

1 **Title**

2 Diagnostic accuracy of Chest X-Ray Computer Aided Detection software and blood biomarkers for
3 detection of prevalent and incident tuberculosis in household contacts followed up for 5 years

4

5 **Authors**

6 Liana Macpherson¹, Sandra V. Kik², Matteo Quartagno¹, Francisco Lakay³, Marche Jaftha⁴, Nombuso
7 Yende⁴, Shireen Galant⁴, Saalikha Aziz³, Remy Daroowala³, Richard Court³, Arshad Taliep³, Keboile
8 Serole³, Rene T. Goliath³, Nashreen Omar Davies³, Amanda Jackson³, Emily Douglass⁵, Bianca
9 Sossen³, Sandra Mukasa³, Friedrich Thienemann^{3,6}, Taeksun Song⁴, Morten Ruhwald², Robert J.
10 Wilkinson^{3,7,8}, Anna K. Coussens^{3,9}, Hanif Esmail^{1,3,10#} on behalf of the Imaging of TB household
11 contacts group*

12

- 13 1. MRC Clinical Trials Unit at University College London, UK
14 2. FIND, Switzerland
15 3. Centre for Infectious Diseases Research in Africa, Institute of Infectious Disease and
16 Molecular Medicine and Department of Medicine, University of Cape Town, Republic of
17 South Africa
18 4. Institute of Infectious Disease and Molecular Medicine and Department of Medicine,
19 University of Cape Town, Observatory 7925, Republic of South Africa
20 5. Rutgers- New Jersey Medical School, Center for Emerging Pathogens, Newark, NJ, United
21 States.
22 6. Department of Internal Medicine, University Hospital of Zurich, University of Zurich,
23 Switzerland.
24 7. Francis Crick Institute London NW1 1AT, London, UK
25 8. Department of Infectious Diseases, Imperial College London W12 0NN, UK
26 9. Infectious Diseases and Immune Defence Division, Walter and Eliza Hall Institute of Medical
27 Research (WEHI), Parkville 3052, Australia
28 10. WHO Collaborating Centre for TB Research and Innovation, Institute for Global Health,
29 University College London.
30

31 # Corresponding author: h.esmail@ucl.ac.uk

32 *Listed in appendix

33

34 **Summary**

35

36 We found that the diagnostic accuracy of CAD-CXR to identify prevalent TB cases in household TB
37 contacts was high but >30% with scores above recommended CAD thresholds who were
38 bacteriologically negative on routine testing baseline were subsequently diagnosed suggest that the
39 potential of CAD-CXR screening is not maximised.

40

41 **Abstract**

42 Background

43 WHO Tuberculosis (TB) screening guidelines recommend computer-aided detection (CAD) software
44 for chest radiograph (CXR) interpretation. However, studies evaluating their diagnostic and
45 prognostic accuracy are limited.

46

47 Methods

48 We conducted a prospective cohort study of household TB contacts in South Africa. Participants all
49 underwent baseline CXR and sputum investigation (routine [single spontaneous] and enhanced
50 [additionally 2-3 induced] sputum investigation and passive and active follow-up for incident TB. CXR
51 were processed comparing 3 CAD softwares (CAD4TBv7.0, qXRv3.0.0, and Lunit INSIGHT CXR
52 3.1.4.111). We evaluated their performance to detect routine and enhanced prevalent, and incident
53 TB, comparing the performance to blood-based biomarkers (Xpert MTB host-response, Erythrocyte
54 Sedimentation Rate, C-Reactive Protein, QuantiFERON) in a subgroup.

55

56 Findings

57 483 participants were followed-up for 4.6 years (median). There were 23 prevalent (7 routinely
58 diagnosed) and 38 incident TB cases. The AUC ROC to identify prevalent TB for CAD4TB, qXR and
59 Lunit INSIGHT CXR were 0.87 (95% CI 0.77-0.96), 0.88 (95% CI 0.79-0.97) and 0.91 (95% CI 0.83-0.99)
60 respectively. >30% with scores above recommended CAD thresholds who were bacteriologically
61 negative on routine baseline sputum were subsequently diagnosed by enhanced baseline sputum
62 investigation or during follow-up. The AUC performance of baseline CAD to identify incident cases
63 ranged between 0.60-0.65. The diagnostic performance of CAD for prevalent TB was superior to
64 blood-based biomarkers.

65

66 Interpretation

67 Our findings suggest that the potential of CAD-CXR screening for TB is not maximised as a high
68 proportion of those above current thresholds but with a negative routine confirmatory sputum have
69 true TB disease that may benefit intervention.

70

71 Funding

72 UKRI-MRC

73 **Introduction**

74 The World Health Organisation (WHO) estimated that there are up to 4 million people living with
75 undiagnosed tuberculosis (TB) (1). Active case finding aims to proactively identify individuals with
76 pulmonary TB not seeking healthcare to enable earlier treatment, thus reducing morbidity,
77 mortality, and transmission. As 50% of undiagnosed bacteriologically positive cases in the
78 community are symptom screen negative, and it is increasingly evident that CXR is more effective to
79 identify the undiagnosed TB cases (2–7). The sensitivity of CXR-based screening for active TB is
80 emphasised in the updated 2021 WHO screening guidelines, which also endorse for the first time the
81 use of computer aided detection (CAD) software for automated radiographic interpretation in those
82 aged >15 years (8). The Stop TB Partnership’s *Global Plan to Stop TB* also highlights the critical role
83 active case finding in reducing TB incidence (9).

84

85 CAD software use trained deep learning algorithms and artificial intelligence to interpret CXR for
86 signs of TB with several studies having reported equivalent accuracy compared to human readers
87 (8,10–12). However, high quality, prospective studies evaluating the diagnostic accuracy of such
88 software are limited, particularly in the screening setting and with comprehensive sampling and
89 follow-up for TB (13,14).

90

91 Confirmatory bacteriological testing is often done by assessment of a single spontaneously produced
92 sputum sample with molecular detection of *Mycobacterium tuberculosis* (*Mtb*) DNA (e.g. Xpert
93 MTB/RIF) (8). This approach is insensitive as not all those with a positive triage test are able to
94 expectorate sputum for confirmatory testing. In a recent research study evaluating CAD, only 29%
95 were able to produce a valid sputum (15). Bacteriologically confirmation increases with multiple
96 samples, sputum induction and/or use of culture, but such enhanced measures are not typically
97 performed during routine screening.

98

99 Those with CXR abnormalities without sputum confirmation are at high risk of disease progression, a
100 recent meta-analysis demonstrated that individuals with CXR changes suggestive of TB but with
101 negative sputum bacteriology subsequently have a 10% risk per year to being diagnosed with
102 bacteriologically positive TB disease (16). However, the majority of the contributory studies were
103 historical, pre-HIV epidemic, and used conventional CXR and no contemporary studies have
104 evaluated this risk utilising digital CXR and CAD approaches.

105

106 In addition, blood tests that are predictive of future TB risk have recently been proposed as a priority
107 for development with blood based transcriptional markers now in late stages of development. Such
108 blood-based assays could potentially be used alongside CXR screening for prevalent TB disease to
109 identify individuals with future TB risk. However, evaluation of these tests in a screening population
110 alongside or in combination with CXR has not previously been undertaken.

111

112 In this study we screened household contacts (HHC) of rifampicin-resistant (RR) TB patients by digital
113 CXR and undertook long term follow-up for development of TB disease in the absence of preventive
114 therapy. Our aims were to: (1) Evaluate the diagnostic accuracy of three CAD software packages
115 against microbiological reference standards to detect prevalent and incident pulmonary TB on
116 baseline CXR. (2) Determine the sensitivity and specificity of CAD software using the recommended
117 threshold scores. (3) Compare and combine CAD scores with a blood tests (Xpert MTB host-
118 response, Erythrocyte Sedimentation Rate (ESR), C-Reactive Protein (CRP), QuantiFERON) to detect
119 prevalent and incident TB.

120 **Methods**

121 Setting and participants

122 This prospective cohort study was conducted in Khayelitsha, South Africa. Recruitment of household
123 contacts (HHC) of at least rifampicin resistant (RR) TB index cases took place between November
124 2014 and September 2017 with follow up until May 2021 (17) (see Supplementary method for
125 details). Ethical approval for this study was received from: University of Cape Town (449/2014),
126 Boston University (H-35831), Rutgers University (Pro2018001966), NIH (DMID 16-0112) and
127 University College London (19219/001). This report follows the STARD guidelines for diagnostic
128 accuracy studies (18).

129

130 Procedures

131 Eligible individuals were HHC aged ≥ 18 years. At screening participants underwent medical history,
132 physical examination, HIV, TB symptoms screen (unexplained cough for ≥ 2 weeks, fever, night
133 sweats and weight loss), and a digital CXR. All participants were extensively investigated for TB at
134 baseline regardless of symptoms, with 1 attempted spontaneous, spot, sputum sample followed by
135 2-3 induced sputum samples to a total of 3 samples. Sputum induction was by nebulisation of 3%
136 hypertonic saline. Samples were processed in the accredited laboratories where auramine sputum
137 smear, Xpert MTB/RIF and mycobacterial growth indicator tube (MGIT) liquid TB culture were
138 performed according to local standard operating procedures. Female participants of child-bearing
139 potential underwent urinary pregnancy testing and if positive were excluded from the study.

140

141 As part of a nested study, participants, who were HIV-uninfected and asymptomatic, additionally
142 consented to also undergo blood sampling for a biomarker sub-study (17). Full details of inclusion
143 and exclusion criteria can be found in the Supplementary Methods. Blood samples were taken for
144 analysis, including full blood count (FBC), serum C-reactive protein (CRP), erythrocyte sedimentation
145 rate (ESR), QuantiFERON Gold (QFT-Gold) (Qiagen) and Tempus Tubes (Applied Biosystems) stored at
146 -80°C to preserve blood RNA for transcriptomic analysis including the 3-gene RNA Xpert MTB Host
147 Response [MTB-HR] (Cepheid, Sunnyvale, Ca, USA). Further details are in the Supplementary
148 Methods.

149

150 No participants received preventive therapy as they were contacts of RR-TB in line with national and
151 some international guidelines at the time of the study recommending close follow-up. (19)
152 Participants who were bacteriologically positive or with clinical concern of TB were referred to the
153 statutory TB clinic where the decision to start TB treatment was determined.

154

155 Follow up

156 Participants were asked to attend the clinic for assessment if they developed TB symptoms. All
157 participants were then invited for systematic re-screening between 24 and 36 months irrespective of
158 symptoms with 3 sputum samples taken (induced if needed) sent for smear, Xpert MTB/RIF and
159 culture. To ensure that all episodes of treated TB within the Western Cape Province were captured,
160 participants consented to access to their health service records via the Provincial Health Data Centre
161 (PHDC) and clinical medical records (20). This search took place on 19th May 2021.

162

163 Chest radiograph reading

164 Posterior-anterior CXR in full inspiration were performed using a digital X-Ray machine (Phillips
165 Essenta DR). All CXR were reported by a medical officer blinded to clinical details and
166 microbiological outcomes and were classified as abnormal – TB related, abnormal – not TB related
167 and normal.

168

169 Three commercially available CAD software: CAD4TB version 7.0 (CAD4TBv7, Delft Imaging, 's-
170 Hertogenbosch Netherlands), qXR version 3.0.0 (qXRv3, qure.ai, Mumbai, India) and Lunit INSIGHT
171 CXR version 3.1.4.111 (Lunit INSIGHT CXRv3, Lunit, Seoul, South Korea) (see Supplementary
172 Methods for details and manufacturer recommended threshold scores).

173

174 Reference Standards

175 For the determination of the diagnostic accuracy a case of prevalent TB was defined as at least 1
176 baseline sputum sample culture and/or Xpert MTB/RIF positive for *Mtb* where the participant was
177 treated for TB. We created subgroups of participants with prevalent disease to reflect cases that
178 would be identified through routine screening (routine prevalent) and those who would only be
179 identified through more intensive investigation (enhanced prevalent). Routine prevalent cases were
180 defined as those cases initiated on TB treatment in whom *Mtb* was detected using Xpert MTB/RIF on
181 the first single spontaneously produced baseline sputum sample, these are cases that would be
182 detected by current screening practices in South Africa (21). Enhanced prevalent cases were defined
183 as those initiated on TB treatment in whom *Mtb* was detected in any other baseline sputum sample
184 by either Xpert MTB/RIF or culture. Incident TB was defined as cases initiated on TB treatment in
185 whom at least 1 follow up sputum sample culture and/or Xpert MTB/RIF was *Mtb* positive or where
186 a clinician independent to the study made a clinical decision to start TB treatment. The inclusion of
187 clinically diagnosed TB was done to capture participants diagnosed and treated elsewhere during the

188 follow up period ascertained through the PHDC records. We classified participants as not having TB if
189 all baseline and follow up samples were negative for *Mtb*, and they were not initiated on TB
190 treatment.

191

192 Statistical analysis

193 Sample size was determined by the parent study (17). Area under the receiver operator curve (AUC
194 ROC) was calculated to evaluate diagnostic performance for the CAD software and other biomarkers.
195 We calculated the sensitivity and specificity of CAD software using the manufacturers pre-specified
196 or commonly used thresholds. Analyses were also performed for a number of pre-specified
197 subgroups chosen because of known associations with TB risk, clinical presentation, and plausibility
198 that they might affect CXR findings: people living with HIV (PLHIV), participants with a history of
199 previous TB, and smokers. For missing data, participants were excluded from analysis if any one of
200 the CAD software did not provide a score due to error, or data was incomplete for the baseline
201 sputum results. Statistical analyses were done with R version 4.0.5 (2021-03-31).

202 **Results**

203 983 HHC of RR-TB were identified, of whom 511 eligible adult participants consented and were
204 screened for TB. 483 participants were included in the analysis following exclusions (Figure 1).
205 Median age was 33 years, 308 (61%) were female, 109 (23%) had a history of previous TB, and 136
206 (28%) were PLHIV (Table 1). For those PLHIV 62.5% were on antiretroviral therapy (ART) and a
207 recent CD4 count was available for 54 participants (median 413.5/mm³ (IQR 236-562)). The
208 participants were followed up for a median of 4.6 years (IQR 3.9-5.2 years) 247 HIV-uninfected,
209 participants without TB symptoms met the inclusion criteria to undergo biomarker blood sampling.

210

211

212 Yields of different screening strategies

213 Of the 483 participants, at baseline screening, 60% produced a spontaneous sputum sample, 96% an
214 induced sample, and 84% had at least 3 sample results available (Table 1). 23 (4.7%) had
215 bacteriologically confirmed TB infection at baseline (prevalent TB), 7 (30%) of these were routine
216 prevalent cases, and the remaining 16 were enhanced prevalent cases.

217

218 A further 38 (7.9%) cases of incident TB, including 18 diagnosed elsewhere and captured via PHDC,
219 were identified during follow-up. Thirteen (34%) were diagnosed within 12 months, and 18 (47%)
220 between 24-36 months, median time to diagnosis was 24 months (IQR 10-34). Six incident cases
221 were clinically diagnosed, but not bacteriologically confirmed.

222

223 Fifty-one (11%) participants reported at least 1 TB symptom at baseline, of these, 8 were confirmed
224 TB at baseline (35% of the 23 prevalent cases). Of the 23 prevalent cases, 8 (35%) were smear
225 positive, 14 (61%) were Xpert MTB/RIF positive and 20 (87%) were culture positive in baseline
226 samples (Supplementary Tables 1a and b). Of the 38 incident cases 79% reported TB symptoms at
227 the time of diagnosis and 3 had extra-pulmonary disease.

228

229 Participants in this study underwent more intensive screening investigations compared to
230 international screening recommendations (22). We explored the yields of hypothetical screening
231 strategies: a) Positive symptom screen followed by Xpert MTB/RIF on a spontaneously produced
232 sputum sample and b) positive CXR (human read) followed by Xpert MTB/RIF on a spontaneously
233 produced sputum sample. These strategies detected 4/23 (17%) and 5/23 (22%) of the prevalent TB
234 cases in this cohort respectively. (Figure 2).

235

236

237

238 CAD results

239 All CAD software performed with highest accuracy for identifying prevalent TB and there were no
240 statistically significant differences between CAD systems for any of the comparisons. The AUC ROC
241 were 0.85-0.95 for identifying routine prevalent and 0.85-0.89 for identifying enhanced prevalent
242 cases (Figure 3). The CAD software performed less well at identifying participants with incident TB on
243 baseline CXR with AUC ROC of 0.60-0.65. For incident disease, there was no substantial difference in
244 the accuracy to detect TB occurring between 1-12 months, 13-24 months or over 24 months after
245 study entry (Supplementary Figure 1).

246

247 All CAD systems detected prevalent TB with significantly greater accuracy in participants with no
248 history of previous TB. There was also a trend to better performance in those who were not living
249 with HIV (see Supplementary Table 2).

250

251 Using the manufacturer recommended thresholds, the sensitivity/specificity to detect routine
252 prevalent TB cases was 0.71/0.91 for CAD4TB, 0.72/0.93 for qXR and 0.86/0.83 for Lunit INSIGHT
253 CXR. The sensitivity/specificity to detect all prevalent cases was 0.7/0.93 for CAD4TB, 0.57/0.94 for
254 qXR and 0.87/0.86 for Lunit INSIGHT CXR (see Supplementary Table 3).

255

256 Between 8% and 18% of participants had CAD scores above recommended threshold and of those 7-
257 13% were diagnosed as prevalent cases with routine sampling. However, notably 30-35% of those
258 above threshold, not diagnosed routinely, were subsequently diagnosed and treated for TB either as
259 prevalent cases identified through enhanced sampling or as incident cases (i.e. cannot be considered
260 as a "false positive"). For participants above threshold not diagnosed or treated for TB over 5 years
261 all had a previous history of TB with CAD4TB and qXR, and 87% (46/53) with Lunit INSIGHT CXR
262 (Figures 4 and 5).

263

264 For each product, we calculated thresholds using the WHO TPP optimal sensitivity (0.95) and
265 specificity (0.8) for a TB triage test for detecting any prevalent TB case. These derived thresholds
266 were substantially lower than those recommended or currently used in practice (see Supplementary
267 Table 4)

268

269 Comparison and combination with blood-based diagnostic tests

270 Of the 247 HIV uninfected participants in the *biomarker subgroup*, 245 had CRP, 247 had ESR, 242
271 had MTB-HR and 247 had QuantiFERON-Gold measured. Overall, 18/247 participants were
272 diagnosed and treated for TB (6 prevalent [1 routine, 5 enhanced] and 12 with incident TB). In this
273 subgroup, AUC of the blood-based biomarkers for all prevalent cases was lower than for CAD
274 software – CRP 0.75 (0.55-0.96), ESR 0.78 (0.60-0.96), QFT-Gold 0.58 (0.39-0.78) and MTB-HR 0.67
275 (0.39-0.96) in comparison to AUC of 0.90 (0.74-1) to 0.98 (0.93-1) for the CAD software.
276 Incorporating each blood biomarkers individually into a predictive model with CAD software did not
277 significantly improve the AUC to detect prevalent TB (Supplementary data Table 5).

278 **Discussion**

279 This is the first study reporting performance of CAD where CXR, symptom screen, routine and
280 enhanced sputum investigation have been undertaken to screen all consenting household contacts
281 for TB with subsequent follow-up for incident disease. We show that routine confirmatory testing
282 with single spontaneous Xpert MTB/RIF to diagnose PTB captures only 30% (7/23) of total prevalent
283 disease. CXR screening with CAD interpretation has excellent diagnostic performance to detect all
284 those with prevalent disease (diagnosed routinely and with enhanced investigation) with no
285 significant differences between the three software solutions we assessed, AUC 0.87-0.91. For all
286 software, performance was reduced in those with previous TB and people living with HIV. Following
287 baseline enhanced sputum investigation the performance of CAD to identify incident cases was
288 limited (range AUC 0.60-0.65). However, using the manufacturers recommended or widely used
289 thresholds we also showed that 30-35% of those not positive by routine (spontaneously produced)
290 sputum Xpert MTB/RIF were either sputum positive by more intensive investigations or were
291 diagnosed with incident TB over the follow-up period, with almost all the remainder having previous
292 TB. Furthermore, we highlight that despite excellent AUC, the recommended CAD thresholds
293 sensitivity is relatively low (71-86% for routine prevalent and 57-87% for all prevalent) and that
294 thresholds could be lowered to be able to meet the WHO optimal sensitivity and/or specificity
295 targets. Finally, we showed in a subgroup of asymptomatic HIV uninfected participants that the
296 diagnostic performance of all CAD products to detect all prevalent TB was superior to a range of
297 blood-based biomarkers and did not substantially improve with the addition of any of these blood
298 based biomarkers.

299
300 Our findings suggest that the potential of CAD-CXR based screening is not being maximised as a high
301 proportion of those above current thresholds with a negative routine confirmatory test have true TB
302 disease that may benefit from treatment. They also show for the first time in a contemporary setting
303 that those who have CXR changes suggestive of TB but who are bacteriologically negative by routine
304 approaches have a similar risk of progressing to bacteriologically positive disease as shown in the
305 recently published historical systematic review (16).

306
307 A systematic review conducted by Harris and colleagues in 2019 found three diagnostic accuracy
308 studies that used a microbiological reference standard in *screening* settings (23–25). These studies
309 examined CAD4TB only and reported sensitivities for the detection of prevalent TB ranged from 0.53
310 to 0.95 with specificities from 0.56 to 0.98. The authors also identified significant sources of bias,
311 methodological limitations, and heterogeneity between the studies. More recently a prospective
312 diagnostic accuracy study examining CAD4TB version 6, qXR version 3 and Lunit INSIGHT CXR version

313 3.1.0.0 was conducted by Soares and colleagues in prison settings in Brazil. They found that AUC
314 ranged from 0.88 to 0.9 for microbiologically confirmed TB, however, excluded 70% of participants
315 who were not able to produce spontaneous sputum samples from their primary analysis (15). Like
316 our findings, they showed that CAD performed with greater accuracy in participants who did not
317 have a history of previous TB. These participants are likely to have residual abnormalities visible on
318 CXR indistinguishable from active disease.

319

320 Our study has several strengths that differentiate it from other diagnostic accuracy studies
321 performed in screening settings: We included all household contacts who were eligible and
322 consented for TB screening and performed thorough examination of sputum with multiple samples
323 for investigation at baseline, we also followed participants up over a 5-year period. This enabled us
324 to avoid misclassification of TB cases and provided a better assessment of accuracy. However, at the
325 same time we were able to represent routine sampling to ensure wider applicability of our results.
326 This was also the first diagnostic accuracy study to incorporate and compare a range of blood-based
327 biomarkers, including the point-of-care Xpert MTB-HR test alongside assessment of CAD, although
328 this analysis was restricted to a subset of HIV uninfected participants who underwent additional
329 testing. Finally, our study was designed to minimise biases that are commonly seen in other
330 diagnostic accuracy studies, for example, we maintained independence from CAD system operators
331 and evaluated the software on a new dataset that had not been used in training of the software.

332

333 Our study has a number of limitations. Although we demonstrated that lower thresholds could be
334 used and still meet WHO TPP such cut-offs were not externally validated. We screened for
335 symptoms of TB using unexplained cough for ≥ 2 weeks, fever, night sweats and weight loss, for both
336 HIV infected and uninfected persons rather than the recommended W4SS in HIV infection (8). This
337 approach was taken for consistency and all participants were investigated at baseline regardless of
338 symptoms, thus reducing the bias in our study findings. As described, this study was not powered to
339 assess the impact of incorporating additional biomarkers on diagnostic accuracy and these need
340 further assessment in future studies. Our study was performed in household contacts in a high
341 burden setting where there is increased risk of re-infection over the study period, and although it is
342 likely that many of our findings would be applicable in other screening settings, additional studies in
343 low burden settings should be undertaken.

344

345 Following the updated WHO screening guidelines and publication of Stop TB Partnership's Global
346 Plan, we will continue to see scale up of CXR based screening and CAD with inevitable significant

347 cost. Our work has shown that although CAD software is diagnostically accurate its utility is not
348 being maximised as ~30% of those with CAD scores above threshold with negative Xpert MTB/RIF on
349 a spontaneous sputum sample, either have prevalent TB diagnosed by intensive sampling or develop
350 incident TB over time. Further work is needed to establish how best to follow up and manage these
351 patients.

352 **Legends**

353 **Figure 1: Study population.** Showing the details of participants who were included in this analysis
354 including a description of TB disease status.

355

356 **Table 1: Baseline characteristics.** Detailing the participants demographic data, sputum samples, and
357 scores from the computer aided detection (CAD) software for baseline chest radiograph
358 interpretation. This information is displayed for all participants, those with routinely diagnosed
359 prevalent Tuberculosis (TB), enhanced prevalent TB, and incident TB.

360

361 **Figure 2: Yields of three different TB screening strategies.** In this representation of the yields of
362 these screening each dot represents one study participant. The colour coding of the participants
363 represents Tuberculosis (TB) disease status as defined by bacteriological investigation. Participants
364 have also been grouped by HIV status to show the relative representation of TB in each population.
365 The figure is accompanied by a table which shows the sensitivity and specificity of the different
366 screening approaches (the numbers represent proportions and 95% confidence intervals).

367

368 **Figure 3: Performance of three different computer-aided detection (CAD) software for the**
369 **detection of prevalent and incident Tuberculosis (TB) cases.** The CAD software evaluated were
370 CAD4TB version 7.0 (CAD4TBv7, Delft Imaging, 's-Hertogenbosch Netherlands), qXR version 3.0.0
371 (qXRv3, qure.ai, Mumbai, India) and Lunit INSIGHT CXR version 3.1.4.111 (Lunit INSIGHT CXRv3 ,
372 Lunit, Seoul, South Korea). Routine prevalent TB was defined as those initiated on TB treatment in
373 whom *Mycobacterium tuberculosis (Mtb)* was detected using Xpert MTB/RIF on the first single
374 spontaneously produced baseline sputum sample, all prevalent TB was defined as those initiated on
375 TB treatment in whom at *Mtb* was detected using Xpert MTB/RIF and/or culture on any other
376 baseline sputum sample, and incident TB was defined as those cases initiated on TB treatment in
377 whom *Mtb* was detected in at least one follow up sputum sample by Xpert MTB/RIF and/or culture
378 or where the initiation of TB treatment was based on clinical grounds. Accuracy was measured using
379 area under receiver operator curve (AUC ROC). The numbers represent proportion (95% confidence
380 interval).

381

382 **Figure 4: Representation of study participants above and below the manufacturer or commonly**
383 **used threshold for each CAD software (CAD4TB, qXR and Lunit INSIGHT CXR) by TB status.** This
384 figure provides a visual representation of the distribution of computer-aided detection software
385 scores for each software used (CAD4TB v7.0, qXR v3.0.0, Lunit INSIGHT CXR v3.1.4.111), broken

386 down by TB disease status. Routine prevalent TB was defined as those initiated on TB treatment in
387 whom *Mycobacterium tuberculosis* (*Mtb*) was detected using Xpert MTB/RIF on the first single
388 spontaneously produced baseline sputum sample, all prevalent TB was defined as those initiated on
389 TB treatment in whom *Mtb* was detected using Xpert MTB/RIF and/or culture on any other
390 baseline sputum sample, and incident TB was defined as those cases initiated on TB treatment in
391 whom *Mtb* was detected in at least one follow up sputum sample by Xpert MTB/RIF and/or culture
392 or where the initiation of TB treatment was based on clinical grounds. Each dot represents a study
393 participant, and the colour coding of the dot represents the bacteriological status of that participant.
394 For example, some incident cases are represented by black dots indicating that these participants
395 were bacteriologically negative, these represent the cases of incident TB that were diagnosed
396 clinically. Horizontal lines have been added to represent three thresholds – the threshold above
397 which the score is consistent with TB infection as recommended by the manufacturer or commonly
398 used in practice, the threshold derived by using the WHO target product profile (TPP) optimal
399 specificity for a TB triage test, and the threshold derived by using the WHO target product profile
400 (TPP) optimal sensitivity for a TB triage test. The latter two thresholds were generated using all study
401 participants/all prevalent cases of TB.

402

403 **Figure 5: Representation of participants with baseline CAD scores for pulmonary TB above the**
404 **manufacturers recommended thresholds (or in the case of CAD4TB, a commonly used threshold).**
405 *Each participant with a score above recommended thresholds above which a diagnosis of TB is likely*
406 *is represented as a human figure. Each figure is accurately represented as male or female, and the*
407 *proportions of participants with prevalent (routine and enhanced) and incident TB are shown.*
408 *Numbers represent proportion (95% confidence interval). The chest radiograph illustrations show an*
409 *example of the typical output from each of the CAD products.*

410 **Acknowledgments**

411 The study was conceived by HE, SK, MR, AC and RJW. The study methods were determined by HE,
412 LM, SK, MQ and AC. Clinical activities and data collection were undertaken by RD, RC, AT, KS, RG,
413 NOD, AJ, ED, BS, SM, FT. Laboratory activities were undertaken by FL, MJ, NY, SG, SA and TS. The
414 data were curated by LM, ED, AC and AJ. Analysis was performed by LM with supervision and support
415 from HE, SK and MQ. The first draft of the manuscript was prepared by LM, HE and SK with all co-
416 authors subsequently reviewing and contributing to the manuscript.

417

418 **Funding**

419 This work was supported by a Medical Research Council award (grant number: MR/V00476X/1) to
420 HE. HE is partially supported by a Medical Research Council unit grant (grant number: MC UU
421 00004/04). RJW is supported by the Francis Crick Institute which receives funding from Cancer
422 Research UK (grant number: FC2112), UK Research and Innovation-Medical Research Council (grant
423 number: CC2112) and Wellcome (grant number: CC2112). He also receives funding from Wellcome
424 (grant number: 203135), South African Medical Research Council via its Strategic health Partnerships
425 and the Bill and Melinda Gates Foundation. Work from authors affiliated with FIND (MR, SK) was
426 supported with funding provided by the German Ministry for Education and Research through KfW
427 Development Bank. For the purposes of open access the authors have applied a CC-BY public
428 copyright to any author accepted manuscript arising from this submission.

429 References

430

- 431 1. World Health Organization. Global tuberculosis report 2019. 283 p.
- 432 2. Nguyen TBP, Nguyen TA, Luu BK, Le TTO, Nguyen VS, Nguyen KC, et al. A comparison of digital
433 chest radiography and Xpert MTB/RIF in active case finding for tuberculosis. *The International*
434 *Journal of Tuberculosis and Lung Disease*. 2020 Sep 1;24(9):934–40.
- 435 3. Madhani F, Maniar RA, Burfat A, Ahmed M, Farooq S, Sabir A, et al. Automated chest
436 radiography and mass systematic screening for tuberculosis. *The International Journal of*
437 *Tuberculosis and Lung Disease*. 2020 Jul 1;24(7):665–73.
- 438 4. Onozaki I, Law I, Sismanidis C, Zignol M, Glaziou P, Floyd K. National tuberculosis prevalence
439 surveys in Asia, 1990–2012: an overview of results and lessons learned. *Tropical Medicine &*
440 *International Health*. 2015 Sep 7;20(9):1128–45.
- 441 5. Law I, Floyd K, Abukaraig EAB, Addo KK, Adetifa I, Alebachew Z, et al. National tuberculosis
442 prevalence surveys in Africa, 2008–2016: an overview of results and lessons learned. *Tropical*
443 *Medicine & International Health*. 2020 Nov 12;25(11):1308–27.
- 444 6. Habib SS, Asad Zaidi SM, Jamal WZ, Azeemi KS, Khan S, Khowaja S, et al. Gender-based
445 differences in community-wide screening for pulmonary tuberculosis in Karachi, Pakistan: an
446 observational study of 311 732 individuals undergoing screening. *Thorax*. 2022
447 Mar;77(3):298–9.
- 448 7. Frascella B, Richards AS, Sossen B, Emery JC, Odone A, Law I, et al. Subclinical Tuberculosis
449 Disease-A Review and Analysis of Prevalence Surveys to Inform Definitions, Burden,
450 Associations, and Screening Methodology. In: *Clinical Infectious Diseases*. Oxford University
451 Press; 2021. p. E830–41.
- 452 8. World Health Organisation. Module 2: Screening WHO operational handbook on tuberculosis
453 Systematic screening for tuberculosis disease.
- 454 9. The Stop TB Partnership. THE GLOBAL PLAN TO END TB. 2023.
- 455 10. Zhen Qin Z, Ahmed S, Shahnewaz Sarker M, Paul K, Shafiq A, Adel S, et al. Can artificial
456 intelligence (AI) be used to accurately detect tuberculosis (TB) from chest x-ray? A
457 multiplatform evaluation of five AI products used for TB screening in a high TB-burden
458 setting.
- 459 11. Qin ZZ, Sander MS, Rai B, Titahong CN, Sudrungrot S, Laah SN, et al. Using artificial
460 intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of
461 the diagnostic accuracy of three deep learning systems. *Sci Rep*. 2019 Dec 1;9(1).
- 462 12. Khan FA, Majidulla A, Tavaziva G, Nazish A, Abidi SK, Benedetti A, et al. Chest x-ray analysis
463 with deep learning-based software as a triage test for pulmonary tuberculosis: a prospective
464 study of diagnostic accuracy for culture-confirmed disease. *Lancet Digit Health*. 2020 Nov
465 1;2(11):e573–81.
- 466 13. Harris M, Qi A, Jeagal L, Torabi N, Menzies D, Korobitsyn A, et al. A systematic review of the
467 diagnostic accuracy of artificial intelligence-based computer programs to analyze chest x-rays
468 for pulmonary tuberculosis. *PLoS One*. 2019 Sep 1;14(9).
- 469 14. Khan FA, Pande T, Tessema B, Song R, Benedetti A, Pai M, et al. Computer-aided reading of
470 tuberculosis chest radiography: Moving the research agenda forward to inform policy. Vol.
471 50, *European Respiratory Journal*. European Respiratory Society; 2017.
- 472 15. Soares TR, Oliveira RD de, Liu YE, Santos A da S, Santos PCP dos, Monte LRS, et al. Evaluation
473 of chest X-ray with automated interpretation algorithms for mass tuberculosis screening in
474 prisons: A cross-sectional study. *The Lancet Regional Health - Americas*. 2023 Jan;17:100388.
- 475 16. Sossen B, Richards AS, Heinsohn T, Frascella B, Balzarini F, Oradini-Alacreu A, et al. The
476 natural history of untreated pulmonary tuberculosis in adults: a systematic review and meta-
477 analysis. *Lancet Respir Med*. 2023 Apr;11(4):367–79.
- 478 17. Esmail H, Coussens AK, Thienemann F, Sossen B, Mukasa SL, Warwick J, et al. High resolution
479 imaging and five-year tuberculosis contact outcomes. *medRxiv*. 2023 Jul 3;

- 480 18. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an
481 updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015 Oct
482 28;h5527.
- 483 19. World Health Organization. WHO consolidated guidelines on tuberculosis. Module 2:
484 screening – systematic screening for tuberculosis disease. Geneva: World Health
485 Organization; 2021. Licence: CC BY-NC-SA 3.0 IGO. 2021;
- 486 20. Boule A, Heekes A, Tiffin N, Smith M, Mutemaringa T, Zinyakatira N, et al. Data Centre
487 Profile: The Provincial Health Data Centre of the Western Cape Province, South Africa. *Int J*
488 *Popul Data Sci*. 2019 Nov 20;4(2):1143.
- 489 21. South African Department of Health. National Tuberculosis Management Guidelines, 2014.
490 [cited 2022 Nov 11]; Available from:
491 [https://www.tbonline.info/media/uploads/documents/national_tuberculosis_management_](https://www.tbonline.info/media/uploads/documents/national_tuberculosis_management_guidelines_%282014%29.pdf)
492 [guidelines_%282014%29.pdf](https://www.tbonline.info/media/uploads/documents/national_tuberculosis_management_guidelines_%282014%29.pdf)
- 493 22. World Health Organization. WHO operational handbook on tuberculosis. Module 2: screening
494 - systematic screening for tuberculosis disease. Geneva: World Health Organization; 2021.
495 Licence: CC BY-NC-SA 3.0 IGO. 2021.
- 496 23. Melendez J, Philipsen RHHM, Chanda-Kapata P, Sunkutu V, Kapata N, van Ginneken B.
497 Automatic versus human reading of chest X-rays in the Zambia National Tuberculosis
498 Prevalence Survey. *The International Journal of Tuberculosis and Lung Disease*. 2017 Aug
499 1;21(8):880–6.
- 500 24. Melendez J, Sánchez CI, Philipsen RHHM, Maduskar P, Dawson R, Theron G, et al. An
501 automated tuberculosis screening strategy combining X-ray-based computer-aided detection
502 and clinical information. *Sci Rep*. 2016;6:25265.
- 503 25. Koesoemadinata RC, Kranzer K, Livia R, Susilawati N, Annisa J, Soetedjo NNM, et al.
504 Computer-assisted chest radiography reading for tuberculosis screening in people living with
505 diabetes mellitus. *The International Journal of Tuberculosis and Lung Disease*. 2018 Sep
506 1;22(9):1088–94.

507
508
509
510
511
512
513
514
515
516
517
518
519
520

152 index cases

983 household contacts

511 screened with CXR/ 3x sputum

483 included in analysis (247 underwent blood sampling)

307 attended active follow up between 24-36 months

TB status:
7 routine prevalent
16 enhanced prevalent:
38 incident (18 via PHDC)
61 TB cases in total

Excluded prior to screening:

271 Age <18
161 Did not consent to participate
32 Unable to be followed up
8 On TB treatment

Excluded at screening:

25 pregnant
2 CAD reading error
1 CXR incorrect format (not digital)

Screening strategy: TB symptoms

– positive symptom screen.

- Confirmatory test eligible : 51
- Routine confirmatory test positive : 4
- Enhanced confirmatory test positive : 4
- Total cases confirmed: 8

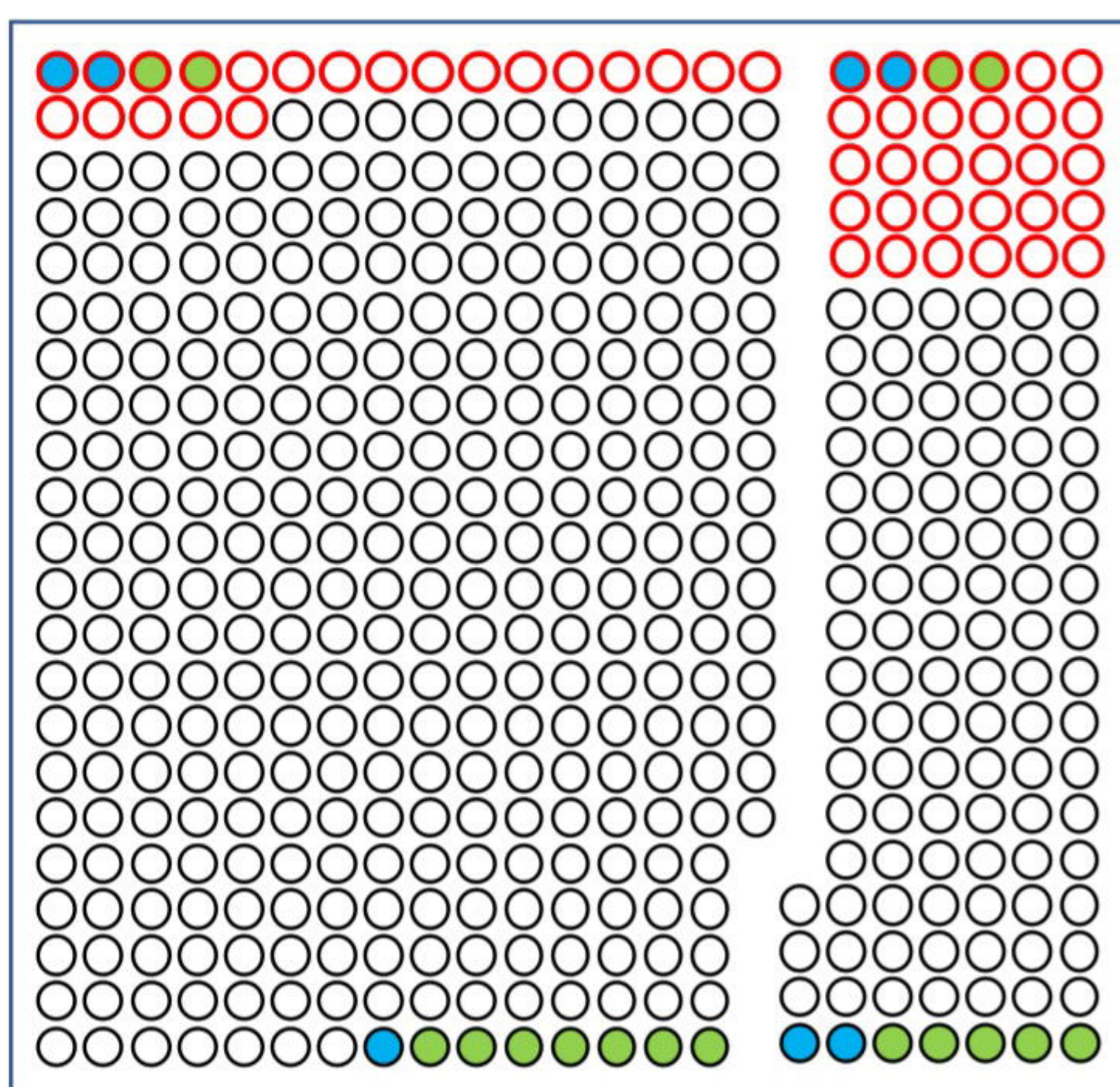
Screening strategy: Abnormal CXR consistent with TB

– determined by human reader.

- Confirmatory test eligible : 74
- Routine confirmatory test positive : 5
- Enhanced confirmatory test positive : 9
- Total cases confirmed: 14

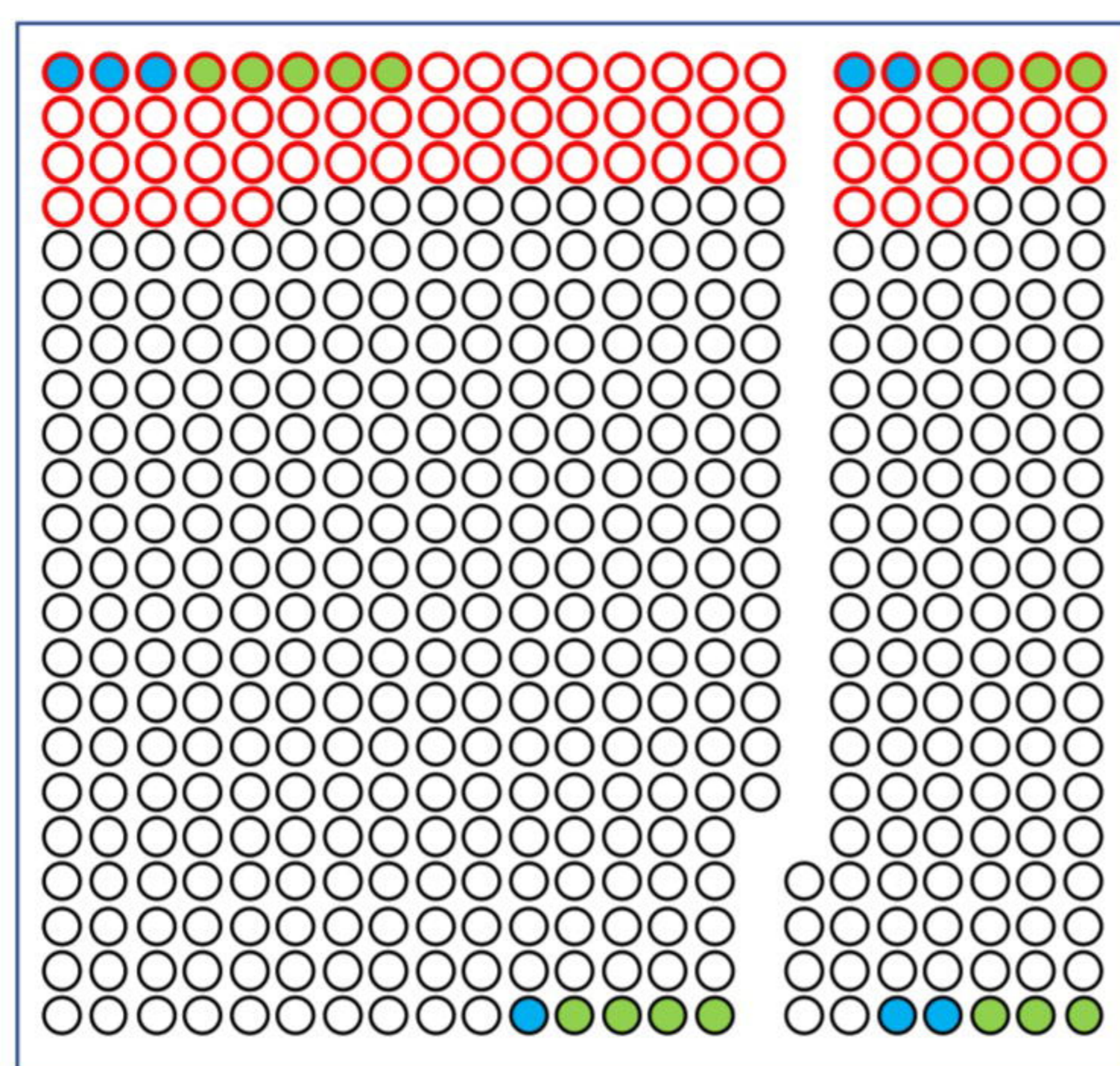
Screening strategy: All – up to 3* sputa for Xpert MTB/RIF and culture for all participants regardless of symptoms or baseline CXR findings.

- Confirmatory test eligible: 483
- Routine confirmatory test positive: 7
- Enhanced confirmatory test positive : 16
- Total cases confirmed: 23



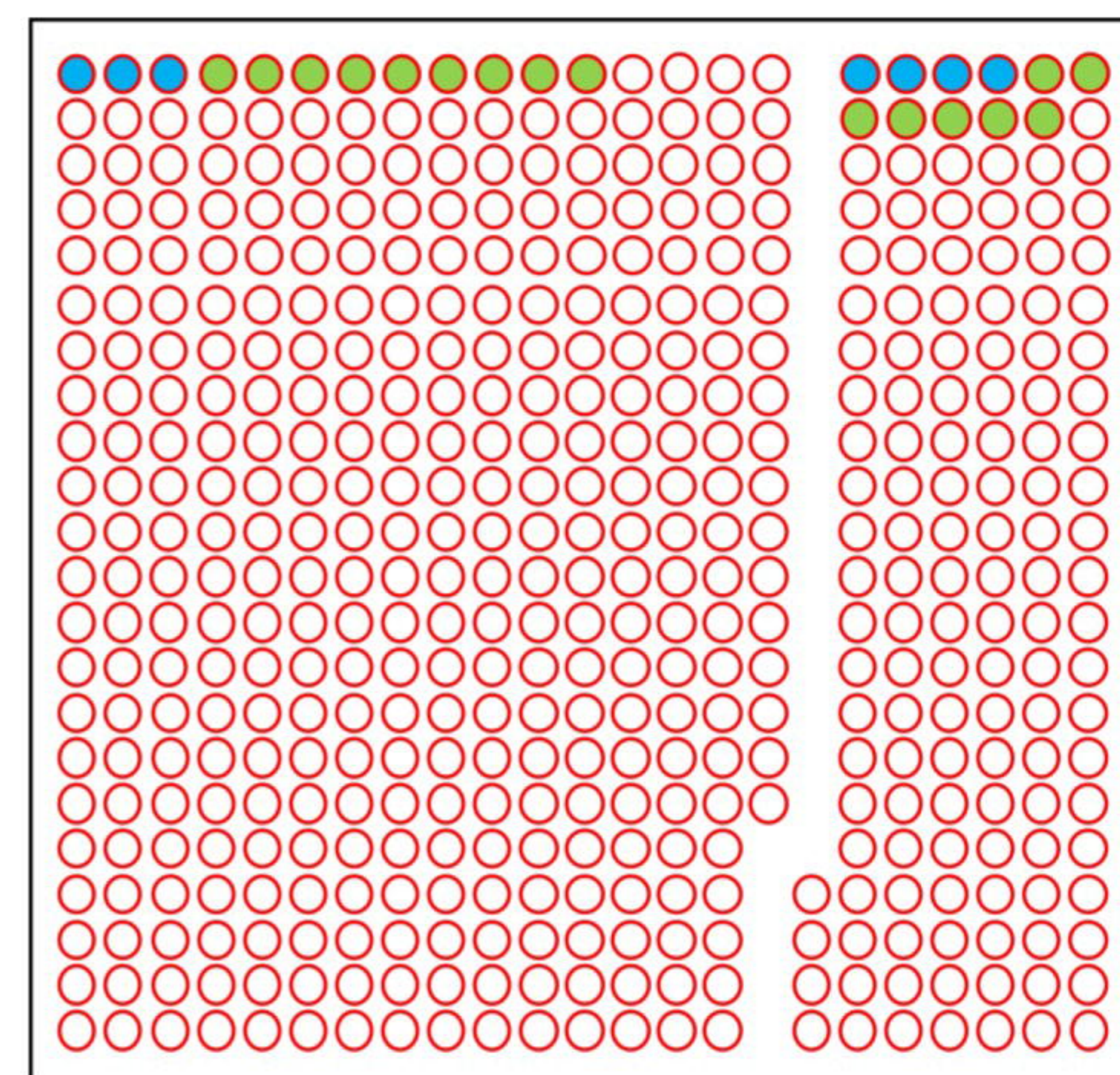
No HIV

HIV



No HIV

HIV



No HIV

HIV

- Initial screening test positive
- Routine prevalent TB: Single spontaneous Xpert positive
- Enhanced prevalent TB: Other sputum sample positive

Screening strategy:

TB symptoms

Abnormal CXR consistent with TB

Routinely diagnosed prevalent TB cases:

Sensitivity

0.57 (0.18 to 0.90)

0.71 (0.29 to 0.96)

Specificity

0.90 (0.87 to 0.93)

0.85 (0.82 to 0.89)

All prevalent TB cases:

Sensitivity

0.35 (0.16 to 0.57)

0.61 (0.39 to 0.80)

Specificity

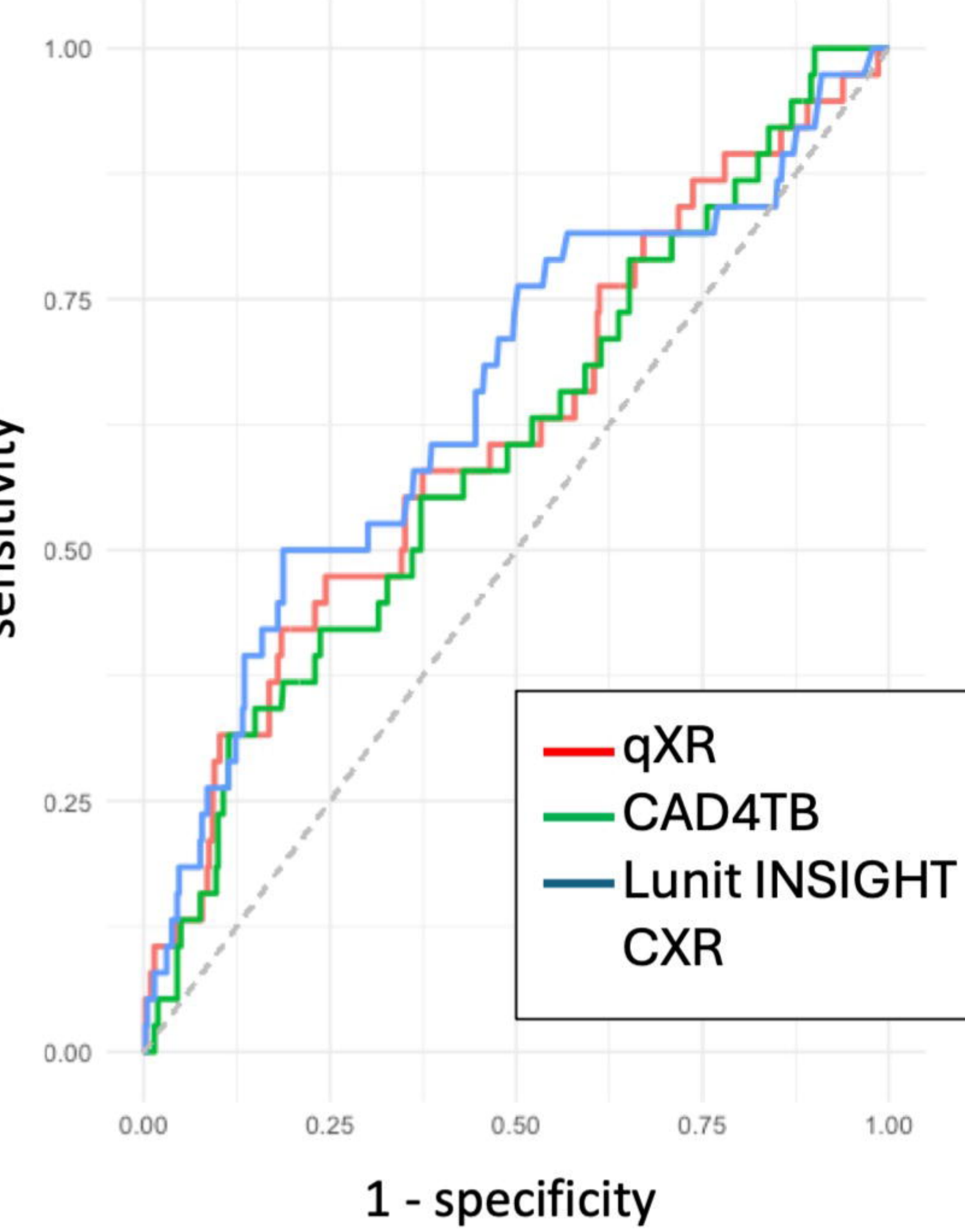
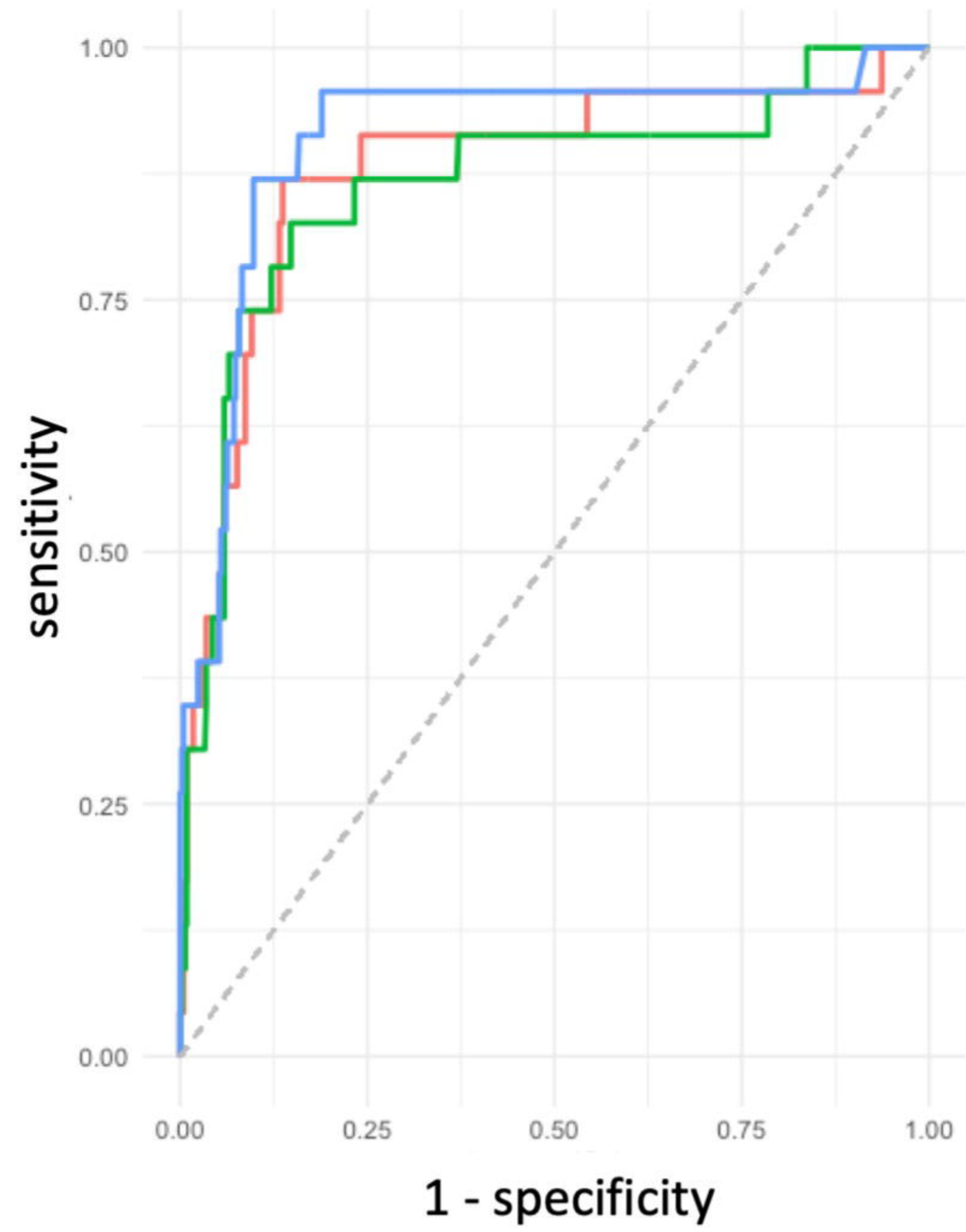
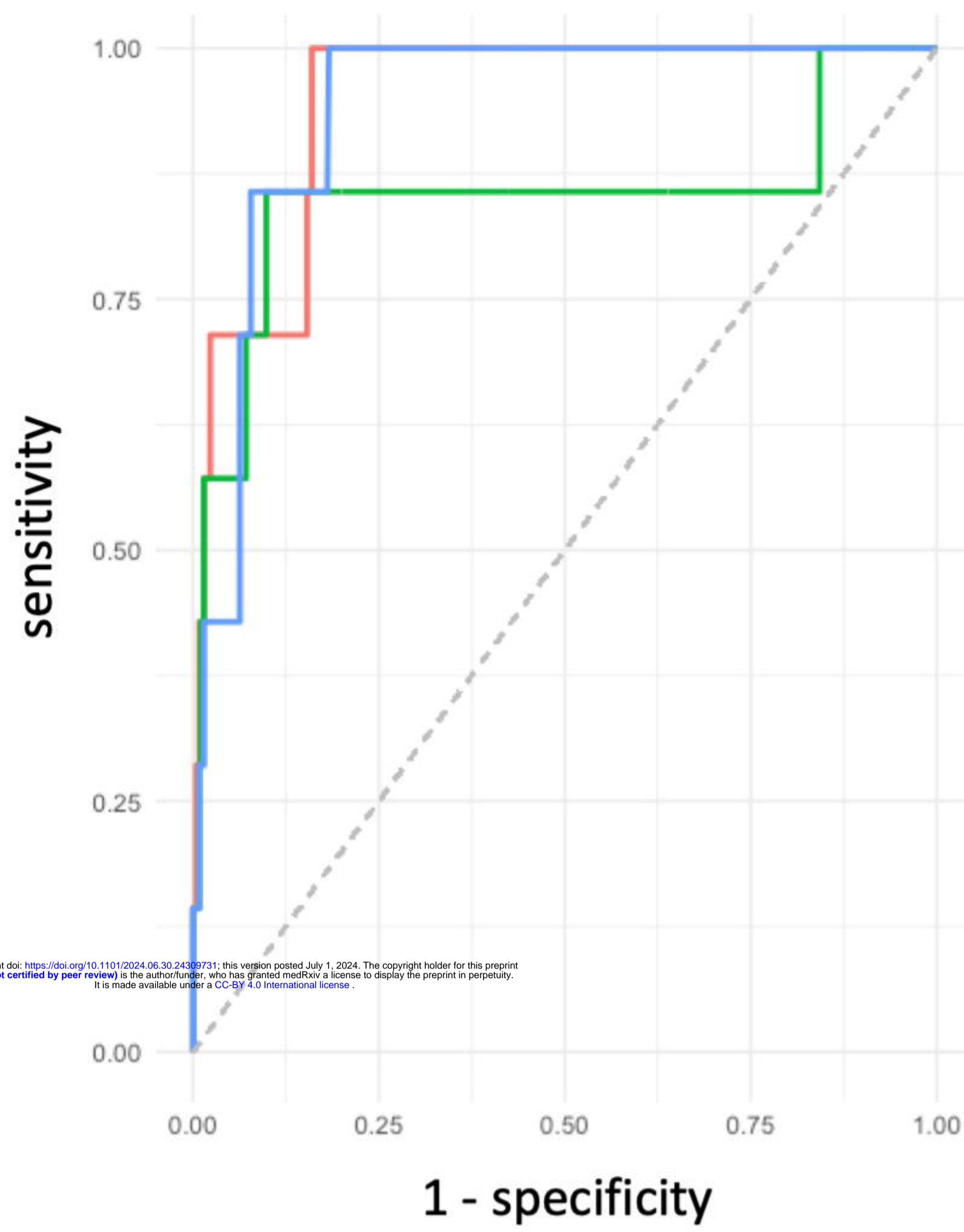
0.91 (0.88 to 0.93)

0.87 (0.84 to 0.90)

Routine prevalent TB

All prevalent TB

Incident TB



— qXR
— CAD4TB
— Lunit INSIGHT
— CXR

AUC ROC: Routine prevalent TB

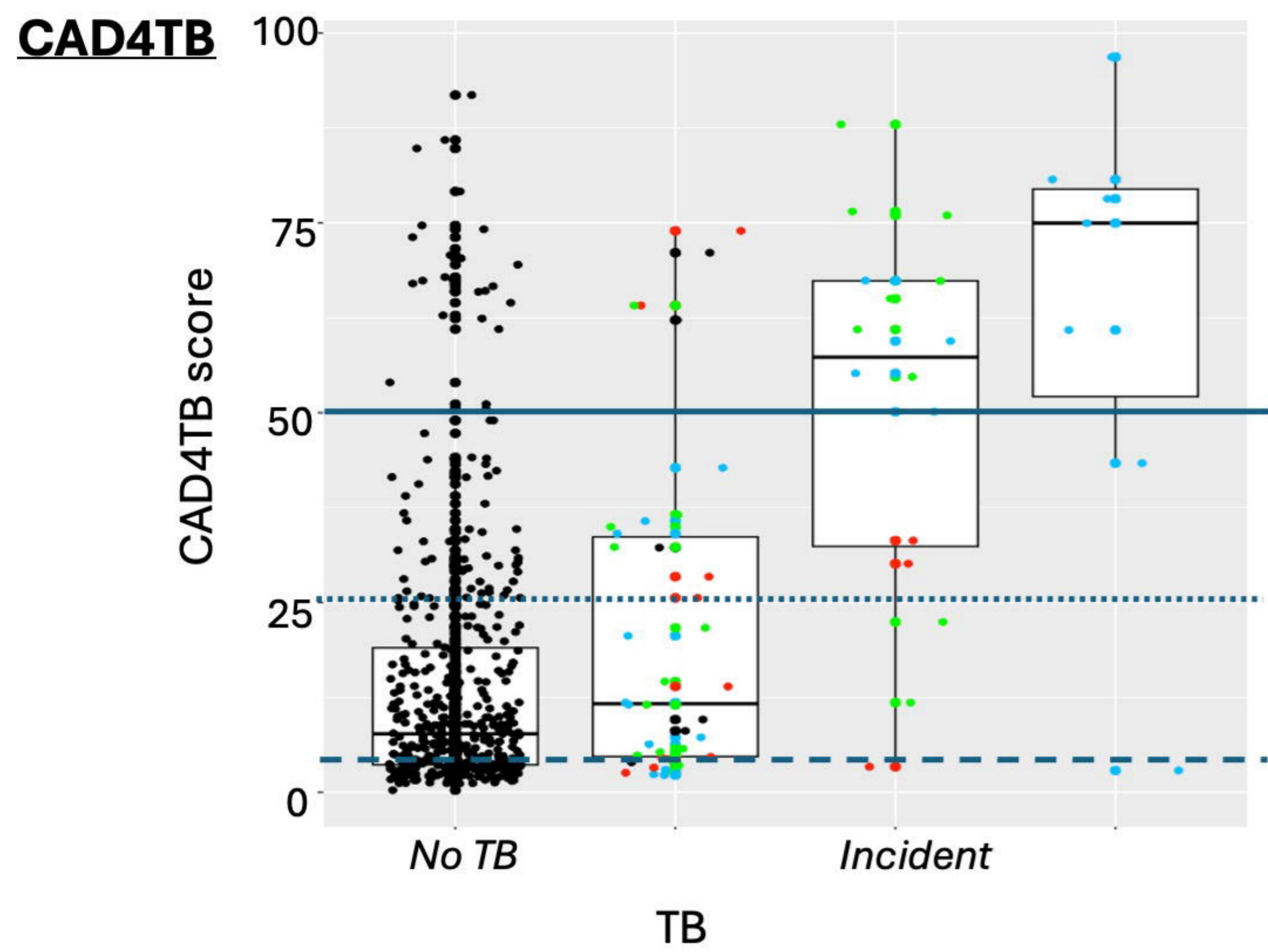
AUC ROC: All prevalent TB

AUC ROC: Incident TB

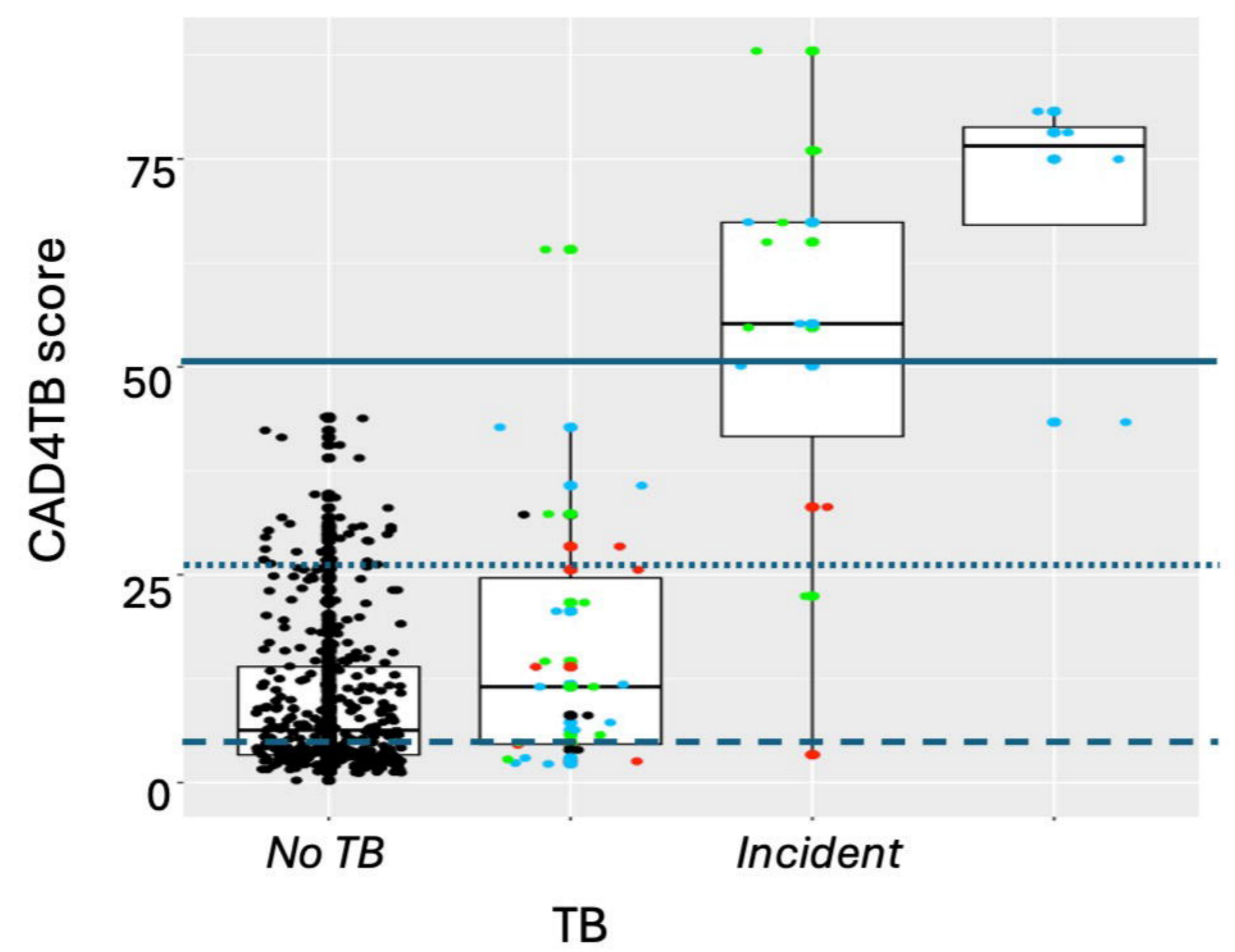
CAD4TB	0.85 (0.62-1)	0.87 (0.77-0.96)	0.60 (0.50-0.70)
qXR	0.95 (0.89-1)	0.88 (0.79-0.97)	0.62 (0.52-0.72)
Lunit	0.94 (0.89-0.99)	0.91 (0.83-0.99)	0.65 (0.55-0.75)

medRxiv preprint doi: <https://doi.org/10.1101/2024.06.30.24309731>; this version posted July 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

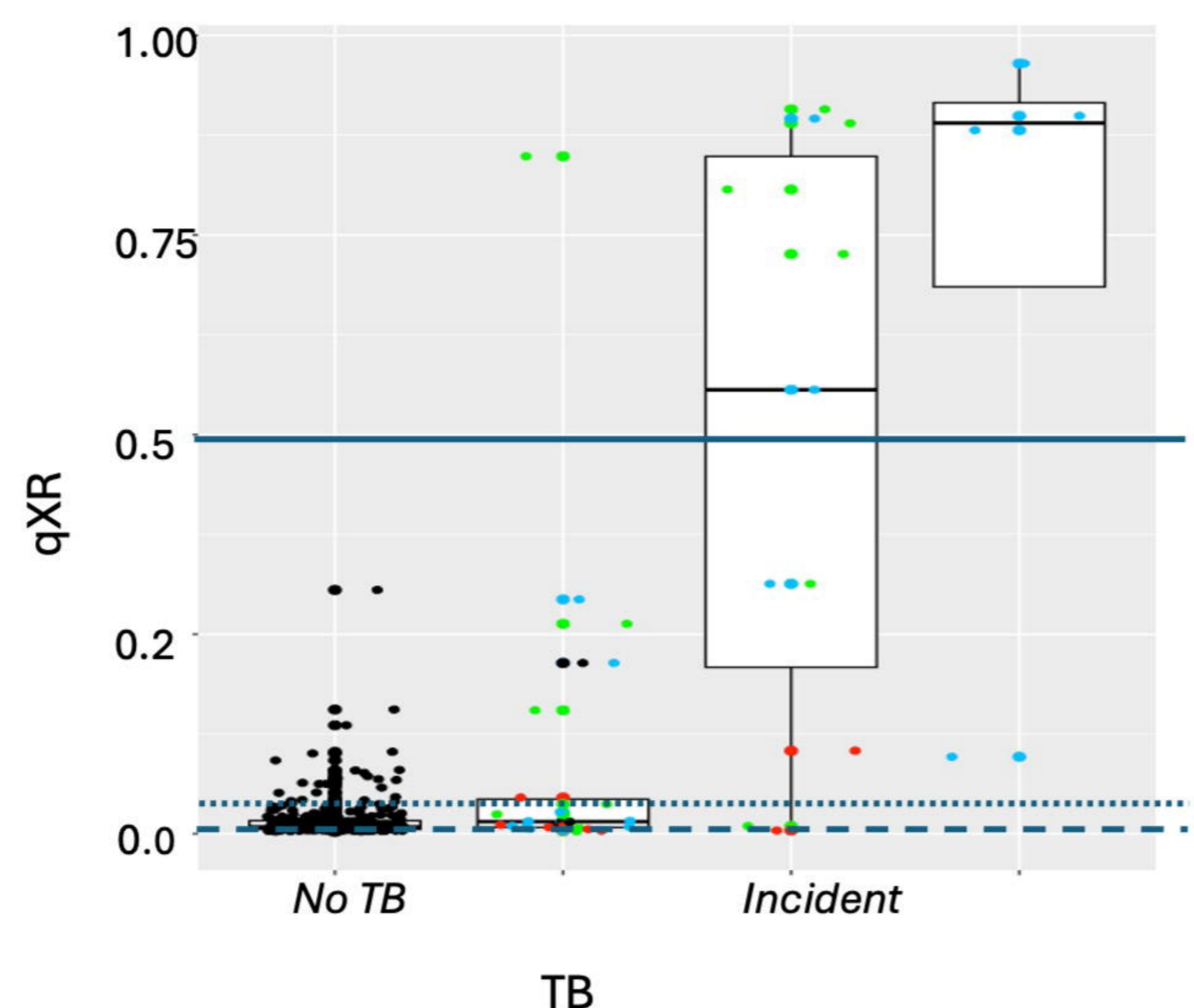
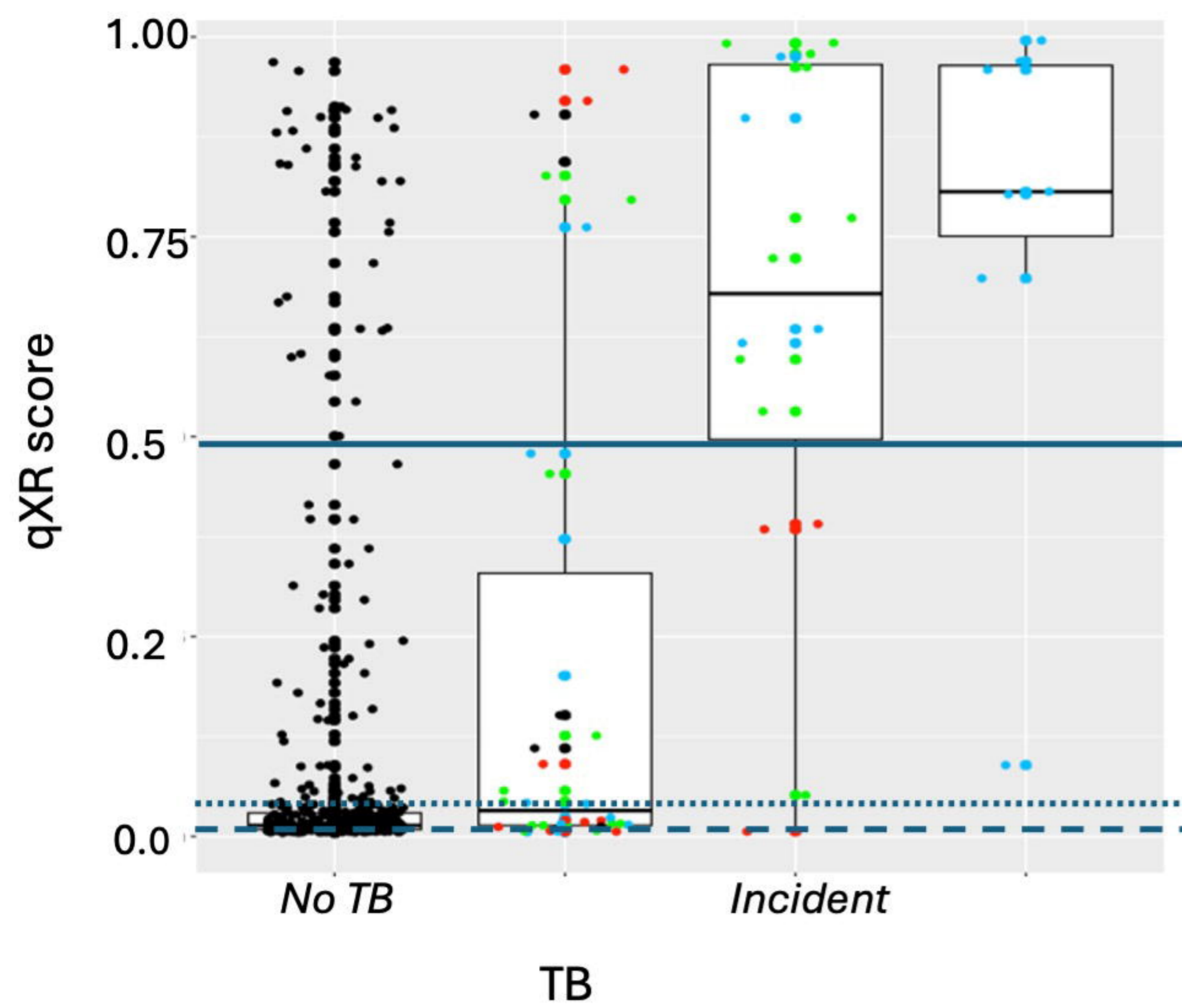
(A) All participants (n=483)



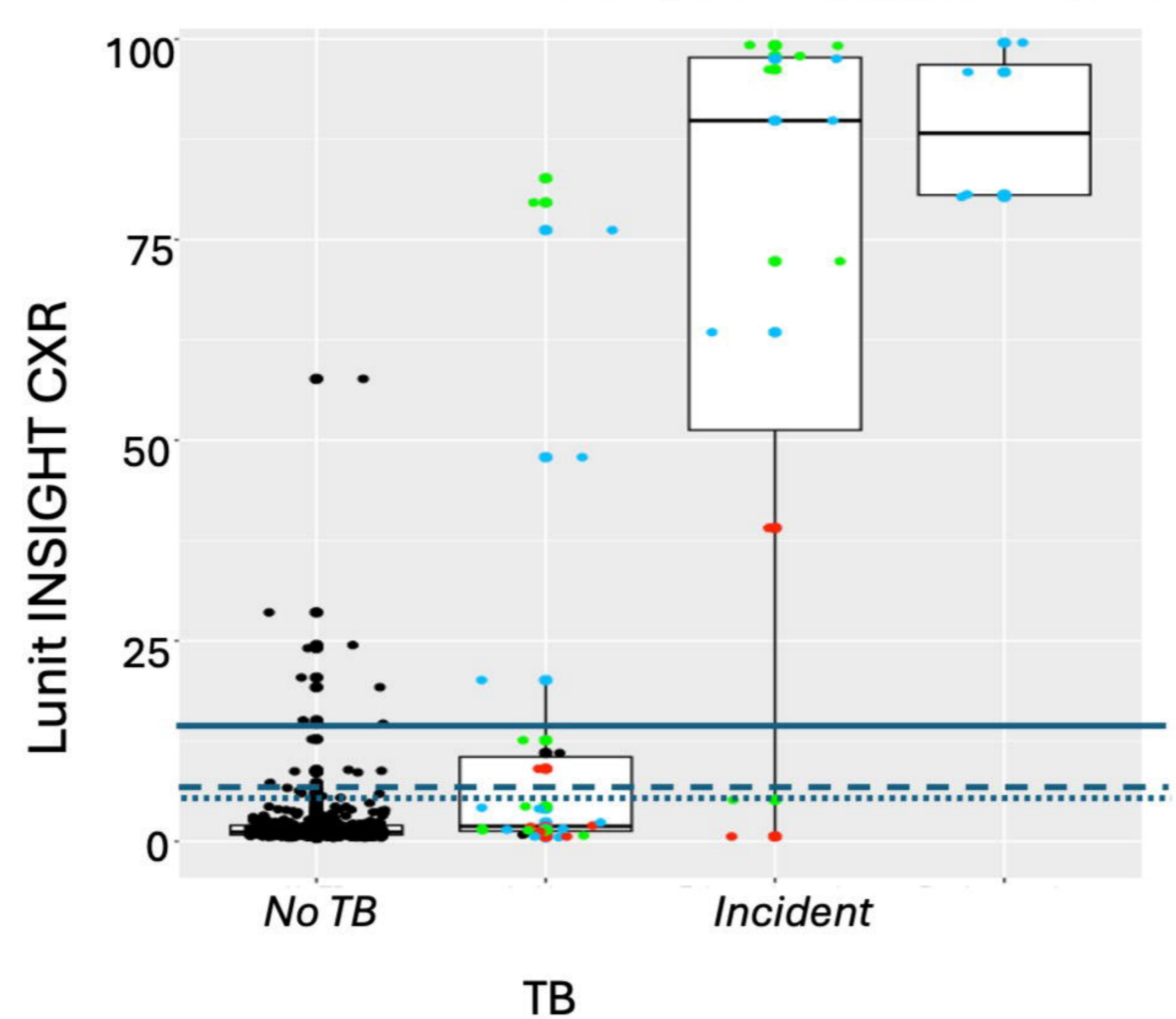
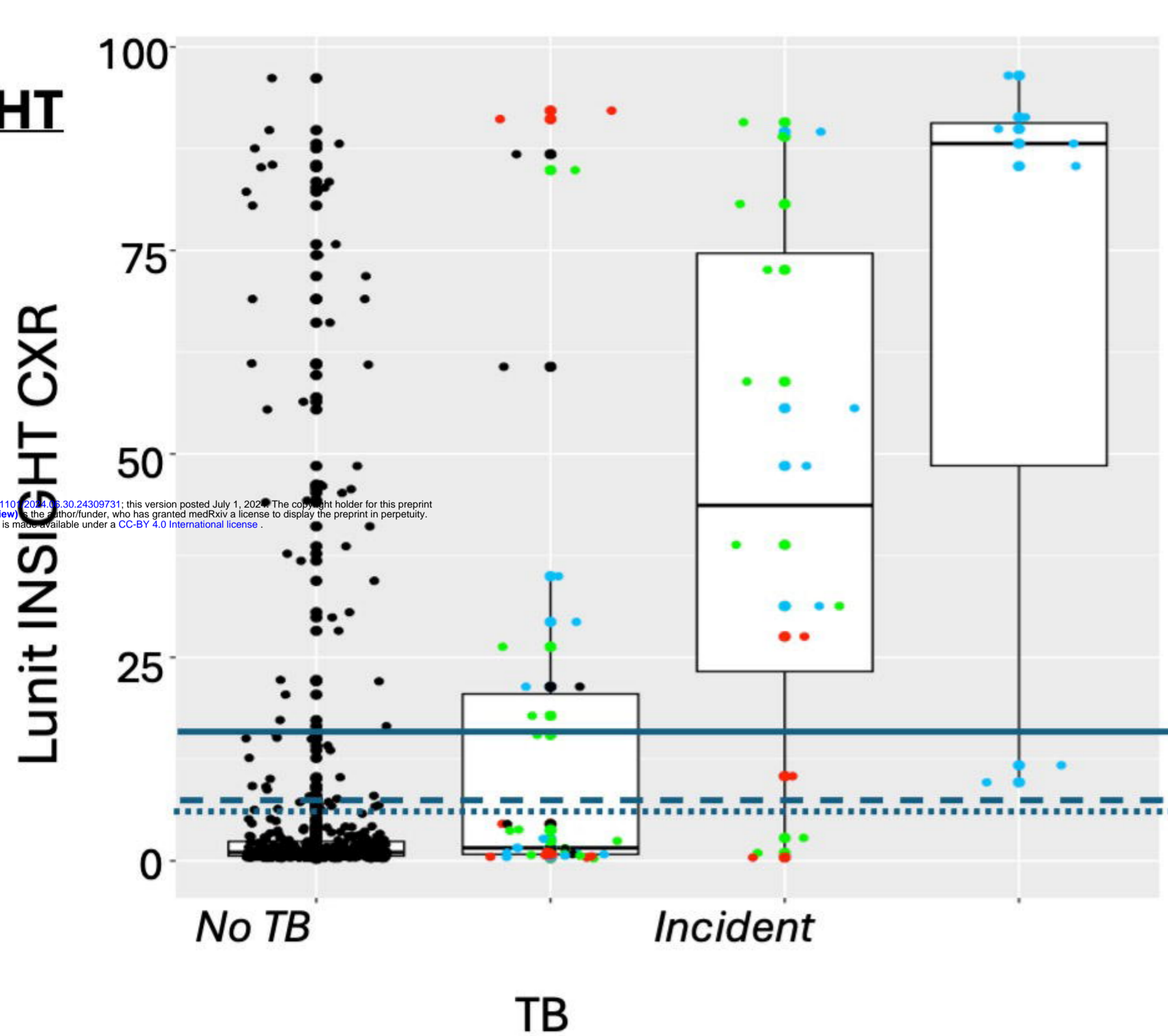
(B) Participants with no previous TB diagnosis (n=374)



qX



Lunit INSIGHT CXR

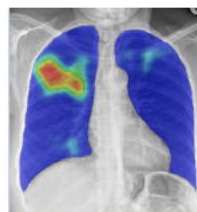


medRxiv preprint doi: <https://doi.org/10.1101/2024.06.24.24289721>; this version posted July 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

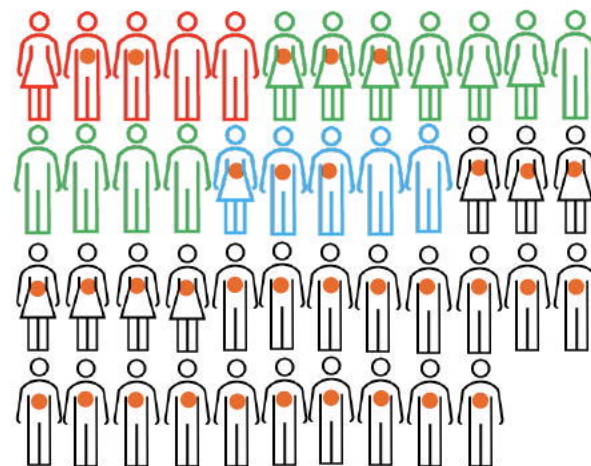
- Negative Xpert MTB/RIF, negative culture
- Positive culture, negative Xpert MTB/RIF
- Positive Xpert MTB/RIF, negative culture
- Positive Xpert MTB/RIF, positive culture

- Manufacturer or commonly used threshold
- ⋯ Threshold derived using the WHO TPP optimal specificity for a TB triage test (0.8)
- - - Threshold derived using the WHO TPP optimal sensitivity for a TB triage test (0.95)

CAD4TB



Proportion above threshold: **10%** (46/483)

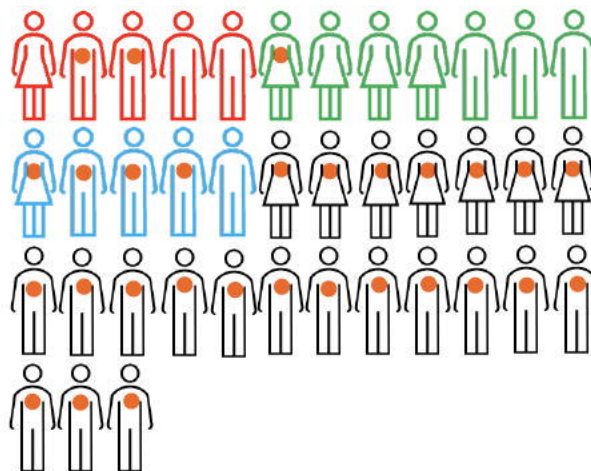


Routine prevalent: 5/46 (11%)
Enhanced prevalent: 11/46 (24%)
Incident case: 5/46 (11%)
History of previous TB: 34/46 (74%)

qXR

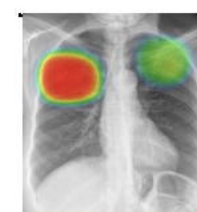


Proportion above threshold: **8%** (39/483)

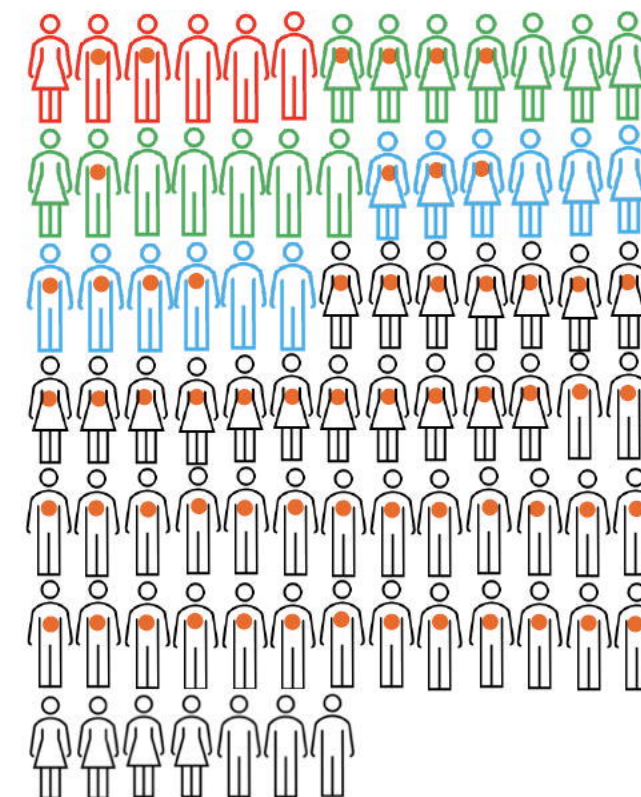


Routine prevalent: 5/39 (13%)
Enhanced prevalent: 7/39 (18%)
Incident case: 5/39 (13%)
History of previous TB: 29/39 (74%)

Lunit INSIGHT CXR



Proportion above threshold: **18%** (85/483)



Routine prevalent: 6/85 (7%)
Enhanced prevalent: 14/85 (16%)
Incident case: 12/85 (14%)
History of previous TB: 60/85 (71%)

- Routine prevalent case (M, F)
- Enhanced prevalent case (M, F)
- Incident case (M, F)
- No TB (M, F)
- History of previous TB

CAD software

CAD4TB

qXR

Lunit INSIGHT CXR

Routinely diagnosed prevalent TB:

Sensitivity
Specificity

0.71 (0.29-0.96)
0.91 (0.89-0.94)

0.71 (0.29-0.96)
0.93 (0.90-0.95)

0.86 (0.42-1)
0.83 (0.80-0.87)

All prevalent TB:

Sensitivity
Specificity

0.70 (0.47-0.87)
0.93 (0.91-0.96)

0.57 (0.34-0.77)
0.94 (0.92-0.96)

0.87 (0.66-0.97)
0.86 (0.82-0.89)