

Artificial intelligence-assisted prediction of Xpert-confirmed pulmonary tuberculosis using routine clinical and laboratory data in Southern Zambia: a retrospective pilot study

Halukala Lubobya¹

halukalalubobya@hotmail.com

Namwala District Health Office

Bislom Chikwanka

Mulungushi University

Mweene² Sepiso Masenga

Livingstone University Teaching Hospital

Research Article

Keywords: artificial intelligence, machine learning, tuberculosis, Xpert MTB/RIF, laboratory surveillance, Zambia, clinical prediction

Posted Date: September 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10591563/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Artificial intelligence has shown promise for tuberculosis screening, particularly through automated chest radiograph interpretation. Its usefulness when applied to routinely collected clinical and laboratory data in district-level African settings is less certain. We assessed whether simple machine-learning models could predict Xpert-confirmed pulmonary tuberculosis using routine surveillance data from a clinical laboratory in Southern Zambia.

Methods

We performed a retrospective pilot prediction study using tuberculosis surveillance records from Namwala District Hospital and referring facilities. Records without a valid GeneXpert result were excluded, and duplicate study numbers were reduced to one diagnostic episode. The outcome was *Mycobacterium tuberculosis* detected versus not detected by Xpert MTB/RIF. Descriptive comparisons, univariable logistic regression and a parsimonious multivariable model were undertaken. Logistic regression, random forest and gradient boosting models using body mass index and specimen type were evaluated with repeated stratified five-fold cross-validation (20 repeats). Model performance was described by the area under the receiver operating characteristic curve (AUROC), precision-recall AUC and Brier score. Bootstrap internal validation was used for the final logistic model.

Results

Of 129 records, 8 lacked a GeneXpert outcome and 3 duplicate records were removed, leaving 118 unique diagnostic episodes. MTB was detected in 17 patients (14.4%, 95% CI 9.2–21.9). Median age was 30 years, 63 (53.4%) participants were male, 42 (35.6%) were HIV positive and 95 (80.5%) were inpatients. Saliva-labelled specimens were positive in 4 of 7 cases (57.1%), compared with 13 of 109 non-saliva specimens (11.9%). Saliva remained associated with MTB detection after adjustment for body mass index (adjusted OR 10.55, 95% CI 2.09–53.37; $P = 0.004$), although the estimate was imprecise. Cross-validated AUROC was 0.593 for logistic regression, 0.522 for random forest and 0.596 for gradient boosting. The final logistic model had an optimism-corrected AUROC of 0.666 and Brier score of 0.121.

Conclusions

Routine clinical and laboratory variables had limited ability to predict Xpert-confirmed tuberculosis in this small real-world dataset. The association with saliva-labelled specimens is hypothesis-generating and requires verification. The models are not suitable for clinical deployment or exclusion of patients from

molecular testing. Larger, prospectively curated, multicentre Zambian datasets are required before artificial intelligence can support tuberculosis diagnostic prioritisation in routine care.

Background

Tuberculosis (TB) remains a major cause of infectious-disease morbidity and mortality. The World Health Organization (WHO) estimated that 10.7 million people developed TB in 2024, underscoring persistent gaps in case detection and timely access to treatment [1]. Zambia has made progress in reducing TB incidence, but diagnostic and care-cascade losses remain important. A national analysis estimated that only 45% of people with TB completed the care cascade in 2018, with the largest losses occurring before access to diagnostic testing [2].

Rapid molecular testing has substantially improved bacteriological confirmation. Xpert MTB/RIF and Xpert MTB/RIF Ultra detect *Mycobacterium tuberculosis* complex and rifampicin resistance directly from clinical specimens. Xpert Ultra has greater sensitivity than the original Xpert assay, particularly in paucibacillary disease, although interpretation may be complicated in patients with previous TB and in trace-positive results [3]. Zambia has expanded GeneXpert access through decentralised and referral networks, but laboratory demand, specimen transport, incomplete request forms, cartridge availability and instrument downtime may still affect turnaround time [4].

Artificial intelligence (AI) and machine learning can combine multiple clinical variables to estimate disease probability. In TB, the most mature application is computer-aided detection of abnormalities on digital chest radiographs. WHO recommends computer-aided detection as an alternative to human interpretation for TB screening in eligible populations [5]. Studies in Zambia and other high-burden settings have demonstrated promising performance of AI-assisted radiographic triage, while also showing that performance may vary according to previous TB, HIV status and patient characteristics [6–9].

AI models using routine non-imaging data may be less costly and easier to integrate into district-level laboratory workflows, but they are also vulnerable to missing information, coding errors and limited predictor signal. Previous studies have used routine laboratory indicators to distinguish active TB from latent infection, generally in larger and more intensively characterised populations [10, 11]. Evidence from rural Zambian clinical laboratories remains limited.

This study evaluated the feasibility of using routine clinical and laboratory surveillance data to predict Xpert-confirmed pulmonary TB in a district-level clinical setting in Southern Zambia. The objectives were to describe the surveillance population, identify factors associated with MTB detection, compare three simple machine-learning approaches and assess whether the available data were sufficient to justify future model development.

Methods

Study design and reporting

This was a retrospective, exploratory prediction-modelling study using routinely collected laboratory surveillance data. The study was designed and reported with reference to the TRIPOD + AI statement for prediction models using regression or machine-learning methods [12]. Because the dataset was small and contained few outcome events, the analysis was explicitly treated as a pilot feasibility study rather than development of a deployable diagnostic tool.

Study setting and period

The study was conducted at Namwala District Hospital, Southern Province, Zambia. The laboratory receives specimens from hospital departments and surrounding referring facilities. Data collection timestamps ranged from 10 to 17 December 2020. Valid specimen dates ranged mainly from 11 November to 17 December 2020. One specimen date recorded in 2022 despite a 2020 collection timestamp was considered a data-entry error and was not used to define the study period.

Data source and study population

The uploaded surveillance workbook contained 129 patient records and 33 variables, including demographic characteristics, anthropometry, selected laboratory measurements, clinical status, referring facility, specimen type, HIV and antiretroviral therapy status, and GeneXpert outcome. Records were eligible when a valid GeneXpert result was recorded as MTB detected or MTB not detected. Records without an outcome were excluded. When the same study number appeared more than once, the earliest record was retained as the representative diagnostic episode because duplicate outcomes were concordant.

Outcome and predictors

The primary outcome was Xpert-confirmed pulmonary TB, coded as 1 for MTB detected and 0 for MTB not detected. Candidate predictors available before or at specimen processing included age, sex, body mass index (BMI), hypertension status, white-cell count, lymphocyte percentage, haemoglobin concentration, inpatient or outpatient status, specimen type, reported TB symptoms, new or old client status, previous TB treatment, antiretroviral therapy status and HIV status.

Several variables were excluded from modelling after data-quality review. Total cholesterol and neutrophil values were identical in all records in which both were available, suggesting duplication or column misalignment. C-reactive protein frequently reproduced the height value, and one height and one weight value were implausible. Plasma sodium and fasting glucose also had substantial missingness. TB microscopy was unavailable for the analytical sample. These variables were not used in the prediction models.

Statistical analysis

Continuous variables were summarised as medians and interquartile ranges and compared using the Mann-Whitney U test. Categorical variables were summarised as counts and percentages. Fisher exact tests were used where cell counts were small. Univariable associations with MTB detection were estimated using logistic regression and reported as odds ratios (ORs) with 95% confidence intervals (CIs).

Only 17 MTB-positive outcomes were available. Contemporary prediction-model guidance indicates that model complexity must be restricted when the number of events is small [13]. Therefore, model development was intentionally limited to BMI and specimen type, the variables with the most stable exploratory signal and acceptable completeness. Specimen type was represented by one-hot indicators in machine-learning models and by saliva versus all other recorded specimen types in the parsimonious regression model. This data-driven restriction is acknowledged as a limitation.

Three classifiers were compared: logistic regression, random forest and gradient boosting. Missing specimen type in two MTB-negative records was handled by complete-case exclusion for model development, leaving 116 records. Numerical variables were standardised within each training fold, and categorical encoding was learned only from training data. Performance was assessed using repeated stratified five-fold cross-validation with 20 repeats. Metrics were mean AUROC, precision-recall AUC and Brier score. A final logistic model was fitted to all 116 complete records and internally validated using 500 bootstrap samples to estimate optimism. A high-sensitivity exploratory threshold was selected as the threshold with the greatest specificity while maintaining sensitivity of at least 90%. These operating characteristics were calculated in the development data and were not considered externally validated.

Analyses were performed in Python using NumPy, SciPy, scikit-learn and statsmodels. Two-sided P values below 0.05 were considered statistically significant for exploratory association analyses; no adjustment was made for multiple comparisons.

Ethical considerations

Ethical approval

was obtained from Mulungushi University Ethics Committee, reference number SMHS-MU2-2026 -83. Administrative permission was obtained from Namwala District Health Office and Namwala District Hospital. The requirement for individual informed consent was waived because the analysis used retrospective, de-identified routine records. All procedures were conducted in accordance with applicable Zambian regulations and the Declaration of Helsinki.

Results

Record selection and data quality

The workbook contained 129 records. Eight records without a valid GeneXpert outcome were excluded. Study numbers 25 and 26 were repeated, resulting in three additional duplicate records beyond the first diagnostic episode. The primary analytical sample therefore contained 118 unique episodes. Seventeen were MTB positive and 101 were MTB negative, corresponding to an overall positivity of 14.4% (95% CI 9.2–21.9%). Two MTB-negative records lacked specimen type, leaving 116 records for prediction modelling (Fig. 1).

Participant characteristics

The median age was 30 years (IQR 23–45 years). Sixty-three participants (53.4%) were male, 42 (35.6%) were HIV positive and 95 (80.5%) were inpatients. Median BMI was 24.0 kg/m² (IQR 21.1–28.2), median white-cell count was 8.6 x 10⁹/L (IQR 5.7–10.3), median lymphocyte percentage was 25.9% (IQR 22.2–33.1) and median haemoglobin was 14.0 g/dL (IQR 11.0–16.0). None of these demographic or laboratory variables differed materially between Xpert-positive and Xpert-negative patients (Table 1).

Table 1
Characteristics of the analytical sample according to GeneXpert outcome

Characteristic	Overall (N = 118)	MTB detected (n = 17)	MTB not detected (n = 101)	P value
Age, years, median (IQR)	30.0 (23.0–45.0)	34.0 (23.0–45.0)	30.0 (23.0–45.0)	0.963
Male sex, n (%)	63 (53.4)	9 (52.9)	54 (53.5)	0.968
HIV positive, n (%)	42 (35.6)	7 (41.2)	35 (34.7)	0.604
Inpatient, n (%)	95 (80.5)	15 (88.2)	80 (79.2)	0.392
BMI, kg/m ² , median (IQR)	23.9 (21.1–28.1)	26.0 (23.0–28.9)	23.5 (21.0–27.7)	0.163
White-cell count, x10 ⁹ /L, median (IQR)	8.6 (5.7–10.3)	8.6 (5.9–10.8)	8.6 (5.7–9.7)	0.527
Lymphocytes, %, median (IQR)	25.9 (22.2–33.1)	26.0 (22.3–31.1)	25.8 (22.2–33.3)	0.991
Haemoglobin, g/dL, median (IQR)	14.0 (11.0–16.0)	13.0 (12.0–17.0)	14.0 (11.0–16.0)	0.556
Reported TB symptoms, n/N (%)	81/113 (71.7)	10/15 (66.7)	71/98 (72.4)	0.644
Previous TB treatment, n/N (%)	8/109 (7.3)	0/15 (0.0)	8/94 (8.5)	0.597*
ART use, n (%)	37 (31.4)	5 (29.4)	32 (31.7)	0.852

*IQR, interquartile range; ART, antiretroviral therapy. P values are exploratory. *Fisher exact test; records with missing values were excluded from the relevant comparison.*

Specimen type and MTB detection

Among 116 records with specimen type, 76 (65.5%) were mucopurulent, 33 (28.4%) were mucoid and 7 (6.0%) were labelled saliva. MTB positivity was 13.2% (10/76) in mucopurulent specimens, 9.1% (3/33) in mucoid specimens and 57.1% (4/7) in saliva-labelled specimens. When saliva was compared with all other specimens, the unadjusted OR for MTB detection was 9.85 (95% CI 1.98–49.02; $P = 0.005$).

In the parsimonious logistic model, saliva-labelled specimens remained associated with MTB detection after adjustment for BMI (adjusted OR 10.55, 95% CI 2.09–53.37; $P = 0.004$). BMI was not independently associated with MTB detection (adjusted OR per 5 kg/m² increase 1.20, 95% CI 0.80–1.78; $P = 0.377$). The wide interval around the saliva estimate reflected the very small number of saliva specimens. The exploratory univariable and adjusted logistic-regression associations with MTB detection are presented in Table 2.

Table 2
Exploratory logistic-regression associations with MTB detection

Predictor	Odds ratio (95% CI)	P value
Age, per 10-year increase	0.99 (0.69–1.43)	0.963
BMI, per 5 kg/m ² increase	1.14 (0.78–1.66)	0.511
White-cell count, per 1 x10 ⁹ /L	1.09 (0.89–1.34)	0.418
Lymphocytes, per 10 percentage points	0.94 (0.58–1.54)	0.819
Haemoglobin, per 1 g/dL	1.05 (0.88–1.25)	0.578
Male sex	0.98 (0.35–2.74)	0.968
Hypertension recorded	1.66 (0.56–4.94)	0.364
Inpatient status	1.97 (0.42–9.29)	0.392
Saliva-labelled specimen	9.85 (1.98–49.02)	0.005
Reported TB symptoms	0.76 (0.24–2.43)	0.644
Old client	1.79 (0.34–9.45)	0.492
ART use	0.90 (0.29–2.77)	0.852
HIV positive	1.32 (0.46–3.77)	0.604
Saliva-labelled specimen, adjusted for BMI	10.55 (2.09–53.37)	0.004
BMI per 5 kg/m ² , adjusted for saliva	1.20 (0.80–1.78)	0.377

CI, confidence interval; ART, antiretroviral therapy. The adjusted model contained only BMI and saliva versus non-saliva specimen type.

Machine-learning model performance

Cross-validated discrimination was modest. Mean AUROC was 0.593 for logistic regression, 0.522 for random forest and 0.596 for gradient boosting. Logistic regression had the highest mean precision-recall AUC (0.372) and the lowest cross-validated Brier score (0.123). Differences among the algorithms were small relative to the variability across resamples (Table 3 and Fig. 2).

Table 3
Performance across repeated stratified five-fold cross-validation

Model	AUROC, mean (SD)	Precision-recall AUC, mean (SD)	Brier score, mean (SD)
Logistic regression	0.593 (0.165)	0.372 (0.188)	0.123 (0.020)
Random forest	0.522 (0.148)	0.294 (0.147)	0.131 (0.020)
Gradient boosting	0.596 (0.130)	0.292 (0.150)	0.131 (0.026)

AUROC, area under the receiver operating characteristic curve; SD, standard deviation. Cross-validation comprised five folds repeated 20 times.

When fitted to all 116 complete records, the final logistic model had an apparent AUROC of 0.697, precision-recall AUC of 0.379 and Brier score of 0.113. Bootstrap correction reduced the AUROC to 0.666 and increased the Brier score to 0.121.

At a development-sample threshold of 0.104, selected to preserve sensitivity of at least 90%, sensitivity was 94.1%, specificity 31.3%, positive predictive value 19.0% and negative predictive value 96.9%. The model classified 84 of 116 records as high probability, including 16 of 17 MTB-positive cases. Because the same data were used to fit the model and select this threshold, these operating characteristics are optimistic and should not guide clinical care.

Discussion

Principal findings

In this retrospective pilot study, routinely collected demographic, clinical and basic laboratory variables had limited ability to predict Xpert-confirmed pulmonary TB. None of age, sex, HIV status, inpatient status, white-cell count, lymphocyte percentage or haemoglobin concentration showed a meaningful association with MTB detection. The three machine-learning algorithms produced modest and variable cross-validated discrimination, and the final logistic model remained only moderately discriminative after correction for optimism.

The only strong exploratory signal was specimen type. Saliva-labelled specimens had approximately ten-fold greater odds of MTB detection than other recorded specimens, even after adjustment for BMI. This finding was based on four positive results among only seven saliva specimens and therefore has substantial uncertainty. It should be interpreted as a signal requiring data verification and independent replication, not as evidence that saliva is intrinsically superior for routine TB diagnosis.

Comparison with previous studies

Most successful AI applications in TB diagnosis use information-rich chest radiographs and substantially larger datasets. Prospective and retrospective Zambian studies have shown that computer-aided chest radiograph interpretation can support TB triage, while highlighting reduced specificity in patients with previous TB and the need for setting-specific thresholds [6, 7]. International evaluations have also reported substantially higher discrimination than observed in the present study [8, 9]. The contrast is expected because radiographic images contain direct pulmonary pattern information, whereas this surveillance dataset contained only a small number of general demographic and laboratory variables.

Machine-learning models based on routine laboratory indicators have distinguished active TB from latent infection in larger clinical datasets [10, 11]. Those studies generally included more outcome events and a broader panel of validated biomarkers. The current findings indicate that algorithm complexity cannot compensate for limited sample size, weak predictor signal and uncertain data quality.

The saliva association is not entirely inconsistent with previous Xpert research. In Uganda, Meyer and colleagues observed a trend toward higher Xpert positivity and significantly higher sensitivity in salivary than mucoid sputum among smear-negative adults [14]. Nevertheless, macroscopic specimen labels are subjective, and our dataset lacked culture, Xpert semiquantitative load and independent verification of specimen classification. The present estimate is therefore especially vulnerable to sparse-data bias and misclassification.

Clinical and programmatic implications

The model should not replace clinical assessment, Xpert testing or national diagnostic algorithms. Its low specificity at a high-sensitivity threshold would prioritise most specimens and therefore provide little reduction in testing workload. The high apparent negative predictive value largely reflects the low outcome prevalence and the use of development data, and it must not be used to deny molecular testing.

The most immediate value of this study is methodological. It demonstrates how routine district-level data can be audited before AI development and identifies specific improvements needed for a future study: standardised variable definitions, verified specimen categories, electronic linkage to GeneXpert outputs, consistent symptom documentation, culture or composite reference standards where feasible, and substantially larger multicentre samples.

Strengths and limitations

Strengths include the use of real-world clinical laboratory data, explicit deduplication, transparent reporting of data-quality defects, comparison of several modelling approaches and internal validation using both repeated cross-validation and bootstrap correction. The analysis also avoided presenting accuracy alone and included discrimination, precision-recall performance, calibration-related Brier scores and clinically interpretable operating characteristics.

The study has important limitations. First, the sample contained only 118 unique episodes and 17 positive outcomes, resulting in imprecise associations and unstable model estimates. Second, data were collected over a short period at one district-level laboratory, limiting transportability. Third, the retrospective dataset lacked a full data dictionary, and some variables showed duplication, implausible values or extensive missingness. Fourth, Xpert was the study outcome rather than mycobacterial culture or a composite clinical reference standard; the study therefore predicted Xpert detection rather than all active TB. Fifth, predictor restriction was data driven and no independent test set was available. Sixth, threshold performance was calculated in the development data. Finally, the study did not evaluate effects on turnaround time, treatment initiation, patient outcomes, costs, equity or workflow.

Recommendations for future research

A future Zambian study should prospectively collect several thousand diagnostic episodes across multiple districts, with enough MTB-positive outcomes to support model complexity and subgroup evaluation. Candidate predictors should be specified in advance and include standardised symptom data, previous TB, HIV and CD4 status where available, validated full blood count indices, specimen-quality measures, digital chest radiograph scores and GeneXpert platform metadata. A temporal or geographic external-validation cohort should remain untouched during model development. Decision-curve analysis and prospective silent evaluation should precede any clinical implementation [15]. Model reporting and risk-of-bias assessment should follow TRIPOD + AI and PROBAST [12, 16].

Conclusions

In this district-level Zambian clinical dataset, simple artificial-intelligence and machine-learning approaches had limited ability to predict Xpert-confirmed pulmonary tuberculosis. A strong association between saliva-labelled specimens and MTB detection was observed, but it was based on only seven specimens and requires verification. The current models are not suitable for clinical deployment or for excluding patients from molecular testing. The study supports investment in larger, prospectively curated, multicentre surveillance datasets before AI-assisted TB diagnostic prioritisation is considered in routine Zambian care.

Abbreviations

AI
artificial intelligence
ART

antiretroviral therapy
AUROC
area under the receiver operating characteristic curve
BMI
body mass index
CI
confidence interval
HIV
human immunodeficiency virus
IQR
interquartile range
MTB
Mycobacterium tuberculosis
OR
odds ratio
PR-AUC
precision-recall area under the curve
TB
tuberculosis
TRIPOD + AI
Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis -
Artificial Intelligence
WHO
World Health Organization.

Declarations

Ethics approval and consent to participate

Ethical approval for this study was obtained from the Mulungushi University Biomedical Research Ethics Committee and the Zambian National Health Research Authority (NHRA). The study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki for Medical Research Involving Human Subjects. Permission to access anonymized routine data from the local health data system Smartcare and patient records was granted by the management of Namwala District Hospital. As the study relied on retrospective, anonymized data with no direct patient contact, the requirement for individual informed consent was waived by the approving ethics committees.

Consent for publication

Not applicable. No identifiable individual patient information is presented.

Competing interests

The authors declare that they have no competing interests.

Use of generative artificial intelligence

A generative artificial-intelligence language tool was used to support manuscript organisation and language editing. The authors independently verified the analysis, numerical results, interpretation and references and accept full responsibility for the final manuscript. The tool was not used to create or alter patient data.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Author Contribution

HL conceived the study, curated the data, contributed to the analysis and drafted the manuscript. BCM contributed to clinical interpretation. SM supervised the statistical analysis. All authors critically revised the manuscript, approved the final version and accept accountability for the work.

Acknowledgements

The authors acknowledge the management and staff of Namwala District Hospital, Namwala District Health Office, referring facilities and laboratory personnel responsible for routine tuberculosis testing and data recording.

Availability of data and materials

The de-identified dataset is subject to institutional and Ministry of Health governance requirements and is not publicly posted. A suitably anonymised dataset and analysis code may be available from the corresponding author on reasonable request and following approval by the relevant authorities.

References

1. World Health Organization. Global tuberculosis report 2025. Geneva: World Health Organization; 2025.
2. Lungu P, Kerkhoff AD, Kasapo CC, et al. Tuberculosis care cascade in Zambia - identifying the gaps in order to improve outcomes: a population-based analysis. *BMJ Open*. 2021;11:e044867. 10.1136/bmjopen-2020-044867.
3. Dorman SE, Schumacher SG, Alland D, et al. Xpert MTB/RIF Ultra for detection of *Mycobacterium tuberculosis* and rifampicin resistance: a prospective multicentre diagnostic accuracy study. *Lancet Infect Dis*. 2018;18:76–84. 10.1016/S1473-3099(17)30691-6.
4. Girdwood S, Pandey M, Machila T, et al. The integration of tuberculosis and HIV testing on GeneXpert can substantially improve access and same-day diagnosis and benefit tuberculosis programmes: a diagnostic network optimisation analysis in Zambia. *PLOS Glob Public Health*. 2023;3:e0001179. 10.1371/journal.pgph.0001179.
5. World Health Organization. Use of computer-aided detection software for tuberculosis screening: policy statement. Geneva: World Health Organization; 2025.
6. Kazemzadeh S, Kiraly AP, Nabulsi Z, et al. Prospective multi-site validation of AI to detect tuberculosis and chest X-ray abnormalities. *NEJM AI*. 2024;1(10). 10.1056/Aloa2400018.
7. Kagujje M, Kerkhoff AD, Nteeni M, et al. The performance of computer-aided detection digital chest X-ray reading technologies for triage of active tuberculosis among persons with a history of previous tuberculosis. *Clin Infect Dis*. 2023;76:e894–901. 10.1093/cid/ciac679.
8. 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:e543–54. 10.1016/S2589-7500(21)00116-3.
9. Qin ZZ, Sander MS, Rai B, et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: a multisite evaluation of the diagnostic accuracy of three deep-learning systems. *Sci Rep*. 2019;9:15000. 10.1038/s41598-019-51503-3.
10. Luo Y, Xue Y, Song H, et al. Machine learning based on routine laboratory indicators promoting the discrimination between active tuberculosis and latent tuberculosis infection. *J Infect*. 2022;84:648–57. 10.1016/j.jinf.2021.12.046.
11. Luo Y, Xue Y, Liu W, et al. Development of diagnostic algorithm using machine learning for distinguishing between active tuberculosis and latent tuberculosis infection. *BMC Infect Dis*. 2022;22:965. 10.1186/s12879-022-07954-7.
12. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. 10.1136/bmj-2023-078378.
13. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: Part II - binary and time-to-event outcomes. *Stat Med*. 2019;38:1276–96. 10.1002/sim.7992.
14. Meyer AJ, Atuheire C, Worodria W, et al. Sputum quality and diagnostic performance of GeneXpert MTB/RIF among smear-negative adults with presumed tuberculosis in Uganda. *PLoS ONE*.

2017;12:e0180572. 10.1371/journal.pone.0180572.

15. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak.* 2006;26:565–74. 10.1177/0272989X06295361.
16. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med.* 2019;170:51–8. 10.7326/M18-1376.
17. Muyoyeta M, Moyo M, Kasese N, et al. Implementation research to inform the use of Xpert MTB/RIF in primary health care facilities in high TB and HIV settings in resource constrained settings. *PLoS ONE.* 2015;10:e0126376. 10.1371/journal.pone.0126376.

Figures

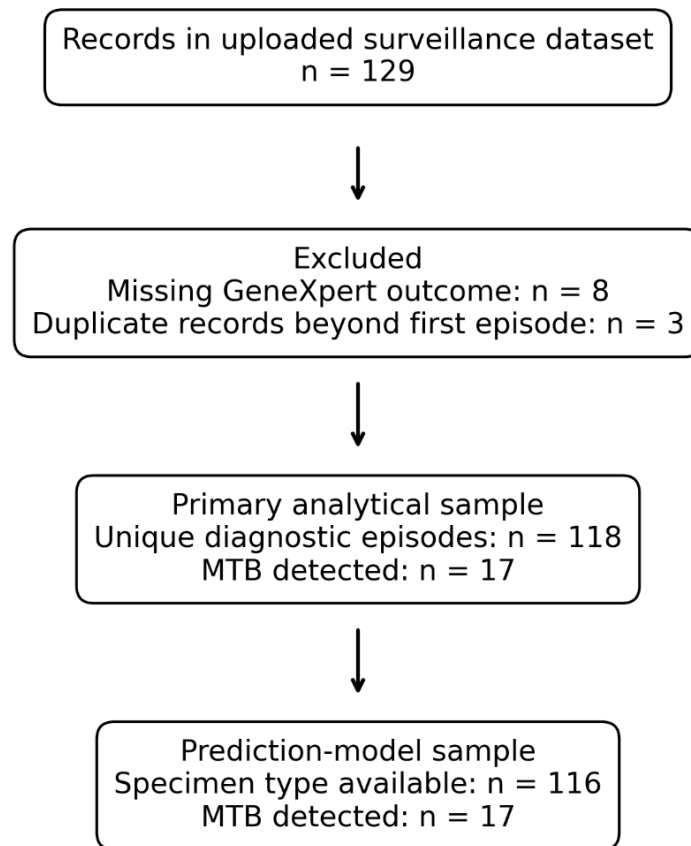


Figure 1

Selection of records for the primary and prediction-model analyses.

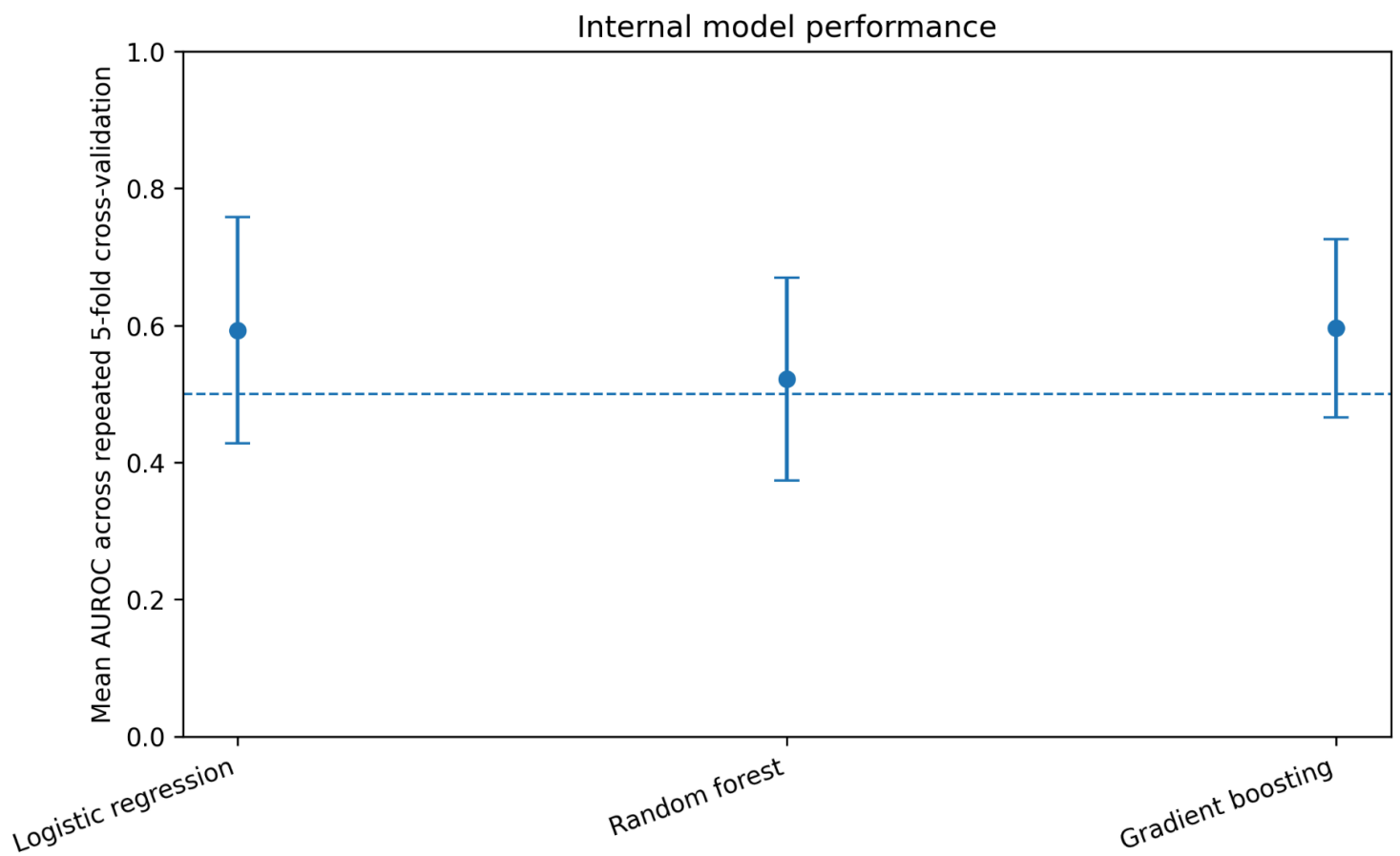


Figure 2

Mean AUROC and standard deviation across repeated stratified five-fold cross-validation. The dashed line represents no-discrimination performance.