Computer-aided detection thresholds for digital chest radiography interpretation in tuberculosis diagnostic algorithms

Fiona Vanobberghen[1,2], Alfred Kipyegon Keter[3,4,5], Bart K.M. Jacobs[3], Tracy R. Glass[1,2], Lutgarde Lynen[3], Irwin Law[6], Keelin Murphy[7], Bram van Ginneken[7], Irene Ayakaka[8], Alastair van Heerden[4,9], Llang Maama[10] and Klaus Reither[1,2]

[1]Swiss Tropical and Public Health Institute, Allschwil, Switzerland. [2]University of Basel, Basel, Switzerland. [3]Institute of Tropical Medicine, Antwerp, Belgium. [4]Centre for Community Based Research, Human Sciences Research Council, Pietermaritzburg, South Africa. [5]Ghent University, Ghent, Belgium. [6]Global Tuberculosis Programme, World Health Organization, Geneva, Switzerland. [7]Radboud University Medical Center, Nijmegen, Netherlands. [8]SolidarMed, Partnerships for Health, Maseru, Lesotho. [9]SAMRC/Wits Developmental Pathways for Health Research Unit (DPHRU), Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. [10]Disease Control Directorate, National Tuberculosis Program, Ministry of Health, Maseru, Lesotho.

Corresponding author: Fiona Vanobberghen (fiona.vanobberghen@swisstph.ch)

Shareable abstract (@ERSpublications)
Determination of computer-aided detection thresholds for tuberculosis diagnostic algorithms remains challenging. Prevalence surveys can be leveraged for threshold determination in population-based settings. https://bit.ly/3se02LZ

Cite this article as: Vanobberghen F, Keter AK, Jacobs BKM, *et al*. Computer-aided detection thresholds for digital chest radiography interpretation in tuberculosis diagnostic algorithms. *ERJ Open Res* 2024; 10: 00508-2023 [DOI: 10.1183/23120541.00508-2023].

Received: 19 July 2023
Accepted: 25 Oct 2023

Abstract

Objectives Use of computer-aided detection (CAD) software is recommended to improve tuberculosis screening and triage, but threshold determination is challenging if reference testing has not been performed in all individuals. We aimed to determine such thresholds through secondary analysis of the 2019 Lesotho national tuberculosis prevalence survey.

Methods Symptom screening and chest radiographs were performed in participants aged ⩾15 years; those symptomatic or with abnormal chest radiographs provided samples for Xpert MTB/RIF and culture testing. Chest radiographs were processed using CAD4TB version 7. We used six methodological approaches to deal with participants who did not have bacteriological test results to estimate pulmonary tuberculosis prevalence and assess diagnostic accuracy.

Results Among 17 070 participants, 5214 (31%) had their tuberculosis status determined; 142 had tuberculosis. Prevalence estimates varied between methodological approaches (0.83–2.72%). Using multiple imputation to estimate tuberculosis status for those eligible but not tested, and assuming those not eligible for testing were negative, a CAD4TBv7 threshold of 13 had a sensitivity of 89.7% (95% CI 84.6–94.8) and a specificity of 74.2% (73.6–74.9), close to World Health Organization (WHO) target product profile criteria. Assuming all those not tested were negative produced similar results.

Conclusions This is the first study to evaluate CAD4TB in a community screening context employing a range of approaches to account for unknown tuberculosis status. The assumption that those not tested are negative – regardless of testing eligibility status – was robust. As threshold determination must be context specific, our analytically straightforward approach should be adopted to leverage prevalence surveys for CAD threshold determination in other settings with a comparable proportion of eligible but not tested participants.

Introduction

Over four million of the estimated 10.6 million people who developed tuberculosis in 2021 remained undiagnosed [1]. Intensified case-finding strategies are impeded by high operational costs and logistical challenges, particularly since tuberculosis disproportionally affects vulnerable and underserved populations [2]. The World Health Organization (WHO) recommends incorporating chest radiography in screening and triaging algorithms but this relies on expert interpretation thus limiting its reach [3, 4]. Since 2021, the

WHO has recommended the use of computer-aided detection (CAD) software for screening individuals aged ⩾15 years [5]. These software systems use artificial intelligence to replace human readings of chest radiographs, and produce abnormality scores indicating likelihood of tuberculosis. Some developers recommend a threshold score for universal use, above which an individual would be classified as having abnormalities suggestive of tuberculosis [6, 7]. Other developers, such as CAD4TB (Delft Imaging Systems, Netherlands), recommend scores tailored to each setting due to variation in CAD abnormality scores between different software and populations [5, 6, 8–13], yet context-specific determination is often not done due to logistical and financial constraints [8, 14, 15]. While there exists comprehensive guidance on the use of CAD for tuberculosis diagnostic algorithms, determining an appropriate threshold remains one of the main challenges [6, 9].

The majority of studies reporting on CAD threshold determination only included individuals who met some criteria for microbiological testing, such as having symptoms and/or abnormal chest radiographs [8, 16–27]. Other studies have assumed that individuals who did not meet the criteria for microbiological testing did not have tuberculosis [12, 28, 29]. Neither approach correctly accounts for those not undergoing microbiological testing, which may be a substantial proportion of individuals in large community screening settings. To date, only one study has attempted to address this problem in a principled way, using latent class analyses (LCA) to account for the missingness in tuberculosis status [11]. There remains a paucity of guidance for threshold determination when the majority of individuals did not undergo microbiological testing [30].

Our work was motivated by the design of a clinical trial, TB TRIAGE+, which will compare two tuberculosis diagnostic algorithms for detection of active and subclinical pulmonary tuberculosis [31]. We did a secondary analysis of the 2019 Lesotho national tuberculosis prevalence survey to determine a CAD4TB version 7 (CAD4TBv7) threshold for use in the trial, meeting the WHO minimum target product profile criteria for screening and diagnostic algorithms of >90% sensitivity, and preferably >70% specificity, using nonsputum based rapid tests [32, 33]. We consider key subgroups, such as HIV status, which may affect diagnostic accuracy [8, 16].

Material and methods

Participants and study design

The Lesotho Ministry of Health conducted a nationwide tuberculosis prevalence survey in 2019 [34]. All consenting adults aged ⩾15 years were screened using a symptoms questionnaire and referred for a chest radiograph (Innomed units fitted with the Samsung Detector panels and Sedecal Dragon 5 kW Digital X-ray units). Participants with symptoms (cough, fever, night sweats, body weight loss) or a chest radiograph with abnormal lung fields suggestive of tuberculosis (as determined by the clinician in the field) were asked to provide two spot sputum specimens for testing by Xpert MTB/RIF Ultra (Cepheid, Sunnyvale, CA, USA) (hereafter, Ultra) and liquid culture (Mycobacteria Growth Indicator Tube (MGIT); Becton Dickinson, Franklin Lakes, NJ, USA). Of note, clinicians interpreting chest radiographs in the field were instructed to over-read the chest radiographs to avoid missing potential tuberculosis cases. CAD4TB version 5 scores were automatically provided with the digital chest radiograph output, but were not consistently used for field chest radiograph interpretation. Chest radiographs which were determined in the field reading to be abnormal (whether suggestive of tuberculosis or not) were sent for reading by a central radiologist from LTE Medical Solutions, South Africa. In addition, approximately 11% of chest radiographs which were determined in the field to be normal were sent for central reading for quality control purposes. For this secondary analysis, chest radiograph images were processed by Radboud University Medical Center using CAD4TB (Delft Imaging, NL) versions 6 and 7, yielding a continuous score between 0–100 with higher scores indicating higher likelihood of tuberculosis.

Ethical approval was obtained for the survey from the Lesotho Research and Ethics Committee, and participants provided written informed consent [34]. Data are available online [35]. Approval for this analysis was obtained through an agreement between the Lesotho National Leprosy and Tuberculosis Program and SolidarMed Lesotho, supported by the Swiss Tropical and Public Health Institute.

Analysis

We excluded people who were on tuberculosis treatment, did not have complete symptom data, did not have chest radiography done or did not have CAD4TB results. We considered participants as true tuberculosis cases if positive on Ultra (excluding trace) and/or culture; as not having tuberculosis if negative on both tests; and otherwise having unknown tuberculosis status. This differed from the definition of the national survey which had a case definition of culture positive, and/or Ultra positive with chest radiograph suggestive of tuberculosis plus no history of tuberculosis [34] to avoid missing some potentially

true positive cases. Asymptomatic participants without chest radiography abnormalities were not eligible for microbiological testing, although some underwent testing for unknown reasons; their test results were excluded from the analyses (of note, none had tuberculosis). Conversely, some participants who were eligible did not undergo testing for various reasons; their tuberculosis status remained unknown. Participants were offered HIV counselling and testing according to national guidelines; if test results were not available, then HIV status was determined by self-report.

We used six methodological approaches to deal with participants whose tuberculosis status remained unknown, depending on their eligibility for testing (table 1). In the first approach, we performed a complete case (CC) analysis, excluding all participants with unknown tuberculosis status (hereafter, CC/CC, indicating that CC was used for the participants who were ineligible for testing along with CC for those who were eligible for testing but did not have results). In the second approach, we assumed that participants with unknown tuberculosis status did not have tuberculosis (not tested=negative (NN) regardless of eligibility status; NN/NN). In the third approach, participants who were ineligible for testing were assumed not to have tuberculosis, while those who were eligible for testing but did not have results were excluded (NN/CC). In the fourth approach, we used multiple imputation (MI) to impute unknown tuberculosis statuses under a missing at random assumption (MI/MI) [36]. We imputed missing results of HIV status, central reading of chest radiographs, Ultra and culture, using logistic regression models, incorporating as independent variables sex, age (linear), cough (no cough/cough of duration <14 days/cough duration ⩾14 days), fever, night sweats, body weight loss, field reading of chest radiograph, CAD4TBv6 and CAD4TBv7 scores (both linear) [37]. After Ultra and culture results were imputed, we derived tuberculosis status as above. We estimated diagnostic test sensitivity and specificity across CAD4TBv7 thresholds, with estimates from the multiple imputed datasets combined using Rubin's rules [36]. In the fifth approach, we assumed that those who were ineligible for testing did not have tuberculosis, and then used multiple imputation in a similar way as above for the participants who were eligible for testing but did not have results (NN/MI). In the sixth approach, we used Bayesian LCA to model and impute the probabilities of diagnostic test results (Ultra and culture) conditional on unobserved (latent) true tuberculosis status and observed participant data including other diagnostic test results [38, 39]. Classical LCA assumes that diagnostic tests are independent conditional on the true disease status, which is unlikely to hold [38, 39]. Therefore, we extended classical LCA by allowing conditional dependence between any tuberculosis symptoms, chest radiograph abnormality suggestive of tuberculosis (based on field results confirmed by central reading if abnormality detected in the field), CAD4TBv6 score ⩾56, CAD4TBv7 score ⩾K where K varied within the range 0–100, Ultra and culture among the latent class of true tuberculosis cases; and conditional dependence between any tuberculosis symptoms, chest radiograph abnormality suggestive of tuberculosis, CAD4TBv6 score ⩾56, and CAD4TBv7 score ⩾K among the latent class of true nontuberculosis cases [38, 39]. Modelling was performed among participants with Ultra and culture results. A further 162 participants who had abnormal field chest radiograph but no central reading were excluded, because complete data were required for the LCA. We used probit regression methods with unknown model parameters assigned Gaussian priors. The final sensitivity and specificity estimates were obtained by combining the estimates based on the microbiologically tested subset with the imputed estimates of the microbiologically-untested subset.

For each methodological approach, we estimated the tuberculosis prevalence and compared it to our "best" estimate of prevalence. The latter was determined by assuming that the prevalence among those eligible but not tested was the same as for those with test results available, and that the sensitivity of chest radiograph readings among those without symptoms was the same as among those symptomatic [40]. We used this comparison to inform the relative performance of each methodological approach. We estimated sensitivity and specificity with respect to CAD4TBv7 score cut-offs estimating logit-transformed 95% confidence intervals, plotted receiver operating characteristic (ROC) curves, and estimated area under the curves (AUC).

Under the NN/NN approach (see table 1), we visually assessed the difference in sensitivity and specificity within the following subgroups as the CAD4TBv7 score thresholds varied: sex, any symptoms (cough, fever, night sweats or weight loss), HIV status and history of tuberculosis. We compared diagnostic accuracy between subgroups for the threshold determined as above. Participants missing subgroup information were excluded from the corresponding analyses. Analyses were done in Stata version 16 [41] and R version 4.2.1. Data and code are available on request *via* the corresponding author.

Results

Among 21 719 people who consented to the survey [34], we included in this analysis 17 070 (79%) (figure 1), across 11 069 households in 54 clusters. Overall, 10 209 (60%) participants were female, 3066 (22% of

https://doi.org/10.1183/23120541.00508-2023

TABLE 1 Methodological approaches to dealing with participants with unknown tuberculosis status, and estimated tuberculosis prevalences

Methodological approach	Description	Handling of participants not eligible (and not tested)	Handling of participants eligible but not tested	Number of participants included	Estimated tuberculosis prevalence, %
Complete case (CC/CC)	All participants with unknown tuberculosis status were excluded from the analyses regardless of testing eligibility status	Excluded	Excluded	5214	2.7
Not tested=negative (NN/NN)	All participants with unknown tuberculosis status were assumed to not have tuberculosis regardless of testing eligibility status, since the majority of them did not have symptoms nor abnormalities on chest radiograph	Assumed negative	Assumed negative	17 070	0.83
Combination of NN and CC (NN/CC)	Participants who were ineligible for testing were assumed not to have tuberculosis, while those who were eligible for testing but did not have results were excluded	Assumed negative	Excluded	16 391	0.87
Multiple imputation (MI/MI)	Multiple imputation was used to impute the unknown tuberculosis statuses for all regardless of testing eligibility status, under the assumption that the data were missing at random, using chained equations (MICE) with 80 imputations[#] [36]	Multiply imputed	Multiply imputed	17 070	1.1
Combination of NN and MI (NN/MI)	Participants who were ineligible for testing were assumed not to have tuberculosis, then multiple imputation was used to impute the unknown tuberculosis statuses for those who were eligible for testing but did not have results, in a similar way as above[#]	Assumed negative	Multiply imputed	17 070	0.89
Latent class analysis (LCA)	Bayesian latent class analysis was used to model the dependencies between true underlying tuberculosis status, symptoms, chest radiograph, CAD4TB scores and microbiological test results, among microbiologically tested individuals, and then used to impute the probability of testing positive on Ultra or culture for those who did not have microbiological results, under the assumption that data were missing at random[¶]	Bayesian LCA	Bayesian LCA	16 908	1.2

[#] Missing results for HIV status, central reading of chest radiographs, Ultra and culture, were imputed using logistic regression models, incorporating as independent variables sex, age (linear), cough (no cough/cough of duration <14 days/cough duration ⩾14 days), fever, night sweats, body weight loss, field reading of chest radiograph, CAD4TBv6 and CAD4TBv7 scores (both linear) [37]. The linearity assumption of CAD4TB scores was assessed using splines and deemed reasonable; [¶] Diagnostic tests included were any tuberculosis symptoms, chest radiograph abnormality suggestive of tuberculosis (based on field results confirmed by central reading if abnormality detected in the field), CAD4TBv6 score ⩾56 (categorised for parsimony with this threshold chosen pragmatically to combine high sensitivity and specificity in these data), CAD4TBv7 score ⩾K where K varied within the range 0–100, Ultra and culture. CAD4TBv6/7: CAD4TB version 6/7. Ultra: Xpert MTB/RIF Ultra.

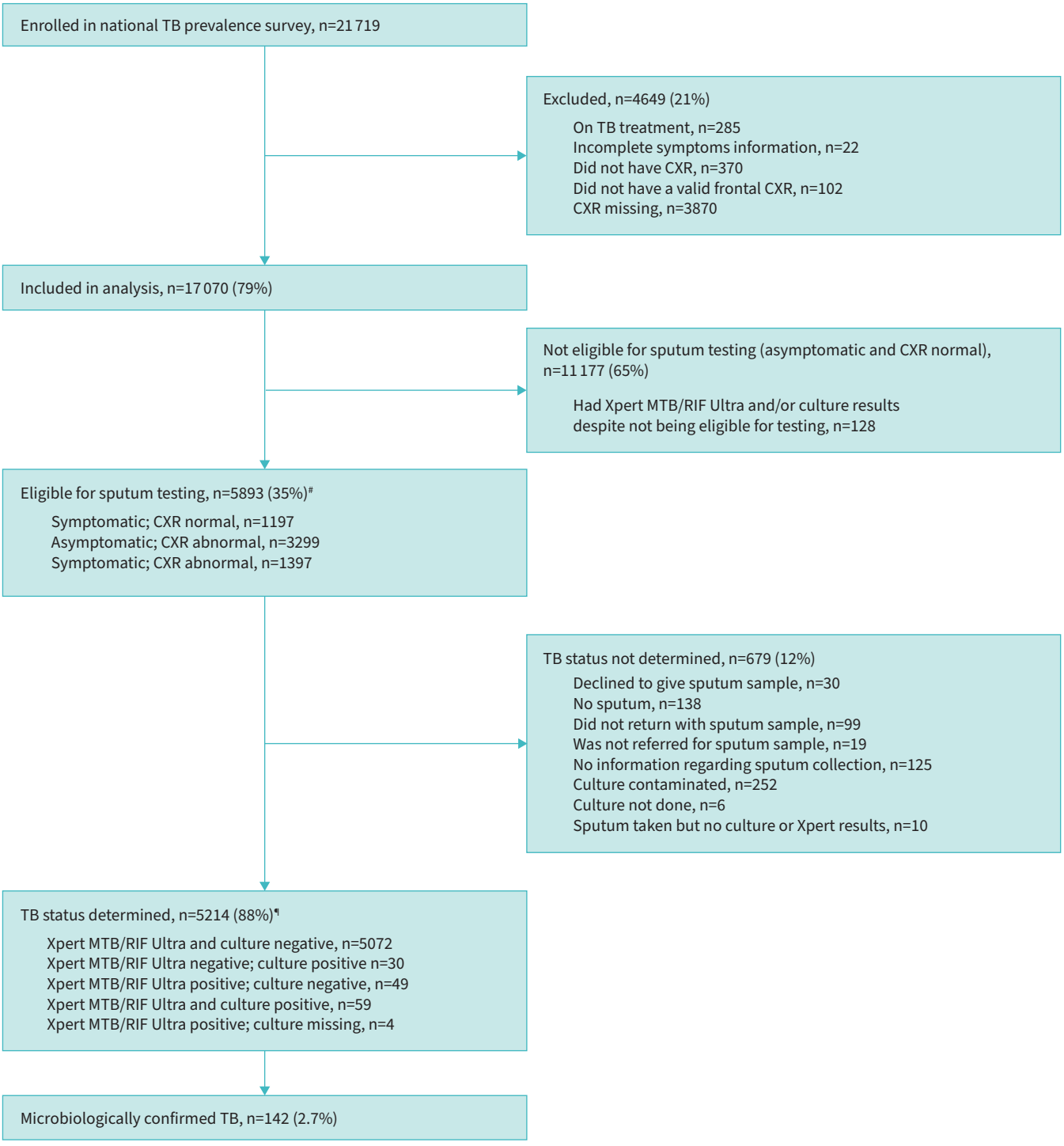

FIGURE 1 Inclusion of participants and tuberculosis determination. # Determination of symptomatic based on cough, fever, night sweats and body weight loss. Chest radiography based on field reading, with abnormal indicating lung fields suggestive of tuberculosis; ¶ Determined as having tuberculosis for this secondary analysis, if Ultra (excluding trace) or culture positive; not having tuberculosis if both negative; otherwise unknown tuberculosis status. CXR: chest radiograph; TB: tuberculosis.; Ultra: Xpert MTB/RIF Ultra.

14 159 with results) were HIV positive, 1485 (9%) reported a history of tuberculosis, 3595 (21%) reported ⩾1 symptom and 991 (6%) had a cough lasting ⩾14 days (table 2). Of the participants, 28% (4696) had a chest radiograph field reading suggestive of tuberculosis; 4035 of these, and 6540 (38%) participants in total, had their chest radiographs interpreted centrally. There was poor agreement between field and central

TABLE 2 Participant characteristics.

	Unknown tuberculosis status (N=11 856)	No tuberculosis (N=5072)	Bacteriologically confirmed tuberculosis (N=142)	Total (N=17 070)
Setting				
Peri-urban	498 (4%)	459 (9%)	9 (6%)	966 (6%)
Rural	7310 (62%)	3080 (61%)	96 (68%)	10 486 (61%)
Urban	4048 (34%)	1533 (30%)	37 (26%)	5618 (33%)
Sex, female	7599 (64%)	2561 (50%)	49 (35%)	10 209 (60%)
Age, years	33 (22–50)	52 (34–65)	58 (40–69)	38 (24–57)
Miner				
No	11 055 (96%)	3958 (81%)	89 (65%)	15 102 (91%)
Yes	518 (4%)	931 (19%)	47 (35%)	1496 (9%)
Missing	283	183	6	472
Ever smoked				
No	8212 (69%)	2568 (51%)	56 (40%)	10 836 (64%)
Yes	3611 (31%)	2493 (49%)	85 (60%)	6189 (36%)
Missing	33	11	1	45
HIV status				
Negative	7504 (80%)	3506 (76%)	83 (68%)	11 093 (78%)
Positive	1919 (20%)	1108 (24%)	39 (32%)	3066 (22%)
Missing	2433	458	20	2911
History of tuberculosis	654 (6%)	803 (16%)	28 (20%)	1485 (9%)
Any of the four tuberculosis symptoms: cough, fever, night sweats or weight loss	1103 (9%)	2428 (48%)	64 (45%)	3595 (21%)
Cough of duration ⩾14 days	119 (1%)	835 (16%)	37 (26%)	991 (6%)

Results are number (column percentage of nonmissing data) for categorical variables and median (interquartile range) for continuous variables.

results: of 4035 chest radiographs which were interpreted in the field as suggestive of tuberculosis, only 419 (10%) were interpreted the same centrally, reflecting the directive for field readers to over-read the chest radiographs. Of the participants 35% (5893) were eligible for sputum sampling, among whom tuberculosis status was determined in 5214 (88%), and of those 142 (2.7%) had tuberculosis (figure 1). CAD4TBv7 score distributions differed markedly by tuberculosis status (Supplemental Figure 1). As expected, people who did not get a confirmatory bacteriological test tended to have even lower scores than those who tested negative for tuberculosis. Sensitivity of field chest radiographs was high at 97.9% (95% CI 94.0–99.3) and specificity was 73.1% (72.4–73.7) under the NN/NN approach. Chest radiographs incorporating the central reading where available had much lower sensitivity at 47.2% (39.2–55.4), but specificity of 96.5% (96.2–96.7) under the NN/NN approach.

Assuming that the prevalence among those eligible but not tested was the same as among those tested (2.7%), and that the sensitivity of chest radiographs among those without symptoms was the same as among those symptomatic [40] (94.7%), our prevalence estimate was 0.97%. Under the CC/CC, NN/NN and NN/CC approaches, the estimated prevalences were 2.7%, 0.83% and 0.87%, respectively (table 1). Under the MI/MI approach, the average number of imputed tuberculosis cases across the multiply imputed datasets was 41 (range 15–222), corresponding to an average prevalence of 1.1% (95% CI 0.7–1.4). Under the NN/MI approach, the average number of imputed cases was 9 (range 2–16), equating to an average prevalence of 0.89% (95% CI 0.74–1.0). Under the LCA approach, the underlying tuberculosis prevalence was estimated to be approximately 1.2% (Supplemental Figure S2). Therefore, all approaches except CC/CC yielded fairly similar, realistic prevalence estimates, with the NN/MI approach yielding a prevalence estimate closest to our "best" estimate.

Under the NN/NN approach, the highest CAD4TB score which yielded a sensitivity of >90% was 13, with sensitivity 90.1% (95% CI 84.0–94.1) (figure 2 and table 3). The associated specificity for this threshold was 74.2% (95% CI 73.6–74.9). Results were almost identical under the NN/CC approach. Under the NN/MI approach, the sensitivity was slightly lower (89.7%, 95% CI 84.6–94.8) for the same specificity, but lower thresholds resulted in substantial drops in specificity. Therefore, the threshold of 13 was chosen, and

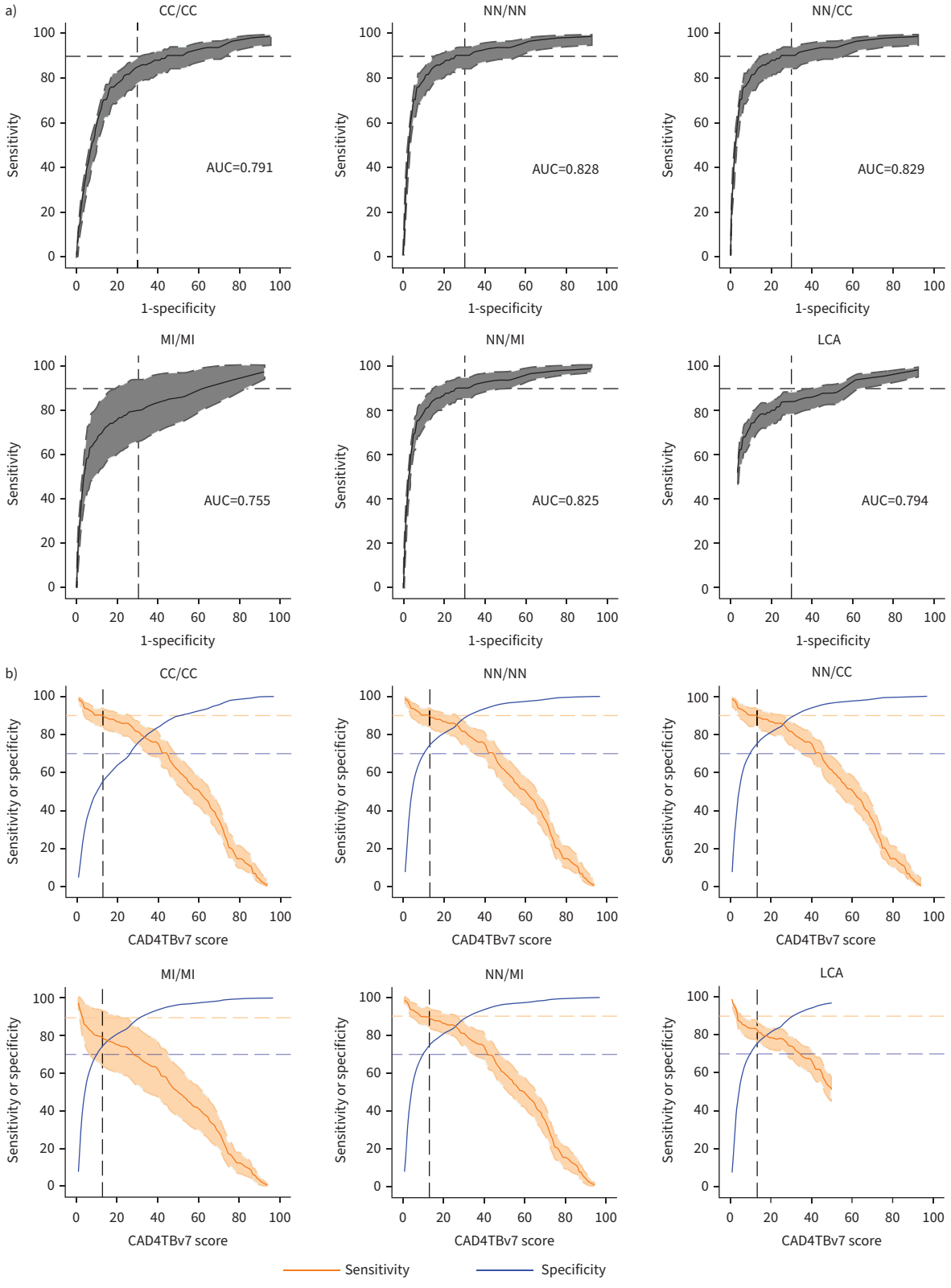

FIGURE 2 Diagnostic accuracy of CAD4TBv7 under different methodological approaches. a) Receiver operating characteristic curves for CAD4TBv7. Horizontal and vertical dashed lines illustrate WHO minimum target product profile criteria of >90% sensitivity and >70% specificity, respectively. b) Sensitivity (orange) and specificity (blue) plotted against CAD4TBv7 scores, under the six methodological approaches (CC/CC, NN/NN, NN/CC, MI/MI, NN/MI, LCA; see table 1) with 95% CIs. Horizontal dashed lines illustrate WHO minimum target product profile criteria of >90% sensitivity

and >70% specificity. Vertical dashed lines are at the CAD4TBv7 threshold of 13. LCA analyses were performed for CAD4TBv7 scores ⩽50 only, so the AUC is approximate and the graphs are truncated. AUC: area under the curve; CAD4TBv7: CAD4TB version 7; CC: complete case; LCA: latent class analysis; MI: multiple imputation; NN: not tested=negative.

the NN/NN approach was deemed preferable for subsequent analyses due to its similar results and analytical simplicity.

Comparing diagnostic accuracy by subgroups as a function of CAD4TBv7 cut-off score under the NN/NN approach, sensitivity tended to be worse for women compared to men, reaching differences of around 30%, while specificity was slightly better (figure 3). For those with symptoms, the sensitivity tended to be better and specificity worse compared to those without symptoms. For HIV status, there was a lot of variability in the sensitivity, with the relative performance depending on the score, however there was little difference in specificity. Specificity was markedly lower among those with a history of tuberculosis. For the chosen CAD4TBv7 threshold score of 13, diagnostic accuracy varied by subgroup although confidence intervals were wide (Supplemental Table S1). Specificity was noticeably poorer among men, those with symptoms and those with a history of tuberculosis.

Discussion

To our knowledge, this is the first study to determine CAD thresholds in a large-scale screening context employing a range of approaches to account for unknown tuberculosis status among a substantial proportion of individuals. In this community screening setting as part of a national tuberculosis prevalence survey, we have demonstrated the limitations of ignoring the participants who did not undergo bacteriological testing, resulting in severe overestimation of the prevalence and underestimation of the specificity of CAD4TB. Assuming that those who did not undergo testing did not have tuberculosis is commonly done [12, 28, 29], but this will misclassify some individuals who have subclinical tuberculosis with normal chest radiograph, and addressing subclinical tuberculosis is essential if WHO incidence targets are to be met [40]. MI has previously been recommended for estimating tuberculosis prevalence from national surveys for the subgroup that is eligible but not tested [37], but as far as we are aware, this is the first time that MI has been applied for CAD threshold determination. The method requires the assumption of data being missing at random, which is unlikely to hold for participants who were ineligible for testing and which is a substantial proportion of the study population. Accordingly, under the MI/MI approach, we observed an overestimation of the number of tuberculosis cases among those not tested. For this approach to be successful, extensions would be required under strong unverifiable assumptions to allow for the data to be missing not at random [36]. In contrast, the NN/MI approach yielded plausible prevalence estimates,

TABLE 3 Sensitivity and specificity of CAD4TBv7 for selected thresholds

CAD4TBv7 threshold	CC/CC N=5214	NN/NN N=17 070	NN/CC N=16 391	MI/MI N=17 070	NN/MI N=17 070	LCA N=16 908
3	96.5; 23.6	96.5; 38.1	96.5; 38.6	89.9; 38.1	96.3; 38.1	93.1; 38.5
4	93.7; 30.4	93.7; 47.6	93.7; 48.2	86.0; 47.6	93.5; 47.6	87.3; 48.1
5	93.7; 35.4	93.7; 54.2	93.7; 54.8	85.1; 54.2	93.4; 54.2	87.1; 54.7
6	93.0; 39.6	93.0; 59.0	93.0; 59.7	83.8; 59.0	92.7; 59.0	85.4; 59.5
7	92.3; 42.8	92.3; 62.6	92.3; 63.3	82.7; 62.6	92.0; 62.6	85.1; 63.2
8	91.5; 45.5	91.5; 65.5	91.5; 66.2	81.7; 65.5	91.3; 65.5	84.3; 66.1
9	90.1; 48.0	90.1; 67.7	90.1; 68.4	80.3; 67.7	89.9; 67.7	83.5; 68.3
10	90.1; 50.2	90.1; 69.7	90.1; 70.4	79.9; 69.7	89.9; 69.7	83.4; 70.3
11	90.1; 52.1	90.1; 71.5	90.1; 72.1	79.7; 71.5	89.8; 71.5	83.3; 72.1
12	90.1; 54.1	90.1; 73.0	90.1; 73.7	79.5; 73.0	89.7; 73.0	83.2; 73.6
13	**90.1; 55.4**	**90.1; 74.2**	**90.1; 74.9**	**79.2; 74.3**	**89.7; 74.2**	**83.2; 74.9**
14	88.7; 56.7	88.7; 75.4	88.7; 76.1	77.9; 75.4	88.4; 75.4	81.0; 76.0
15	88.7; 58.2	88.7; 76.6	88.7; 77.2	77.6; 76.6	88.3; 76.6	81.1; 77.2
16	88.0; 59.2	88.0; 77.4	88.0; 78.1	76.9; 77.4	87.6; 77.4	79.5; 78.0

Results presented are sensitivity; specificity, and correspond to those presented in figure 3. The number of participants contributing to each analysis are shown in the column headers. Methodological approaches are described in table 1. CAD4TBv7: CAD4TB version 7; CC: complete case; LCA: latent class analysis; MI: multiple imputation; NN: not tested=negative.

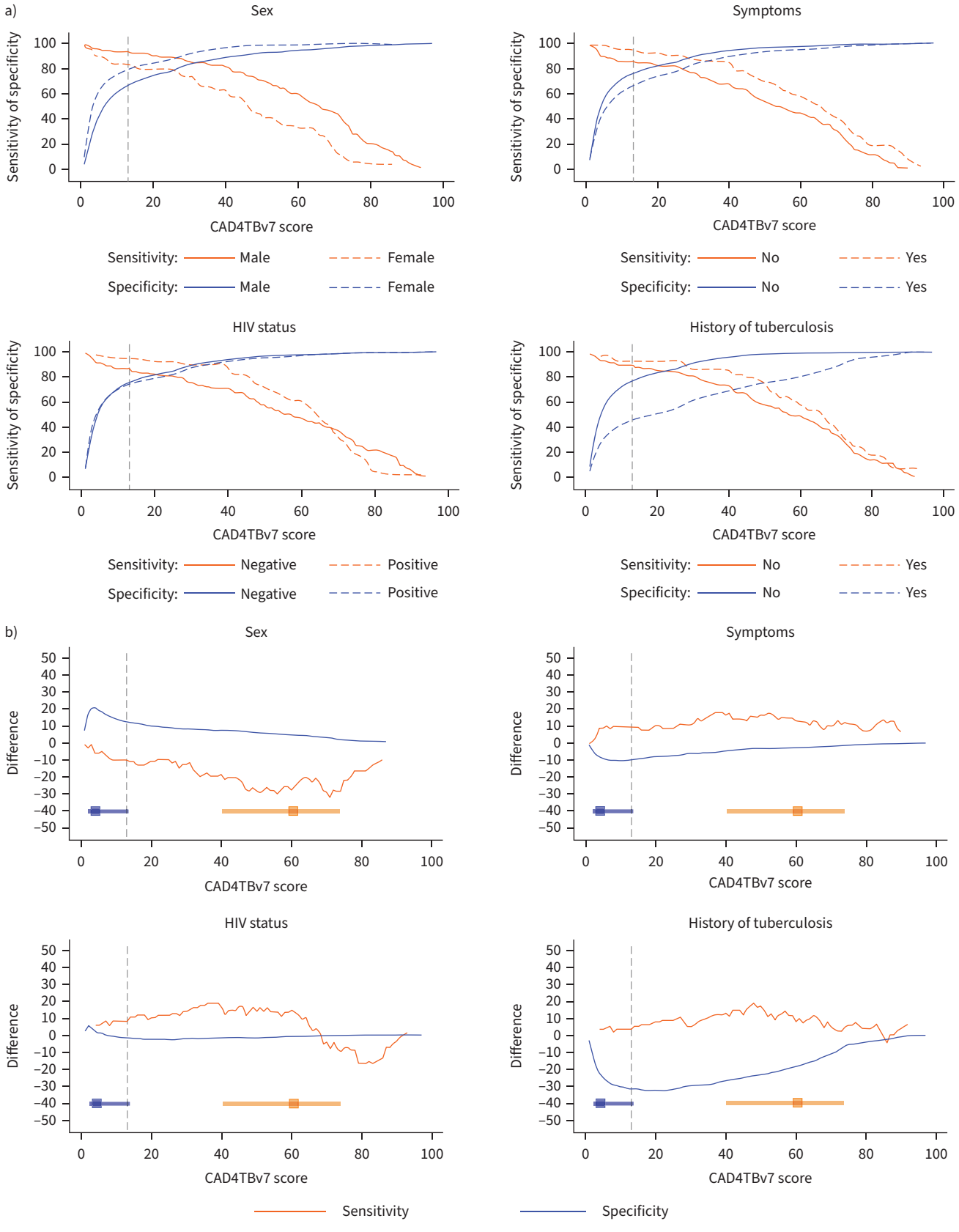

FIGURE 3 Diagnostic accuracy of CAD4TBv7 by subgroups, under NN/NN approach (see table 1). a) Sensitivity and specificity. b) Differences in sensitivity and specificity. Differences are for women *versus* men (*i.e.* values above 0 indicate better diagnostic accuracy for women), symptoms

versus no symptoms, HIV positive *versus* HIV negative, and history of tuberculosis *versus* no history of tuberculosis. Vertical dashed lines are at the CAD4TBv7 threshold of 13. Squares and horizontal lines show the overall median and interquartile range, respectively, of the CAD4TBv7 scores, for those with tuberculosis (orange) and those without tuberculosis (blue), to illustrate where the data lie and hence aid interpretation of the sensitivity and specificity results. CAD4TBv7: CAD4TB version 7; NN/NN: not tested=negative

with the missing at random assumption more likely to hold for the MI only among those eligible for testing but without test results. However, all approaches except the LCA approach assume that the reference standard is perfect. In contrast, LCA assumes that all diagnostic tests are imperfect but contribute information to determine the likelihood of tuberculosis, so it incorporates uncertainties from the diagnostic tests to determine the reference standard. However, because it is in essence used as a probabilistic imputation here, it also requires the (implausible) missing at random assumption like the MI/MI approach. Bayesian LCA has only been used in one previous study of CAD thresholds [11], reflecting its complexity and specialist knowledge required. We have shown that, in our setting, the assumption that individuals who did not undergo testing – usually due to not having symptoms nor abnormalities on chest radiography – did not have tuberculosis is reasonable, even though this also included a smaller group that was eligible for testing, but for other reasons did not have bacteriological test results available. The NN/NN approach is therefore optimistic with respect to sensitivity, but results were similar to those from the NN/MI approach and did not deviate strongly from the LCA approach. Since the overestimation of tuberculosis cases under the LCA approach is typically in those with lower scores, this means that the sensitivity is most unlikely underestimated thus yielding conservative estimates for the sensitivity in this case. Under our best estimate of prevalence, we estimated an additional five cases of undetected tuberculosis among those not eligible for testing, equating to an overestimation of the sensitivity of at most 4% under the NN/MI approach.

To meet our predefined criterion of >90% sensitivity, we fixed the CAD4TBv7 threshold at 13. To our knowledge, only two previous studies have estimated an optimal CAD4TBv7 threshold. In the first, a large study in nearly 24 000 participants, a much higher threshold of 50 was identified (for 90% sensitivity with 73% specificity) [23]. However, the study population was very different, being people who presented or were referred to tuberculosis screening centres in Bangladesh with 98% having symptoms. In the second, a community-based study in South Africa restricted to participants with sputum test results, a threshold of 20 was estimated (to most closely match radiologist sensitivity of 81%, with specificity 57%) [42]. This highlights the importance of threshold determination needing to be context specific.

We evaluated diagnostic accuracy in a range of subgroups, including HIV status for the first time in a community-based prevalence survey, where HIV testing and counselling was offered to all participants and status determined in >80%. We did not observe noticeable differences in diagnostic accuracy among people living with HIV, in line with one other study [12], while others found poorer performance among people living with HIV [8, 16, 21, 22]. We observed poorer specificity among men, likely related to men generally having unhealthier lifestyles and poorer health-seeking behaviour, therefore having higher scores on average than women. These results correspond to those from a previous study [17]. Specificity was poorer among those with symptoms compared to those without symptoms, also likely related to those with symptoms being more likely to have lung damage for any reason and therefore higher scores. Lastly, specificity was poorer in those with history of tuberculosis, as expected and corresponding to findings in other settings [11, 17, 23, 43].

A strength of this study is the large number of individuals included in a nationally-representative survey. There are some limitations. Our tuberculosis case definition differed from that of the national survey, at the risk of including false positives (*e.g.* Ultra positive with recent history of tuberculosis). Additionally, Ultra or culture positivity was used as a reference standard for the sensitivity and specificity estimation, despite both tests, as well as their combination, being imperfect and even more so in a community-based setting, where sensitivity of bacteriological tests is lower [44, 45]. Diagnostic accuracy of chest radiograph was assessed for comparison purposes, but those results should be interpreted cautiously since: 1) field readers were instructed to over-read the chest radiographs therefore artificially inflating the sensitivity and decreasing the specificity; and 2) abnormal chest radiograph was an inclusion criterion for bacteriological testing therefore the results are susceptible to selection bias. While our analyses did not take into account a potential clustering effect, this should not have an impact on the chosen threshold. However, confidence limits of the sensitivity and specificity may be too narrow as a result. It is worth noting that confidence intervals for specificity were narrower than those for sensitivity due to the larger sample size. A large number of individuals had missing chest radiographs; those from urban areas were more likely to be missing than those from rural or peri-urban areas therefore potentially compromising the representativeness

of our sample, although the sex and age distributions were broadly similar to those with chest radiographs available. We did not account for the 19% of individuals who were invited to be part of the survey but refused [34], who were more likely to be younger, male and living in peri-urban or urban areas. Inverse probability weighting methods could be used to address these issues of missing chest radiographs and survey nonparticipation [37], but we would anticipate minimal impact on the threshold determination.

In conclusion, we have demonstrated the need for context-specific threshold determination, supporting recent calls for regulation of companies to provide guidance for effective adoption of CAD with appropriate oversight by WHO [30, 42]. In our setting, a CAD4TBv7 threshold of 13 was close to WHO target product profile criteria. Methodologically, we have illustrated the importance of careful consideration of how to account for untested individuals in the analyses. We have shown that the NN/NN approach – namely assuming all those not tested are negative regardless of testing eligibility status – is robust, yielding similar results to the NN/MI approach. Our analytically straightforward approach should be adopted to leverage prevalence surveys for CAD threshold calibration in other settings, which typically have lower tuberculosis prevalences, provided that the proportion of participants eligible but not tested and the population characteristics are broadly comparable to our study.

Provenance: Submitted article, peer reviewed.

Ethics statement: Ethical approval was obtained for the survey from the Lesotho Research and Ethics Committee, and participants provided written informed consent. Data are available online. Approval for this analysis was obtained through an agreement between the Lesotho National Leprosy and Tuberculosis Program and SolidarMed Lesotho, supported by the Swiss Tropical and Public Health Institute.

Conflict of interest: F. Vanobberghen declares a grant from EDCTP (RIA2018D-2498; TB TRIAGE+). A.K. Keter declares no conflicts of interest. B.K.M. Jacobs declares a grant from EDCTP (RIA2018D-2498; TB TRIAGE+). T.R. Glass declares no conflicts of interest. L. Lynen declares being director and member of the Board of Governors of ITM. I. Law declares no conflicts of interest. K. Murphy declares no conflicts of interest. B. van Ginneken declares grants from EDCTP, royalties from Delft Imaging Systems and Thirona, and stocks in Thirona. I. Ayakaka declares no conflicts of interest. A. van Heerden declares grants from EDCTP, NIH and BMGF. L. Maama declares no conflicts of interest. K. Reither declares a grant from EDCTP (RIA2018D-2498; TB TRIAGE+), a grant from SNSF (CRSII5_213514 Sinergia) and participation on a Data Safety Monitoring Board or Advisory Board (TrENDxTB). The authors alone are responsible for the views expressed in this article and they do not necessarily represent the views, decisions or policies of the institutions with which they are affiliated.

Support statement: This project is part of the European and Developing Countries Clinical Trials Partnership 2 (EDCTP2) programme supported by the European Union (grant number: RIA2018D-2498; TB TRIAGE+). Funding information for this article has been deposited with the Crossref Funder Registry.

References

1 World Health Organization. Global tuberculosis report 2022. Geneva, World Health Organization,2022.
2 Moutinho S. Tuberculosis is the oldest pandemic, and poverty makes it continue. *Nature* 2022; 605: S16–S20.
3 World Health Organization. Consolidated guidelines on HIV prevention, testing, treatment, service delivery and monitoring: recommendations for a public health approach. Geneva, World Health Organization, 2021.
4 Frija G, Blažić I, Frush DP, *et al.* How to improve access to medical imaging in low- and middle-income countries? *EClinicalMedicine* 2021; 38: 101034.
5 World Health Organization. WHO consolidated guidelines on tuberculosis Module 2: Screening – Systematic screening for tuberculosis disease. Geneva, World Health Organization, 2020.
6 Stop TB Partnership. Screening and triage for TB using computer-aided detection (CAD) technology and ultra-portable X-ray systems: a practical guide. Date last accessed: 8 November 2023. https://www.stoptb.org/file/15474/download
7 FIND. Digital chest radiography and computer-aided detection (CAD) solutions for TB diagnostics, 2021. Date last accessed: 12 December 2023. www.finddx.org/wp-content/uploads/2023/02/20210401_technology_landscape_computer_aided_tb_FV_EN.pdf
8 Tavaziva G, Harris M, Abidi SK, *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* 2021; 74: ciab639.
9 World Health Organization. Determining the local calibration of computer-assisted detection (CAD) thresholds and other parameters: a toolkit to support the effective use of CAD for TB screening. Geneva, World Health Organization, 2021.

10 Qin ZZ, Sander MS, Rai B, *et al.* Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. *Sci Rep* 2019; 9: 15000.
11 Mungai B, Ong'angò J, Ku CC, *et al.* Accuracy of computer-aided chest X-ray in community-based tuberculosis screening: Lessons from the 2016 Kenya National Tuberculosis Prevalence Survey. *PLOS Glob Public Health* 2022; 2: e0001272.
12 Fehr J, Konigorski S, Olivier S, *et al.* Computer-aided interpretation of chest radiography reveals the spectrum of tuberculosis in rural South Africa. *NPJ Digital Med* 2021; 4: 106.
13 Ahmad Khan F, Pande T, Tessema B, *et al.* Computer-aided reading of tuberculosis chest radiography: moving the research agenda forward to inform policy. *Eur Respir J* 2017; 50: 1700953.
14 Madhani F, Maniar RA, Burfat A, *et al.* Automated chest radiography and mass systematic screening for tuberculosis. *Int J Tuberc Lung Dis* 2020; 24: 665–673.
15 Odume B, Chukwu E, Fawole T, *et al.* Portable digital X-ray for TB pre-diagnosis screening in rural communities in Nigeria. *Public Health Action* 2022; 12: 85–89.
16 Breuninger M, van Ginneken B, Philipsen RHHM, *et al.* Diagnostic accuracy of computer-aided detection of pulmonary tuberculosis in chest radiographs: a validation study from Sub-Saharan Africa. *PLoS ONE* 2014; 9: e106381.
17 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–ee81.
18 Maduskar P, Muyoyeta M, Ayles H, *et al.* Detection of tuberculosis using digital chest radiography: automated reading vs. interpretation by clinical officers. *Int J Tuberc Lung Dis* 2013; 17: 1613–1620.
19 Melendez J, Sánchez CI, Philipsen RHHM, *et al.* An automated tuberculosis screening strategy combining X-ray-based computer-aided detection and clinical information. *Sci Rep* 2016; 6: 25265.
20 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.
21 Muyoyeta M, Maduskar P, Moyo M, *et al.* The sensitivity and specificity of using a computer aided diagnosis program for automatically scoring chest X-rays of presumptive TB patients compared with Xpert MTB/RIF in Lusaka Zambia. *PLoS ONE* 2014; 9: e93757.
22 Philipsen RHHM, Sánchez CI, Maduskar P, *et al.* Automated chest-radiography as a triage for Xpert testing in resource-constrained settings: a prospective study of diagnostic accuracy and costs. *Sci Rep* 2015; 5: 12215.
23 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–ee54.
24 Rahman MT, Codlin AJ, Rahman MM, *et al.* An evaluation of automated chest radiography reading software for tuberculosis screening among public- and private-sector patients. *Eur Respir J* 2017; 49: 1602159.
25 Zaidi SMA, Habib SS, Van Ginneken B, *et al.* Evaluation of the diagnostic accuracy of computer-aided detection of tuberculosis on chest radiography among private sector patients in Pakistan. *Sci Rep* 2018; 8: 12339.
26 Kik SV, Gelaw SM, Ruhwald M, *et al.* Diagnostic accuracy of chest X-ray interpretation for tuberculosis by three artificial intelligence-based software in a screening use-case: an individual patient meta-analysis of global data. *medRxiv* 2022; preprint
27 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.
28 Koesoemadinata RC, Kranzer K, Livia R, *et al.* Computer-assisted chest radiography reading for tuberculosis screening in people living with diabetes mellitus. *Int J Tuberc Lung Dis* 2018; 22: 1088–1094.
29 Melendez J, Philipsen RHHM, Chanda-Kapata P, *et al.* Automatic versus human reading of chest X-rays in the Zambia National Tuberculosis Prevalence Survey. *Int J Tuberc Lung Dis* 2017; 21: 880–886.
30 Geric C, Qin ZZ, Denkinger CM, *et al.* The rise of artificial intelligence reading of chest X-rays for enhanced TB diagnosis and elimination. *Int J Tuberc Lung Dis* 2023; 27: 367–372.
31 Reither K. Tuberculosis diagnostic trial of CAD4TB screening alone compared to CAD4TB screening combined with a CRP triage test, both followed by confirmatory Xpert MTB/RIF Ultra in communities of Lesotho and South Africa. Date last accessed: 16 June 2023. Date last updated: 2 December 2022. https://ichgcp.net/clinical-trials-registry/nct05526885
32 World Health Organization. High-priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting. Geneva, World Health Organization, 2014.
33 Nathavitharana RR, Yoon C, Macpherson P, *et al.* Guidance for studies evaluating the accuracy of tuberculosis triage tests. *J Infect Dis* 2019; 220: Supplement_3, S116-S125.
34 Matji R, Maama L, Roscigno G, *et al.* Policy and programmatic directions for the Lesotho tuberculosis programme: findings of the national tuberculosis prevalence survey, 2019. *PLoS ONE* 2023; 18: e0273245.

35 Kak N, Matji R, Maama L, *et al.* Policy and programmatic directions for the Lesotho tuberculosis programme: Findings of the national tuberculosis prevalence survey, 2019 [Dataset]. Dryad. https://doi.org/10.5061/dryad.905qfttnq

36 White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011; 30: 377–399.

37 Floyd S, Sismanidis C, Yamada N, *et al.* Analysis of tuberculosis prevalence surveys: new guidance on best-practice methods. *Emerg Themes Epidemiol* 2013; 10: 10.

38 Keter AK, Lynen L, Heerden AV, *et al.* Evaluation of tuberculosis diagnostic test accuracy using Bayesian latent class analysis in the presence of conditional dependence between the diagnostic tests used in a community-based tuberculosis screening study. *PLOS ONE* 2023; 18: e0282417.

39 Keter AK, Lynen L, Van Heerden A, *et al.* Implications of covariate induced test dependence on the diagnostic accuracy of latent class analysis in pulmonary tuberculosis. *J Clin Tuberc Other Mycobact Dis* 2022; 29: 100331.

40 Kendall EA, Shrestha S, Dowdy DW. The epidemiological importance of subclinical tuberculosis. A critical reappraisal. *Am J Respir Crit Care Med* 2021; 203: 168–174.

41 StataCorp. Stata Statistical Software: Release 16. College Station, TX, StataCorp LLC, 2019.

42 Fehr J, Gunda R, Siedner MJ, *et al.* CAD4TB software updates: different triaging thresholds require caution by users and regulation by authorities. *Int J Tuberc Lung Dis* 2023; 27: 157–160.

43 Codlin AJ, Dao TP, Vo LNQ, *et al.* Independent evaluation of 12 artificial intelligence solutions for the detection of tuberculosis. *Sci Rep* 2021; 11: 23895.

44 Saavedra B, Mambuque E, Nguenha D, *et al.* Performance of Xpert MTB/RIF Ultra for tuberculosis diagnosis in the context of passive and active case finding. *Eur Respir J* 2021; 58: 2100257.

45 Floyd S, Klinkenberg E, de Haas P, *et al.* Optimising Xpert-Ultra and culture testing to reliably measure tuberculosis prevalence in the community: findings from surveys in Zambia and South Africa. *BMJ Open* 2022; 12: e058195.