone, 17 (6.6%) were diagnosed through sputum culture alone, and 129 (49.8%) had positive Xpert and culture tests. Sixty-two participants were Xpert positive and culture negative. The presence of any TB symptom did not differ between TB cases and controls (76% vs 73%, $p = 0.28$); however, cough was slightly more common among TB cases (66 vs 60%, $p = 0.04$). Smoking, drug use, history of incarceration, and history of TB were significantly more prevalent among TB cases compared to non-TB cases (Table 1).

Among TB cases, 209 (80.7%) had X-ray abnormality scores above the threshold selected by CAD4TBv6, 207 (79.9%) by, LunitTB, and 193 (74.5%) by qXR (Table 2). At 90% sensitivity, only LunitTB and qXR met the WHO's Target Product Profile (TPP) with specificity of 83.7% (95% CI 72.4–87.3) and 74.2% (95% CI 60.2–81.3), respectively. At a 4% prevalence of TB, LunitTB had the highest PPV (18.7%), followed by qXR (12.7%) and CAD4TBv6 (9.0%). Receiver operating characteristic (ROC) curves for each algorithm are shown in Fig. 2. Compared with CAD4TBv6 (AUC 0.88), LunitTB (AUC 0.91, $p = 0.003$) and qXR (AUC 0.90, $p = 0.01$) had higher AUCs, though AUC did not differ between LunitTB and qXR ($p = 0.17$). In a secondary analysis of accuracy in which we included the 4940 participants unable to provide sputum (total $N = 7015$), AUCs did not differ substantially from the primary analysis, with LunitTB at 0.93, qXR at 0.92, and CAD4TBv6 at 0.90 (Supplementary Figure S1).

We next performed multivariable logistic regression analysis to examine whether the performance of each algorithm varied by sociodemographic characteristics and risk factors, namely: age, race/ethnicity, current smoker, drug use, previous TB, and TB symptoms. Specificity of all three algorithms decreased with age and tended to be lower among current smokers and those without TB symptoms, compared to their respective counterparts (Fig. 3, Supplementary Figures S2 and S3). LunitTB was the only algorithm that met WHO TPP criteria among individuals 45 years and older. Notably, specificity was under 50% across all three algorithms for individuals with a history of TB.

To further investigate diagnostic performance depending on history of previous TB, we analyzed the distribution of abnormality scores for TyB cases versus non-TB cases as confirmed by sputum Xpert or culture,

Variables	Total N = 2075 (%)	TB cases N = 259 (%)	No TB N = 1816 (%)	p value
Median age (IQR)	33 (28, 40)	33 (28, 39)	33 (28, 40)	0.5
Prison Unit				<0.001
PED	889 (42.8)	65 (25.1)	824 (45.0)	
EPJFC	840 (40.5)	144 (55.6)	696 (38.3)	
IPCG	346 (16.9)	50 (19.3)	296 (16.3)	
Race/ethnicity				0.2
Mixed	1279 (61.6)	158 (61.0)	1121 (61.4)	
White	508 (24.5)	56 (21.6)	452 (24.9)	
Black	253 (12.2)	38 (14.7)	215 (11.8)	
Indigenous	33 (1.6)	7 (2.7)	26 (1.4)	
Asian	2 (0.1)	–	2 (0.1)	
<8 years of schooling	1546 (74.5)	198 (76,4)	1348 (74.2)	0.23
Current smoker	1520 (73.3)	208 (80.3)	1312 (72.2)	0.006
Illicit drug use over the last year	1460 (70.4)	203 (78.4)	1257 (69.2)	0.003
Previously incarcerated	1557 (75.0)	214 (82.6)	1343 (74.0)	0.003
BCG vaccinated	1862 (89.7)	223 (86.1)	1639 (90.3)	0.04
Previous TB	293 (14.1)	55 (21.2)	238 (13.1)	<0.001
Report any WHO TB symptoms	1512 (72.9)	196 (75.7)	1316 (72.5)	0.28
Report cough	1255 (60.5)	172 (66.4)	1083 (59.6)	0.04
TB contact	1565 (75.4)	211 (81.5)	1354 (74.6)	0.16

Table 1: **Sociodemographic characteristics and risk factors for TB among study participants, stratified by TB status as determined by sputum Xpert or culture.**

disaggregated by history of TB. We focused on LunitTB for this analysis given its superior overall performance and its relatively stable specificity by subgroup compared to the other two algorithms. Strikingly, the thresholds required to reach WHO TPP benchmarks of 90% sensitivity and 70% specificity varied dramatically by history of TB (Fig. 4). In participants without previous TB, LunitTB score thresholds ≥0.04 had 70% specificity and those ≤0.15 had 90% sensitivity, providing a range of thresholds (0.04–0.15) meeting TPP benchmarks. Conversely, in participants with previous TB, a threshold of at least 0.73 was required for 70% specificity, and there was no score threshold to satisfy both TPP sensitivity and specificity. In participants with previous TB, neither CAD4TBv6 nor qXR had a score that satisfies both TPP sensitivity and specificity (Supplementary Figures S4 and S5).

Finally, we assessed the relationship between sputum bacillary load and algorithm performance by examining x-ray abnormality scores by Xpert semi-quantitative result (negative, very low, low, medium, high). Among TB cases with a positive Xpert test for whom Xpert semi-quantitative results were available (188/242, 77.7%), all three algorithms yielded abnormality scores that were positively correlated with sputum Xpert semi-quantitative levels ($p < 0.0001$) (Fig. 5). Among the 67 participants with a medium or high Xpert result, CAD4TBv6 had 97% sensitivity (65/67) and LunitTB and qXR both had 96% sensitivity (64/67) at the 70% specificity threshold (Supplementary Table S2).

Discussion

Active case finding for tuberculosis in high burden carceral settings is needed to address the substantial excess burden among incarcerated populations. However, despite WHO recommendations for routine TB screening in prisons, most facilities in low- and middle-income countries do not perform systematic active case

System	AUC (95% CI)	At pre-defined thresholds		At 90% sensitivity, 4% prevalence		
		Sensitivity % (95% CI)	Specificity % (95% CI)	Specificity % (95% CI)	PPV %	NPV %
CAD4v6	0.88 (0.85–0.90)	80.7 (75.4–85.3)	82.7 (80.8–84.4)	62.3 (52.0–73.1)	9.0	99.3
LunitTB	0.91 (0.89–0.93)	79.9 (74.5–84.6)	89.8 (88.3–91.2)	83.7 (72.4–87.3)	18.7	99.5
qXR	0.90 (0.88–0.92)	74.5 (68.8–79.7)	89.4 (87.9–90.8)	74.2 (60.2–81.3)	12.7	99.4

Table 2: **Sensitivity, Specificity, Area Under the Curve (AUC), Positive Predictive Value (PPV) and Negative Predictive Value (NPV) of each algorithm at pre-defined thresholds or with thresholds adjusted to 90% sensitivity as specified by the WHO Target Product Profile minimum target.**

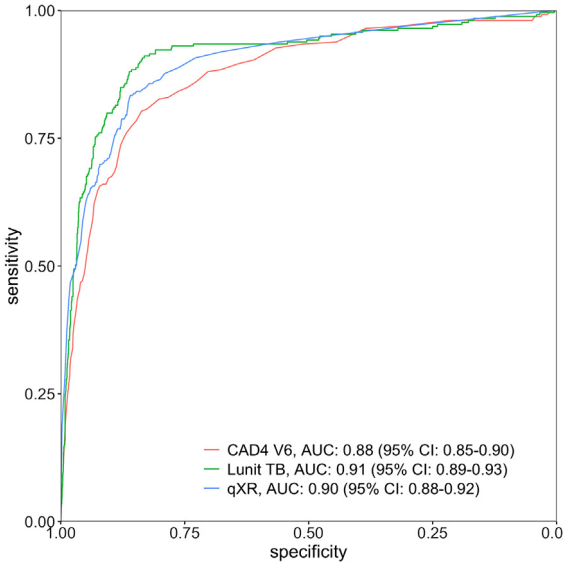

Fig. 2: Receiver operating characteristic (ROC) curves for CAD4v6, LunitTB and qXR.

finding, often citing resource and infrastructure constraints. Despite being an upper-middle-income country, Brazil is considerably unequal in income distribution, so it faces all these limitations of low-middle-income countries. Effective, cost-efficient screening strategies are needed to make active case finding more accessible in such environments. In this study, conducted via a nurse-led mobile diagnostic unit in three prisons in Central-Western Brazil, we found a very high prevalence of undiagnosed, microbiologically confirmed TB (3.7%). Algorithms for automated interpretation of x-rays achieved high sensitivity and specificity as a screening tool, with the LunitTB and qXR systems exceeding the WHO minimal TPP thresholds for a screening or triage test. Sputum molecular testing is still needed to confirm TB, but a limiting factor in the speed and costs of screening has been the number of tests that can be run daily during mass screening of thousands of individuals.[3,18] Our findings suggest that screening by mobile x-ray systems with automated interpretation could reduce the number of confirmatory tests required and enable screening to be more rapid in high burden TB settings, while still maintaining sufficient sensitivity.

Recent studies have evaluated x-ray interpretation algorithms among individuals presenting to clinics with TB symptoms, finding variable results. An individual participant meta-analysis found that none of the systems investigated met the WHO TPP criteria for triage, with specificities ranging from 54 to 61% at 90% sensitivity.[21] By contrast, a study in Bangladesh found that the qXR and CAD4TB systems achieved >70% specificity at the same threshold, and that all algorithms outperformed interpretation by radiologists.[22] Our study differed in that it was performed in the context of active case finding, irrespective of symptoms, which could affect estimates of accuracy in several ways. For instance, the cases identified through systematic screening are often those in early stages of disease, with lower bacillary burden, as evidenced by the fact that 54% of confirmed cases in our cohort had low, very low, or negative Xpert results. Given the association we observed between sputum bacillary load and abnormality scores, this could

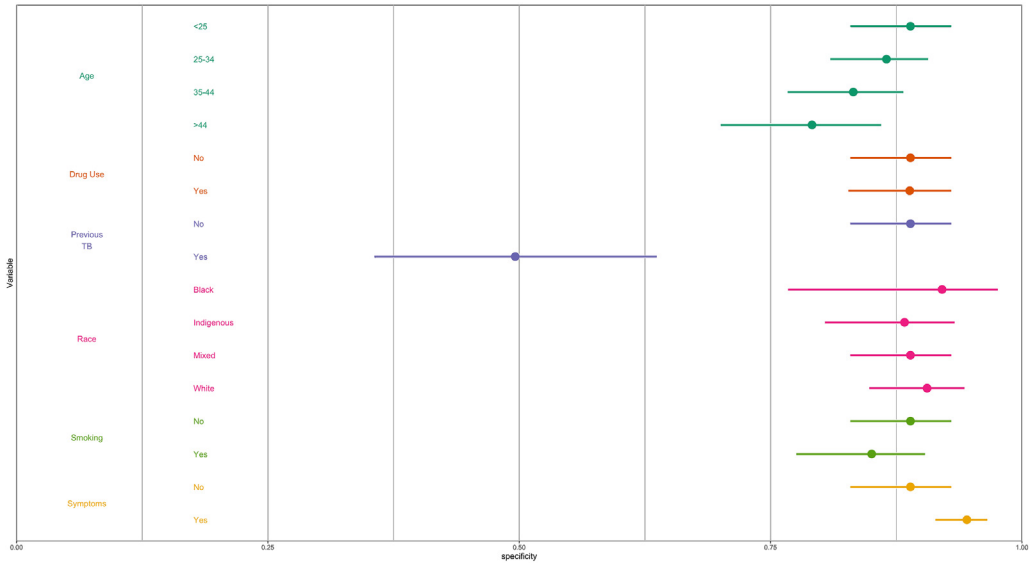

Fig. 3: Specificity of LunitTB, by sociodemographic characteristics and risk factors. Shown are the predicted margins for specificity and 95% confidence intervals from a multivariable logistic regression, holding sensitivity at 90%.

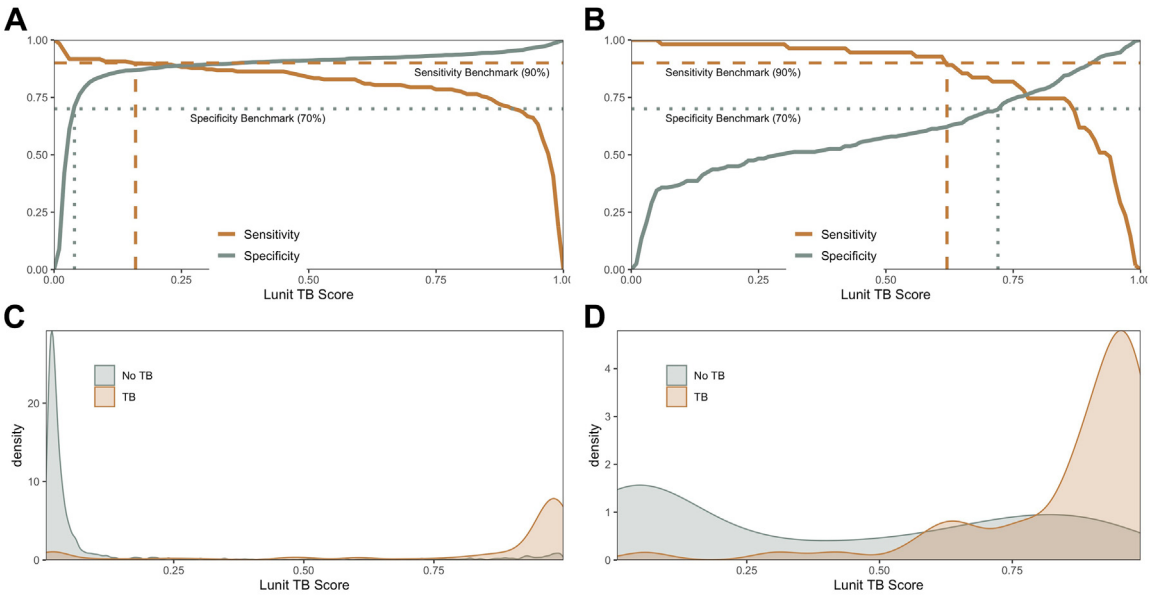

Fig. 4: Sensitivity and specificity according to LunitTB score threshold, with the WHO sensitivity (dashed line) and specificity (dotted line) benchmarks (top) among individuals without (A) and with (B) previous TB. Distribution of LunitTB scores for participants without (C) or with (D) previous TB (bottom).

have resulted in the algorithms having lower sensitivity in our cohort. At the same time, we might expect higher specificity in the context of active case finding, regardless of symptoms, compared to use in clinics among those presenting with TB symptoms, as the latter setting may include more patients with other pulmonary diseases such as bacterial and viral pneumonias that can be challenging to distinguish from TB. Furthermore, we evaluated these algorithms in incarcerated populations, which tend to be younger, predominantly male, and with high prevalence of various risk factors for TB.

LunitTB was the best-performing algorithm in this cohort, with greater accuracy and generalizability among subgroups, with particularly superior robustness to age compared to the other two algorithms. Nonetheless, performance of all three algorithms varied by subgroup, with consistently lower specificity among older individuals and those with previous TB, corroborating previous findings.[22,23] We also found reduced specificity among current smokers and those without TB symptoms. Of note, our pre-defined thresholds for each algorithm led to overall sensitivity under 90%, suggesting

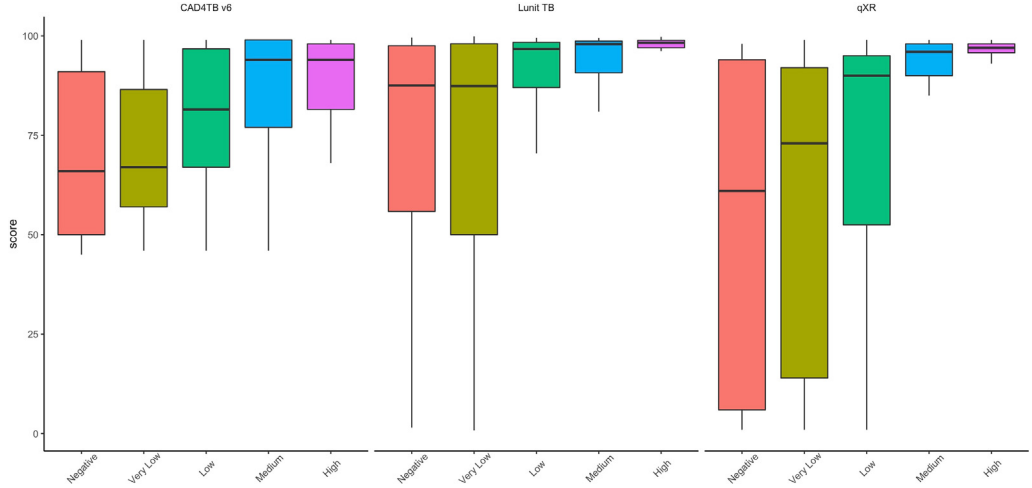

Fig. 5: Relationship between the bacillary load in the sputum and the performance of the algorithm through the stratification of the scores by the semiquantitative Xpert result.

that setting- or population-specific threshold calibration may be an important step in implementation.

Specificity of all three algorithms decreased considerably to less than 50% for those with previous TB, indicating failure to meet the WHO TPP for this subgroup. For the three analyzed software tools, the distribution of abnormality scores among non-TB cases was shifted higher for those with a history of TB, suggesting the algorithm may not distinguish active TB lesions from fibrous scarring of the lung parenchyma and other chest radiograph patterns indicative of previous TB.[24] Thus, in populations with high prevalence of previous TB, Xpert may be more appropriate for screening,[25] though studies have raised concern about its specificity in individuals with prior tuberculosis.[26]

We found that x-ray abnormality scores were higher —suggestive of more abnormalities—in individuals with high sputum bacillary loads. At the 70% specificity threshold, sensitivity for individuals with medium or high bacillary loads exceeded 96% for all three systems. Given that Xpert bacillary load correlates with smear status,[27] and smear status predicts infectiousness,[28,29] it may be reasonable to infer that x-ray automated interpretation algorithms may be more sensitive in identifying the most infectious individuals.

Even with the availability of automated interpretation algorithms, the cost-effectiveness of using x-rays for mass screening in prisons is still unclear. Previous work found that mass screening in prisons with sputum Xpert alone had high yield and was less costly than using x-ray and CAD4TB (version 5) for triage prior to confirmatory Xpert.[18] However, the prior study used a single CAD4TB threshold for all individuals and evaluated an additional strategy where only individuals without symptoms were screened with x-ray and CAD4TB prior to confirmatory Xpert. Our present findings suggest that such strategies may be less effective due to the algorithms' variable performance by subgroup, particularly the reduced accuracy for individuals without TB symptoms. Furthermore, CAD4TB (version 6) was shown to have the lowest performance in this study; thus, screening with a more accurate algorithm like LunitTB could increase cost-effectiveness. Based on the findings of this study, at a prevalence of 4%, LunitTB could detect 90% of cases while reducing the number of individuals requiring confirmatory Xpert testing by 80%. Additionally, emerging technologies for portable, digital radiography could reduce consumable costs, making x-rays more accessible and affordable in resource-constrained environments. Screening programs based on automated interpretation of x-rays still require x-ray equipment and protective equipment, radiographers and other personnel, electricity, and downstream diagnostics for tuberculosis confirmation; further studies are needed to quantify the costs of these components and evaluate the cost-effectiveness of mass screening programs utilizing x-ray automated interpretation.

This study has several limitations. First, in our primary analysis, we only included participants who were able to produce sputum for confirmatory testing, as sputum induction was not able to be undertaken in this environment. The excluded participants were less likely to be current smokers, to have TB symptoms, and to report previous TB; we expect that their inclusion may have affected overall estimates of algorithm performance in this population. In secondary analyses of the entire population, AUCs did not differ significantly. However, future research is needed to evaluate these x-ray interpretation algorithms on this group, given that a strength of x-ray screening is the lack of requirement for sputum. Second, we used solid media culture due to local availability costs; however, solid media culture is less sensitive than liquid media and could have led to missed cases, which could lead to overestimation of sensitivity and underestimation of specificity. We could not compare directly against smear microscopy, as we have replaced it with GeneXpert in our mass screening programs. Additionally, we note that while for LunitTB and qXR we used the manufacturers' recommended thresholds, for CAD4TBv6 we used a threshold determined from a subset of our population; therefore, the thresholds at 90% sensitivity may be more appropriate than our pre-defined thresholds for comparison of the three algorithms. A newer version of CAD4TB (version 7) has been released but was not available to us at the time of analysis. Due to the low prevalence of HIV in our population, we did not consider HIV status in our study.[30] Moreover, our study only included those in male prisons as there are fewer than 10 cases annually among incarcerated women in this state; consequently, the performance of these algorithms for TB screening in female prisons remains unknown.[23] Finally, we did not include the human interpretation of the radiographic images as a comparator, and we also did not evaluate the quality of the images before the CAD analysis.

Overall, our results suggest that the use of chest x-rays and artificial intelligence-based interpretation algorithms can be part of an effective mass screening strategy in high-burden settings like prisons. Although Lunit-TB had the greatest accuracy and robustness in our cohort, all three algorithms exhibited similar performance, particularly as a rule-out-test, and could be used to reduce the need for universal molecular testing. However, our findings suggest the need for future optimization of these algorithms to improve generalizability across subgroups, especially for individuals with a history of TB. Nevertheless, given their high overall accuracy in this population, especially among cases with the greatest sputum bacillary load, automated interpretation algorithms could enable scaling of mass screening to help mitigate disparities in TB diagnosis among incarcerated populations.

Contributors

T.R.S., R.D.d.O., A.d.S.S., P.C.P.d.S., L.R.S.M., and L.M.d.O.: acquisition, analysis, and interpretation of data for the work; drafting the work; final approval of the version to be published. C.M.P. and E.J.H.: acquisition and analysis of data for the work, final approval of the version to be published. Y.E.L.: drafting the work; revising it critically for important intellectual content; final approval of the version to be published. J.R.A.: conception and design of the work; performed data analyses, drafting the work; final approval of the version to be published. J.C.: conception and design of the work; drafting the work; final approval of the version to be published and takes responsibility for the content of the manuscript, including the data and analysis.

Data sharing statement

All data collected for the study are presented in the manuscript and its supplements.

Declaration of interests

The authors declare no conflict of interest.

Acknowledgments

The State Agency for the Administration of the Penitentiary System of the State of Mato Grosso do Sul (AGEPEN) for having authorized the study; the Coordination for the Improvement of Higher Education Personnel (CAPES) and the Federal University of Grande Dourados (UFGD) for having supported the study in their postgraduate programs.

Funding: This study was supported by the US National Institutes of Health (grant numbers R01 AI130058 and R01 AI149620) and the State Secretary of Health of Mato Grosso do Sul.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.lana.2022.100388.

References

1 Cords O, Martinez L, Warren JL, et al. Incidence and prevalence of tuberculosis in incarcerated populations: a systematic review and meta-analysis. *Lancet Public Health*. 2021;6:e300–e308.
2 Walter KS, Martinez L, Arakaki-Sanchez D, et al. The escalating tuberculosis crisis in central and South American prisons. *Lancet*. 2021;397:1591–1596.
3 World Health Organization. *WHO consolidated guidelines on tuberculosis. Module 2: screening – systematic screening for tuberculosis disease*. Geneva: World Health Organization; 2021.
4 Hermans SM, Andrews JR, Bekker L-G, Wood R. The mass miniature chest radiography programme in Cape Town, South Africa, 1948 - 1994: the impact of active tuberculosis case finding. *S Afr Med J*. 2016;106:1263–1269.
5 Comstock GW, Philip RN. Decline of the tuberculosis epidemic in Alaska. *Public Health Rep (1896)*. 1961;76:19–24.
6 Organization WH. *WHO Expert Committee on Tuberculosis [meeting held in Geneva from 11 to 20 December 1973]: ninth report*. 1974.
7 World Health Organization. Systematic screening for active tuberculosis: an operational guide. http://apps.who.int/iris/bitstream/10665/181164/1/9789241549172_eng.pdf?ua=1; 2015. Accessed March 5, 2021.
8 Pinto LM, Pai M, Dheda K, Schwartzman K, Menzies D, Steingart KR. Scoring systems using chest radiographic features for the diagnosis of pulmonary tuberculosis in adults: a systematic review. *Eur Respir J*. 2013;42:480–494.
9 Murphy K, Habib SS, Zaidi SMA, et al. Computer aided detection of tuberculosis on chest radiographs: an evaluation of the CAD4TB v6 system. *Sci Rep*. 2020;10:5492.
10 Jaeger S, Karargyris A, Candemir S, et al. Automatic tuberculosis screening using chest radiographs. *IEEE Trans Med Imaging*. 2014;33:233–245.
11 Velen K, Sathar F, Hoffmann CJ, et al. Digital chest X-ray with computer-aided detection for tuberculosis screening within correctional facilities. *Ann Am Thorac Soc*. 2022;19:1313–1319.
12 Mahler B, de Vries G, van Hest R, et al. Use of targeted mobile X-ray screening and computer-aided detection software to identify tuberculosis among high-risk groups in Romania: descriptive results of the E-DETECT TB active case-finding project. *BMJ Open*. 2021;11:e045289.
13 World Health Organization. *Global tuberculosis report 2021*. Geneva: World Health Organization; 2021.
14 Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. *J Biomed Inform*. 2019;95:103208.
15 Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42:377–381.
16 World Health Organization. *High-priority target product profi les for new tuberculosis diagnostics: report of a consensus meeting*. World Health Organization; 2014.
17 Nash M, Kadavigere R, Andrade J, et al. Deep learning, computer-aided radiography reading for tuberculosis: a diagnostic accuracy study from a tertiary hospital in India. *Sci Rep*. 2020;10:210.
18 Santos ADS, de Oliveira RD, Lemos EF, et al. Yield, efficiency and costs of mass screening algorithms for tuberculosis in Brazilian prisons. *Clin Infect Dis*. 2021;72(5):771–777.
19 dos Santos PCP, da Silva Santos A, de Oliveira RD, et al. Pooling sputum samples for efficient mass tuberculosis screening in prisons. *Clin Infect Dis*. 2022;74:2115–2121.
20 pROC: an open-source package for R and S+ to analyze and compare ROC curves. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-77. Accessed November 16, 2021.
21 Tavaziva G, Harris M, Abidi S, et al. Chest X-ray analysis with deep learning-based software as a triage test for pulmonary tuberculosis: an individual patient data meta-analysis of diagnostic accuracy. *Clin Infect Dis*. 2022;74(8):1390–1400.
22 Qin ZZ, Ahmed S, Sarker MS, et al. Tuberculosis detection from chest x-rays for triaging in a high tuberculosis-burden setting: an evaluation of five artificial intelligence algorithms. *Lancet Digit Health*. 2021;3:e543–e554.
23 Khan FA, Majidulla A, Tavaziva G, et al. Chest x-ray analysis with deep learning-based software as a triage test for pulmonary tuberculosis: a prospective study of diagnostic accuracy for culture-confirmed disease. *Lancet Digit Health*. 2020;2:e573–e581.
24 Piccazzo R, Paparo F, Garlaschi G. Diagnostic accuracy of chest radiography for the diagnosis of tuberculosis (TB) and its role in the detection of latent TB infection: a systematic review. *J Rheumatol Suppl*. 2014;91:32–40.
25 Frascella B, Richards AS, Sossen B, et al. Subclinical tuberculosis disease - a review and analysis of prevalence surveys to inform definitions, burden, associations and screening methodology. *Clin Infect Dis*. 2021;73(3):e830–e841.
26 Mishra H, Reeve BWP, Palmer Z, et al. Xpert MTB/RIF ultra and Xpert MTB/RIF for diagnosis of tuberculosis in an HIV-endemic setting with a high burden of previous tuberculosis: a two-cohort diagnostic accuracy study. *Lancet Respir Med*. 2020;8:368–382.
27 Beynon F, Theron G, Respeito D, et al. Correlation of Xpert MTB/RIF with measures to assess Mycobacterium tuberculosis bacillary burden in high HIV burden areas of Southern Africa. *Sci Rep*. 2018;8:5201.
28 Behr MA, Warren SA, Salamon H, et al. Transmission of Mycobacterium tuberculosis from patients smear-negative for acid-fast bacilli. *Lancet*. 1999;353:444–449.
29 Hernández-Garduño E, Cook V, Kunimoto D, Elwood RK, Black WA, FitzGerald JM. Transmission of tuberculosis from smear negative patients: a molecular epidemiology study. *Thorax*. 2004;59:286–290.
30 Sgarbi RVE, Carbone ADSS, Paião DSG, et al. A cross-sectional survey of HIV testing and prevalence in twelve Brazilian correctional facilities. *PLoS One*. 2015;10:e0139487.