Title

A Comparison of CXR-CAD Software to Radiologists in Identifying COVID-19 in Individuals Evaluated for Sars CoV 2 Infection in Malawi and Zambia.

Short Title

Performance of CXR-CAD in identifying COVID-19 in Malawi and Zambia

Authors

Sam Linsen[1], Aurélie Kamoun[1], Andrews Gunda[2], Tamara Mwenifumbo[2], Chancy Chavula[2], Lindiwe Nchimunya[2], Yucheng Tsai[2], Namwaka Mulenga[2], Godfrey Kadewele[3], Eunice Nahache[3], Veronica Sunkutu[4], Dr. Jane Shawa[5], Rigveda Kadam[1], Matt Arentz[1],

Affiliations

1. FIND
2. Clinton Health Access Initiative
3. Malawi Ministry of Health
4. University Teaching Hospital, Zambia Ministry of Health
5. Levy Mwanawasa Medical University, Zambia Ministry of Health

Abstract

Introduction: AI based software, including computer aided detection software for chest radiographs (CXR-CAD), was developed during the pandemic to improve COVID-19 case finding and triage. In high burden TB countries, the use of highly portable CXR and computer aided detection software has been adopted more broadly to improve the screening and triage of individuals for TB, but there is little evidence in these settings regarding COVID-19 CAD performance.

Methods: We performed a multicenter, retrospective cross-over study evaluating CXRs from individuals at risk for COVID-19. We evaluated performance of CAD software and radiologists in comparison to COVID-19 laboratory results in 671 individuals evaluated for COVID-19 at sites in Zambia and Malawi between January 2021 and June 2022. All CXRs were interpreted by an expert radiologist and two commercially available COVID-19 CXR-CAD software.

Results: Radiologists interpreted CXRs for COVID-19 with a sensitivity of 73% (95% CI: 69%-76%) and specificity of 49% (95% CI: 40%-58%). One CAD software (CAD2) showed performance in diagnosing COVID-19 that was comparable to that of radiologists, (AUC-ROC of 0.70 (95% CI: 0.65-0.75)), while a second (CAD1) showed inferior performance (AUC-ROC of 0.57 (95% CI: 0.52-0.63)). Agreement between CAD software and radiologists was moderate for diagnosing COVID-19, and very good agreement in differentiating normal and abnormal CXRs in this high prevalent population.

Conclusions: The study highlights the potential of CXR-CAD as a tool to support effective triage of individuals in Malawi and Zambia during the pandemic, particularly for distinguishing normal from abnormal CXRs. These findings suggest that while current AI-based diagnostics like CXR-CAD show promise, their effectiveness varies significantly. In order to better prepare for future pandemics, there is a need for representative training data to optimize performance in key populations, and ongoing data collection to maintain diagnostic accuracy, especially as new disease strains emerge.

Author Summary

During the COVID-19 pandemic, AI-based software was developed to help identify and manage cases, including software that assists in reading chest X-rays (CXR-CAD). This technology has also been used in high tuberculosis (TB) burden countries to screen and manage TB cases. However, there's limited information on how well these tools work for COVID-19 in these settings. This study examined chest X-rays from people at risk for COVID-19 in Zambia and Malawi to evaluate the performance of CXR-CAD software against expert radiologists and laboratory COVID-19 tests. The research included X-rays from 671 participants, reviewed by two AI software programs and radiologists.

The results showed that radiologists had a sensitivity of 73% and specificity of 49% in detecting COVID-19. One AI software (CAD2) performed similarly to radiologists, while another (CAD1) performed worse. The agreement between the AI software and radiologists varied, but both were good at distinguishing between normal and abnormal X-rays.

The study suggests that while AI tools like CXR-CAD show potential, their effectiveness can vary. To improve these tools for future pandemics, more representative training data and continuous data collection are necessary.

Introduction

The COVID-19 pandemic, caused by the coronavirus SARS-CoV-2, has had devastating consequences for healthcare systems worldwide. By the end of 2023, there were over 7.7 million reported deaths, and over 18 million estimated deaths globally. (1) A breakdown in global coordination regarding testing, vaccination, and allocation of resources has been described as a major factor involved in the failure of this global response. (2) Limited access to adequate testing and treatment during the pandemic was a consequence of this breakdown, and likely contributed to deaths due to COVID-19 and a number of other communicable diseases. (3, 4)

Use of medical imaging to help diagnose COVID-19 and differentiate it from other respiratory conditions remains a challenge given the overlapping clinical and radiological manifestations of respiratory illnesses. Chest X-rays (CXRs) have been a valuable tool in the evaluation and diagnosis of respiratory diseases for decades. While not as sensitive as computed tomography (CT) scans, CXRs are more widely available, less costly, less infrastructure intensive and can be performed with minimal exposure risk to healthcare providers. Newer, highly portable digital CXR devices have been deployed in many high burden TB settings, improving access to medical imaging outside of regional hospitals, and CXRs may aid in the identification of individuals with communicable respiratory diseases including those at greater risk for worsening. (5),(6, 7) Such interventions may also be beneficial when future respiratory pandemics arise. Yet, the interpretation of CXR findings, especially in a setting where the caseload of patients with respiratory symptoms is high, requires novel strategies given variation in access to radiologists and expert readers. (8)

Leveraging artificial intelligence (AI) to interpret CXRs (computer aided detection, or CXR-CAD) showed promise early in the COVID-19 pandemic. In research settings, use of COVID-19 algorithms with CXR CAD reported very high performance. CXR-CAD software have been deployed in many high burden TB settings and many CXR-CAD developers added COVID-19 specific algorithms during the pandemic. (9-11) However, in many instances, training data for these software leveraged datasets drawn from populations outside of the African continent, and it has been unclear how COVID-19 specific CAD algorithms could perform in populations in the region.

In addition to uncertainty on performance, as the pandemic evolved, policy shifted away from medical imaging (and CXR) as a first step in the diagnostics evaluation of an individual at risk for COVID-19. This has limited the understanding of the benefit of CXR-CAD for this use. (12) As a consequence of a decrease in reimbursement for COVID-19 specific screening and diagnostic tools, many commercially available COVID-19 CAD algorithms have been removed from the market, while the potential benefit for use has never been fully defined.

Understanding how AI based algorithms for COVID-19 that interpret CXRs can benefit *at risk* populations in Africa can better inform their potential for use in this region for future respiratory pandemics, especially in settings where access to radiologists may be limited. Malawi and Zambia are two African countries with innovative digital health strategies which

consider governance of digital technology to ensure local benefit. (13, 14) In this study, we explore the potential benefits, challenges, and applications of CXR-CAD systems specifically for use in diagnosing COVID-19 and CXR abnormalities in populations at risk for COVID-19 from Zambia and Malawi. In doing so, we aim to elucidate the role that AI specific diagnostics for CXR could play in the diagnosis and disease management COVID-19 and future respiratory pandemics.

Materials and Methods

Study design and participants

We performed a multicenter, retrospective cross-over study evaluating CXRs from individuals at risk for COVID-19 with both CXR-CAD software and radiologist in comparison to WHO approved COVID-19 laboratory testing. Individuals were enrolled from one of two sites in Zambia (the University Teaching Hospital and the Levey Mwanawasa University Teaching Hospital, both in Lusaka, Zambia) and one of five sites in Malawi (Mzuzu Central Hospital, Kamuzu Central Hospital, Queen Elizabeth Central Hospital, Nsanje District Hospital, and Mwaiwathu Hospital). Included patients were adults evaluated at the enrolling sites from January 1, 2021 to June 1, 2022 who had symptoms consistent with COVID-19 based on a clinical evaluation; and who also had received both a WHO approved Sars CoV2 test as well as a digital chest radiograph within 72 hours of evaluation. Individuals younger than 18 years of age, and those who did not have a result for COVID-19 testing and a CXR digital image accessible in DICOM format were excluded.

Sample size and sampling

For the comparison of sensitivity and specificity between radiologist and CXR CAD diagnostics, a sample size calculation was generated using a previously established formula for comparison of proportions in cross-over designs. (15) The sample size of 500 total participants per country was calculated based on existing data to demonstrate a difference in accuracy of 10% between CXR-CAD systems and human readers at a COVID-19 prevalence of 20%, and was powered to determine a sensitivity of 90% and a specificity of 60% against laboratory reference standards at a COVID prevalence of 20%.

Data Collection

Records from all patients with COVID-19 testing during the study period were reviewed consecutively by researchers at each site to determine eligibility. In individuals meeting inclusion criteria, data were collected and anonymized for evaluation using Open Clinica (Waltham, USA). Additionally, DICOM CXR images from included individuals were obtained by site researchers. Prior to sharing of DICOM images with FIND, identifying information was removed by use of a previously described DICOM anonymizing tool. (16) Patient clinical data and radiologist interpretations were aggregated and duplicate study subjects were excluded prior to analysis.

Reference standards

The COVID-19 laboratory reference standard was defined based on the result of a WHO-approved Sars CoV2 diagnostic test. (17) COVID-19 cases were defined as positive by one or

more diagnostic test results within 72 hours of clinical evaluation. COVID-19 controls were defined as negative by all diagnostic test results performed within 72 hours of clinical evaluation.

Our radiology reference standard (RRS) for COVID-19 drew on expert radiologists (> 5 years experience) from Malawi and Zambia who were blinded to COVID-19 testing results. A single radiologist from the country where the image was acquired independently evaluated CXRs for findings and recorded results as described below. Radiologists were encouraged to use published guidelines on chest radiograph interpretation for COVID-19 as a component of their interpretation. (18)

Expert radiologists characterized CXRs in one of the following 3 categories:

1. CXR pattern consistent with COVID-19
2. CXR pattern abnormal, but findings not consistent with COVID-19
3. CXR pattern normal

For the primary analysis, RRS were considered positive for COVID-19 if radiologist interpretation determined the CXR pattern was consistent with COVID-19 (category #1). The RRS was considered negative for COVID-19 if the interpretation determined the CXR pattern was normal, or was abnormal, but with findings not consistent with COVID-19 (categories #2 and #3).

For the secondary analysis, radiologist evaluations were considered abnormal if interpretation of the CXR was characterized as abnormal and consistent with COVID-19, or abnormal, but with findings non consistent with COVID-19 (categories #1 and #2). Radiologist evaluations were considered normal only for CXRs interpreted with findings consistent with a normal CXR pattern (category #3).

In instances where multiple DICOM CXR images were available within 72 hours of evaluation, the first CXR chronologically was collected for analysis. In instances where there were duplicate DICOM images, the higher resolution/larger file was selected and used for evaluation. If these parameters were identical, one image was selected for analysis and any duplicates were deleted.

Ethics and Privacy Statement

The study received approval through the Clinton Health Access Initiative Institutional Review Board (CHAI IRB), and received in country ethical/IRB approval. Because of this study's retrospective design, informed consent was not obtainable. In all instances data was anonymized on site prior to sharing with FIND for evaluations. Data and images were stored in an encrypted password protected storage, with full data only available to the study PI and relevant members of the FIND data science team.

Change in approach to developer evaluations during the study

Initially, a comparative analysis of multiple COVID-19 CAD algorithms was planned, including multiple developers FIND previously identified with CXR-CAD algorithms for TB. (10, 19) However, as the pandemic evolved, many CXR-CAD developers removed their COVID 19 algorithms from commercial use. As a result, FIND established an agreement with 2 CXR-CAD developers to independently evaluate performance of their COVID-19 CXR-CAD and normal/abnormal CXR-CAD algorithms that had been in commercial use during the height of the pandemic, in this cohort of individuals at risk for COVID-19. However, given the removal of many of these products from the market, a stipulation to this evaluation was that publicly available results would blind specific developer performance in this analysis. Both developers included in this analysis have CXR-CAD products in use for the evaluation of CXRs for TB, and both have received WHO stringent regulatory approval for at least one diagnostic use for CXR-CAD. For this publication, these software developers are referred to as CAD1 and CAD2.

Data analysis

For the primary analysis, both radiologist and COVID-19 algorithms were evaluated in comparison to a WHO approved laboratory reference standard. For CXR-CAD software interpretation, images were first pre-processed as needed to conform to CXR-CAD software developer specifications for DICOM files. Images were then exposed to CXR-CAD software with outputs/probability scores recorded for COVID-19 algorithms and for normal/abnormal algorithms (i.e. determining if any abnormalities are observed), using the FIND validation platform, as has been described elsewhere. (20) Performance metrics (sensitivity and specificity) of CXR-CAD for COVID-19 were compared to radiologist readings using CXR-CAD thresholds set to observed radiologist sensitivity and, separately, to observed radiologist specificity. This analysis mirrors a similar approach that has been described in other studies evaluating CXR-CAD for TB. (16, 21) Subgroup analysis was performed based on country, site, age group, sex, symptoms, diabetes status (if known), and HIV status (if known). Results were presented for CXR-CAD sensitivity and specificity with 95% confidence intervals.

CXR-CAD scores were used to generate receiver operating characteristic (ROC) curves for both COVID scores and any abnormalities scores . (22) The Area Under the ROC curve (AUC-ROC) for each CAD system was calculated against laboratory reference standards using binomial distribution assumptions for the primary analysis.

Radiologist sensitivity and specificity assessments for COVID-19 were calculated in comparison to laboratory reference standards. CXR-CAD software estimates of sensitivity and specificity were then calculated at the threshold produced by the same sensitivity or specificity achieved by the radiologist.

For the secondary analysis, CXR-CAD was evaluated for agreement with a radiologist in differentiating normal and abnormal CXRs, based on the AC1 coefficient. (23) In this secondary analysis, CXR-CAD algorithms for abnormal CXRs were evaluated using the manufacturer suggested threshold.

Results:

Population characteristics and image assessment

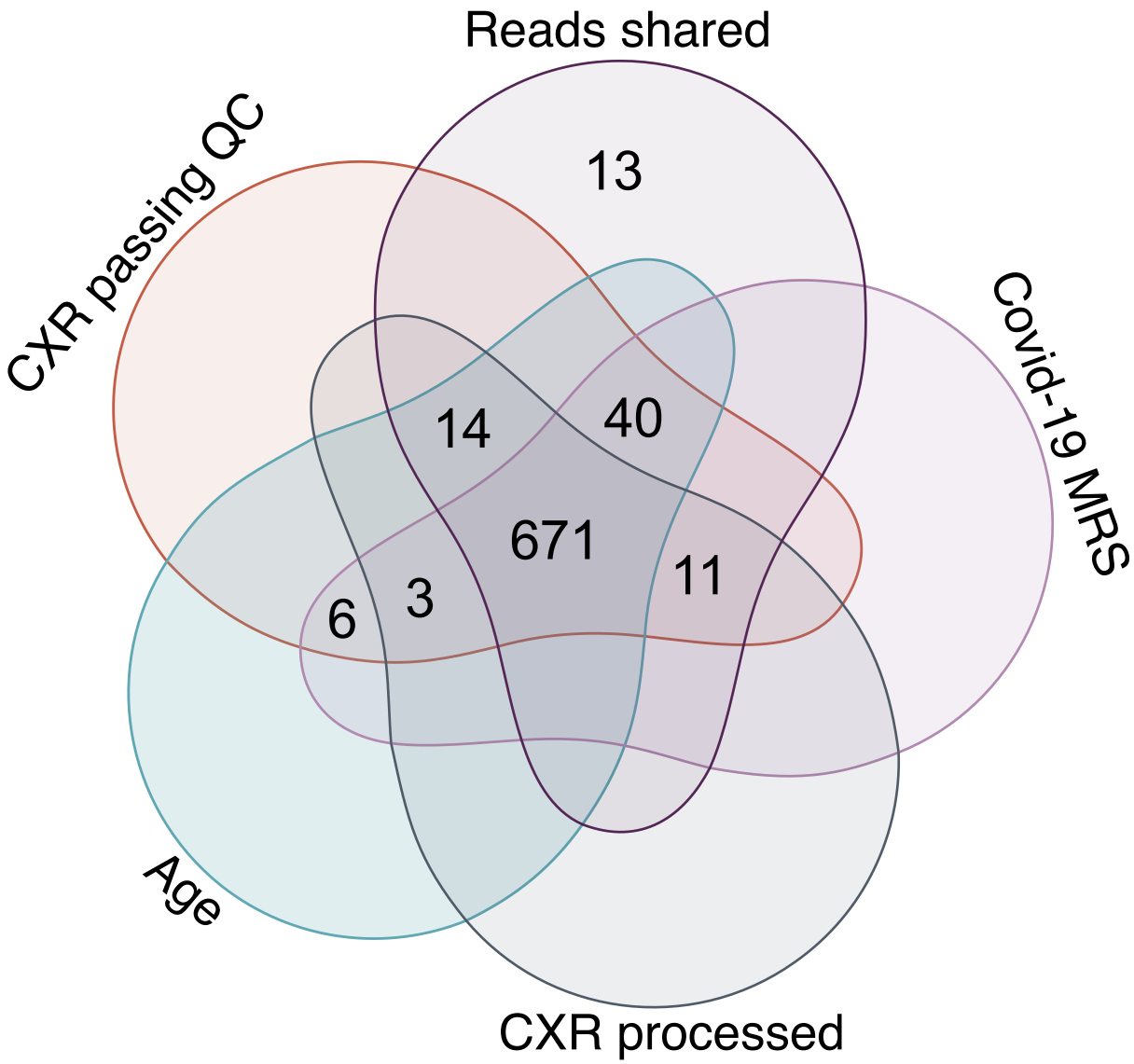


Figure 1 This study, participants were selected based on Reads shared (CXR interpretation by a radiologist); available MRS diagnostic output (Covid-19 MRS); CXRs processed by CAD-1 and CAD-2 (CXR processed); and quality checks (CXR passing QC).

In total 758 CXR images were available for this study. A total of 749 CXR images were read with findings reported by radiologist and 9 were not. Among the shared reads, 13 CXR images had to be excluded from this study as they were not properly de-identified; for 11 patients, age information was missing and these files were excluded; and 14 patient entries did not report a COVID-19 test which was recorded withing 72 hours of evaluation. In the remaining 711 images, there were 40 images which could not be processed by CAD1 or CAD2, and the associated subjects were removed from the evaluation.

This left a total of 671 participant with complete clinical data and appropriate images available for evaluation (89% of the image total) which were included in the comparative covid analysis. A Venn diagram of participants inclusion in the final analyses based on image assessment and processing is shown in **Figure 1.**

The included participants included 269 (40% of the total) from Malawi and 402 (60% of the total) from Zambia. This total included 540 COVID-19 cases and 131 controls, with a considerably higher prevalence of COVID-19 cases than in our pre-study power calculations. Of the total number of cases, 212 (39%) were from Malawi and 328 (61%) were from Zambia. Of the total controls, 57 (44%) were from Malawi and 74 (56%) were from Zambia.

The majority of participants were male (n= 410, 61%) and the average age was 51 years (range 18 to 90 years). In most instances, the test sample was collected via nasopharyngeal swab (N= 474, 71%). A minority of patients were tested HIV positive (n= 77, 11%) or were known to be diabetic (n=86, 13%). A description of the populations by site are described in **table 1.**

Variable	**Total**	**Positive Cases**	**Negative Cases**
All	**671**	**540**	**131**
Study Site ID			
Zambia			
UTH	190	142 (26.3%)	48 (36.6%)
LEVY M. UTH	212	186 (34.4%)	26 (19.8%)
Malawi			
Mzuzu Central Hospital	30	30 (5.6%)	0 (0.0%)
Kamuzu Central Hospital	105	101 (18.7%)	4 (3.1%)
Queen Elizabeth Central Hospital	62	62 (11.5%)	0 (0.0%)
Nsanje District Hospital	8	8 (1.5%)	0 (0.0%)
Mwaiwathu Hospital	64	11 (2.0%)	53 (40.5%)
Age group			
[18-25]	36	30 (5.6%)	6 (4.6%)
[26-35]	87	60 (11.1%)	27 (20.6%)
[36-45]	141	111 (20.6%)	30 (22.9%)
[46-55]	145	118 (21.9%)	27 (20.6%)
[56-65]	128	107 (19.8%)	21 (16.0%)
[>=66]	134	114 (21.1%)	20 (15.3%)
Gender			
Female	260	203 (37.6%)	57 (43.5%)
Male	410	337 (62.4%)	73 (55.7%)

Variable	Total	Positive Cases	Negative Cases
Not recorded	1	0 (0.0%)	1 (0.8%)
Type of COVID-19 test done			
Antigen	154	90 (16.7%)	64 (48.9%)
PCR	516	449 (83.1%)	67 (51.1%)
NA	*1*	*1 (0.2%)*	*0 (0.0%)*
HIV status			
Positive	77	71 (13.1%)	6 (4.6%)
Negative	459	381 (70.6%)	78 (59.5%)
Not recorded	135	88 (16.3%)	47 (35.9%)
Diabetes status			
Positive	86	70 (13.0%)	16 (12.2%)
Negative	450	382 (70.7%)	68 (51.9%)
Not recorded	135	88 (16.3%)	47 (35.9%)
Symptoms			
Present	586	479 (88.7%)	107 (81.7%)
Absent	35	22 (4.1%)	13 (9.9%)
Not recorded	50	39 (7.2%)	11 (8.4%)
Sampling method			
Bronch-alveolar lavage	1	1 (0.2%)	0 (0.0%)
Nasal swab	186	173 (32.0%)	13 (9.9%)
Nasopharyngeal swab	474	356 (65.9%)	118 (90.1%)
Not recorded	3	3 (0.6%)	0 (0.0%)
Throat swab	7	7 (1.3%)	0 (0.0%)

Table 1 Demographic and clinical characteristics from includes participants

Radiologist Reference Standard (RRS) for COVID and radiologist interpretation other abnormalities

Radiologists interpreted 460 images (72.8% of cases and 51.2 % of controls, based on microbiologic reference standard or MRS) as having findings consistent with COVID-19. Of the remaining COVID cases positive by MRS, radiologists interpreted the majority (89, 16.5%) as abnormal but with findings not consistent with COVID-19 pneumonia. Details of radiologist interpretation of CXRs for COVID-19 are shown **in Table 2**.

Human interpretation of CXR	Total	MRS Positive	MRS Negative

COVID-19 Positive	CXR pattern consistent with COVID-19	460	393 (72.8%)	67 (51.2%)
COVID-19 Negative	1. CXR pattern abnormal, but findings not consistent with COVID-19 (non COVID-19)	116	89 (16.5%)	27 (20.6%)
	CXR pattern normal (no abnormalities)	95	58 (10.7%)	37 (28.2%)

Table 2 radiologist interpretation of CXRs in comparison to laboratory testing results. Positive and Negative cases are defined based on the Sars CoV – 2 Microbiologic Reference Standard (MRS).

In comparison to the COVID-19 laboratory testing (MRS), blinded expert radiologists had a pooled specificity of 49% (95% CI 40% - 58%) and a pooled sensitivity of 73% (95% CI 69%-76%) for diagnosis of COVID-19. However, specificity varied significantly by country. In Zambia, the observed specificity was 30% (95% CI 20%-41%) and sensitivity was 73% (95% CI 68%-78%), while in Malawi, radiologist observed specificity was 74% (95% CI 60%-84%) and sensitivity was 73% (95% CI 66%-79%) for COVID-19. Radiologist diagnostic accuracy is shown in **table 3, figure 2 and figure 3**

	N	**RRS Pos**	**RRS Neg**	**Sensitivity (95% CI)**	**Specificity (95% CI)**
Overall	671	460	211	**73% (69-76%)**	**49% (40-58%)**
Zambia	402	291	111	**73% (68-78%)**	**30% (20-41%)**
Malawi	269	169	100	**73% (66-79%)**	**74% (60-84%)**

Table3: Radiologist accuracy in identifying COVID-19 on CXR in comparison to laboratory test results.

Software interpretation of CXR images for COVID-19

Two commercially available CAD software were included in this analysis (deemed CAD1 and CAD2). CAD software interpretation of CXRs was compared to COVID-19 laboratory testing and radiologist performance. The diagnostic outcome of these 2 CAD software is reflected by a score that is generated from a CXR image, and then defined as positive or negative according to a set threshold.

The observed area under the receiver operating curve (AUC-ROC) for CAD1 was 0.57 (95% CI 0.52-0.63) and for CAD2 was 0.70 (95% CI 0.65-0.75). Pooled estimates for CAD1 were inferior to our RRS for COVID-19, when software thresholds were set at the radiologist observed

sensitivity and specificity (**Table 4**). CAD2 had an observed software performance that was comparable to the observed radiologist performance.

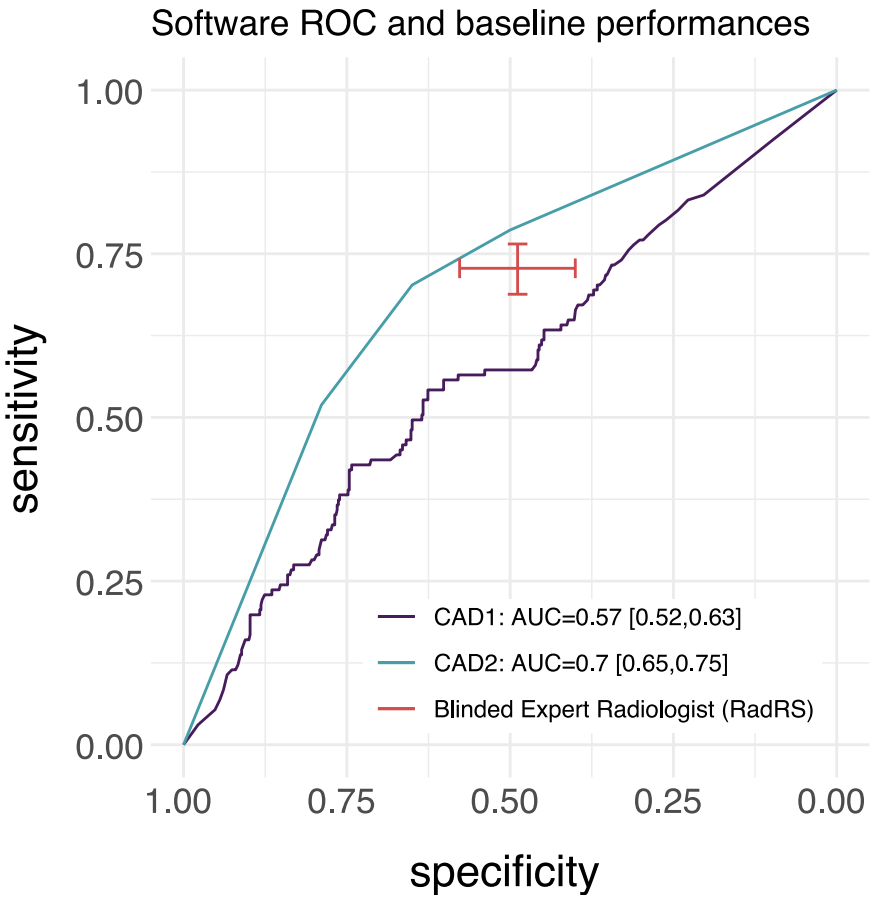


Figure 2 Aggregate performance and 95% confidence intervals of radiologist and software CAD1 and CAD2 at identifying Covid-19 in comparison to baseline molecular testing. AUC: Area under Curve, RadRS: Radiologist reference standard for COVID-19.

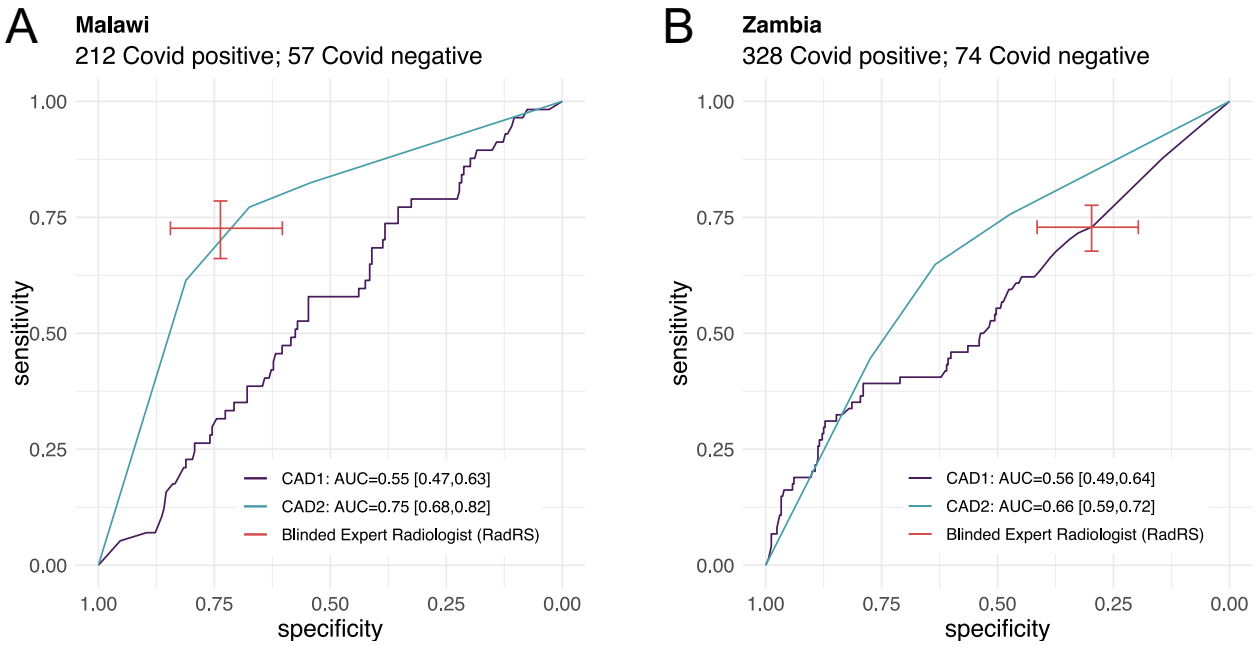


Figure 3: Performance of radiologist and software CAD1 and CAD2 at identifying Covid-19 in comparison to baseline molecular testing. A Performance in the Malawi cohort; B performance in the Zambia cohort

Although CAD1 had a similar performance in Malawi (AUC 0.55, 95% CI 0.47, 0.63) and in Zambia (AUC 0.56, 95% CI 0.49, 0.64), given the lower observed specificity in Zambia, performance of CAD1 was not inferior to RRS in this subgroup (Figure 3). CAD2 demonstrated an AUC of 0.75 (95% CI of 0.68, 0.82) in Malawi and 0.66 (95% CI 0.59, 0.72) in Zambia and was non inferior in performance to a RRS at set specificities in CXRs from both countries.

We evaluated the agreement between CAD software and radiologists at vendor recommended thresholds, and found that agreement was very low for CAD1, with only 60% agreement regarding COVID-19 cases and controls, having an AC1 coefficient value of 0.28. Agreement was moderate for CAD2 (75%) with a AC1 of 0.53. (Appendix table 1a and 1b)

		CAD1			CAD2		
Description	**Fixed Value**	**Th**	**Sens (95% CI)**	**Spec (95% CI)**	**Th**	**Sens (95% CI)**	**Spec (95% CI)**
Fixed Threshold							
Vendor recommended	Th	0.55	47% [38%-55%]	66% [62%-70%]	1.5	70% [62%-78%]	65% [61%-69%]
Fixed Sensitivity							
Radiologist performance	Sens	0.99	73% [66%-80%]	35% [31%-39%]	1.5	70% [62%-78%]	65% [61%-69%]
Fixed Specificity							
Radiologist performance	Spec	0.91	61% [53%-69%]	46% [42%-50%]	2.5	79% [71%-85%]	50% [46%-54%]

Table 4 Software performance at a set radiologist observed sensitivity and specificity. Fixed values (threshold, sensitivity, specificity) were mapped to their closest point from the ROC. This may lead to slight differences between the actual fixed value and the corresponding value from the ROC. Th: threshold Sens: Sensitivity Spec: Specificity CI: confidence interval

Given the poor observed performance, we assessed whether a difference in COVID-19 strains could have resulted in a difference in algorithm performance (24). Therefore, we performed a secondary analysis evaluating performance before and after November of 2021 (when the Omicron strain first superseded the Delta strain as the globally dominant variant). As shown in **figure 4**, The performance of CAD1 demonstrated an AUC of 0.55 (95% CI 0.48, 0.61) before the emergency of the Omicron strain and an AUC of 0.64 (95% CI 0.53,0.75) after the emergence of Omicron. CAD2, which demonstrated a non-significant trend towards better overall performance, had an observed AUC of 0.71 (95% CI 0.65-0.77) before the emergence of Omicron, and an observed AUC of 0.65 (95% CI 0.55- 0.75) after this time. Radiologist sensitivity also decreased after emergence of the Omicron strain, and observed CAD2 software performance remained non-inferior to a radiologist as set sensitivity and specificity.

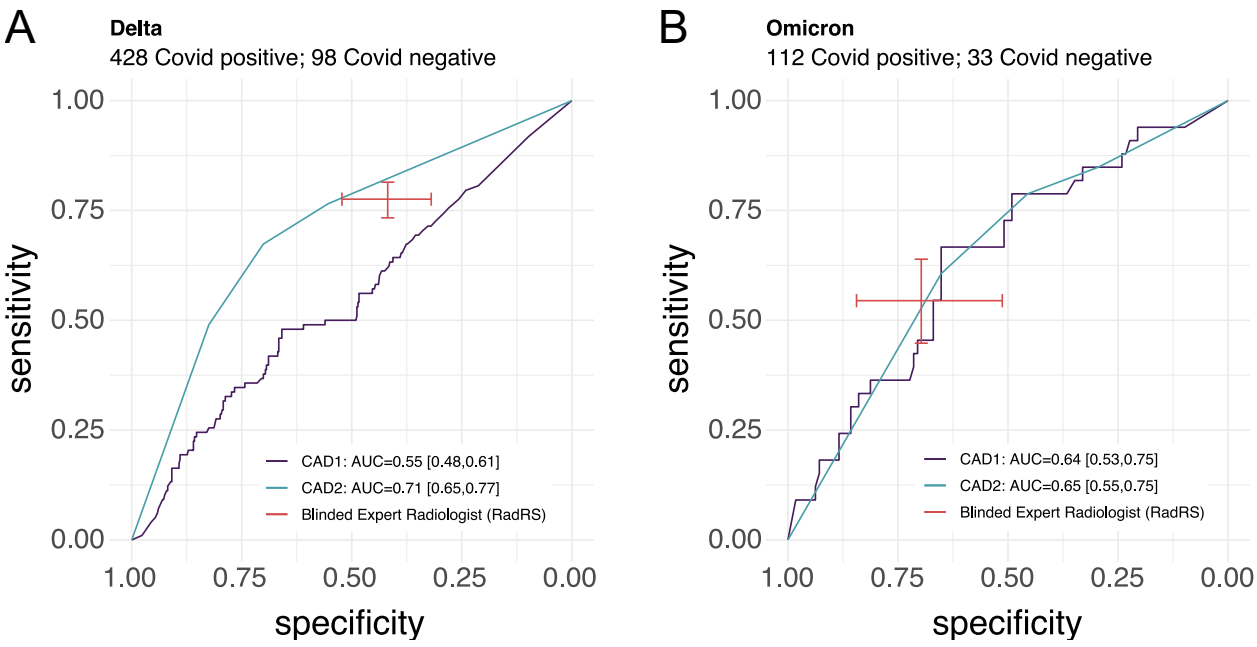


Figure 4: software and radiologist performance at diagnosing COVID-19 before and after the shift to Omicron being the dominant variant.

Software analysis by subgroup

We also evaluated subgroups by age, gender, HIV status, and diabetes status. Given the low numbers in each of these groups, we observed large confidence intervals in AUCs with non significant trends. Further subgroup analysis is (appendix figures 1-4).

Software interpretation of normal vs. abnormal CXR images

For our secondary aim, we evaluated the agreement between CAD software and radiologists at identifying any abnormalities on CXRs. Of the includes CXRs, two lacked a radiologist interpretation of normal vs. abnormal; and 5 additional CXR images read by CAD2 which produced a COVID-19 score did not produce an normal/abnormal score. Therefore, the total number of images included in this secondary analysis were 669 for CAD1 and 664 for CAD2. For this secondary aim, agreement of both CAD software with a radiologist was very good, with 89% agreement and a calculated agreement coefficient of 0.86 (table 5b). In settings where a radiologist identified a CXR as "abnormal" level of agreement with software was better, with 93% of images read by CAD1 and 95% of images read by CAD2 as also identifying CXRs as abnormal. (table 5a)

Table 5

	Radiologist assessment	
	Abnormal	**Normal**
CAD1		
Abnormal	532 (93%)	33 (35%)
Normal	42 (7%)	62 (65%)
CAD2		
Abnormal	544 (95%)	43 (47%)
Normal	29 (5%)	48 (53%)

Table 5a: Agreement of CAD software with radiologists, at manufacturer suggested thresholds for normal vs. abnormal CXRs.

Software	**N**	**Agreement**	**AC1**
CAD1	**669**	**89.00%**	**0.86**
CAD2	**664**	**89.00%**	**0.86**

Table 5b: Percentage agreement of CAD software with radiologists, at manufacturer suggested thresholds for normal vs. abnormal CXRs and the calculated agreement coefficient. AC1: Gwet's Agreement Coefficient.

Discussion

In this study, we independently evaluated performance of CXR-CAD software, in comparison to a radiologist, at identifying COVID-19 or other radiographic abnormalities in individuals evaluated for COVID-19 in two countries in Africa. In doing so, we aim to highlight the potential uses and benefits of CXR-CAD software for COVID-19 and consider the implications for use in future respiratory pandemics. There are a number of key takeaways from this work.

First, we found that performance of CXR-CAD in evaluating CXRs for findings consistent with COVID-19 varied for the two software's evaluated. In comparison to a radiologist against

pooled data from Zambia and Malawi, one software performance (CAD1) was inferior to an expert human radiologist in diagnosing COVID-19, but a second (CAD2) was comparable. It is unclear how different the training data sets were for these two software, or how well they aligned with the populations we evaluated these tools in. As is the case for many commercial products, details related to algorithm development and training are not publicly available. However, these key aspects in software training data are known to be central to appropriate development and deployment of AI based tools. (25, 26) Such issues are critical to address during a rapidly evolving respiratory pandemic, where digital data are limited, and where clinical presentation and radiographic manifestations may change. Our findings suggest that use of CXR-CAD to identify novel respiratory infections can achieve performance comparable to human interpretation. However, it is likely that development of CXR CAD for future pandemics would require more training data from populations in Africa in order to optimize performance in these settings. However, if performance validation is possible, such tools may be considered as an alternative to a human radiologist, or to support through preliminary interpretations in settings in Zambia and Malawi where trained expert readers are scarce.

Second, we found that performance of CAD software and radiologist interpretation varied over time, and a decrease in sensitivity was observed which coincided with emergence of the Omicron wave of COVID-19. This finding highlights the need for novel strategies to re-train and fine tune algorithms when strains with different phenotypes and differing levels of respiratory involvement emerge. (24) In instances where AI based diagnostics are supporting clinical care, including those that use medical imaging, emergence of new strains should be a key consideration for software re-training and version updates.

Third, both software showed strong agreement with a radiologist in differentiating normal and abnormal CXRs in this cohort of individuals at risk for COVID-19. This was particularly true in settings where a radiologist identified no abnormalities on CXR. Algorithms for differentiating normal vs. abnormal CXRs may be an alternative for use early in a pandemic which can identify a subset of individuals with normal CXRs. As a result, when data is scarce and disease specific algorithms will take long to adequately train such models may be valuable in triaging those without respiratory involvement. While such a strategy would likely not have had a large influence on disease transmission, it could rapidly screen or triage a large group of *at risk* individuals to identify those who and do not need close monitoring or further testing, thereby optimize clinical resources and enhance robustness in the diagnostic network.

CXR-CAD software is a promising tool that is currently being used with portable CXR in many settings globally, and for diagnosis of TB and other radiographic findings. Some of these software were adapted for use in identifying COVID-19 during the height of the global pandemic. In these future, consideration of where and how these tools can be used may be a valuable consideration for future pandemics preparedness in Zambia and Malawi, and in other countries in Africa as a component of a digital health strategy. However, as will most AI based diagnostics, key considerations need to be made regarding the collection and use of high quality, representative data for use in the training and testing of these products in the settings where they will be deployed, and in post deployment monitoring. This study demonstrates that some commercially available CAD software could achieve parity with a radiologist in

performance during the pandemic, and that the key considerations mentioned above are central to maintaining benefit.

Limitations

There are a number of limitations to this study. First, the retrospective nature and the inclusion criteria may have selected a subset of individuals which may not be representative of a general population being triaged for COVID-19. Second, there was a significant difference in radiologist performance between Malawi and Zambia. It is unclear whether this represented a difference in approach to CXR interpretation or other factors. Third, although a number of developers had developed algorithms for COVID during the pandemic, many withdrew their software from the market and a comparison of more products could have added more insight into the findings. It is unclear how those other products would have performed against this dataset. Lastly, most CAD developers now offer some strategy for local threshold setting or fine tuning. Local fine tuning on a subset of images would often improves performance for a given setting and population. However, given the independent nature of our assessment, fine tuning on this dataset was not offered, but likely would have improved software performance.

Acknowledgements
This work, including the data collection, protocol development, analysis, and publication was supported with funding provided by German Ministry for Education and Research (BMBF) through KfW. Additionally, we would like to thank Nikhil Jagtiani and Stefano Ongarello at FIND for their support in the study design and execution. Lastly, we would like to thank Qure.ai and VinBrain for their support in this study.

References

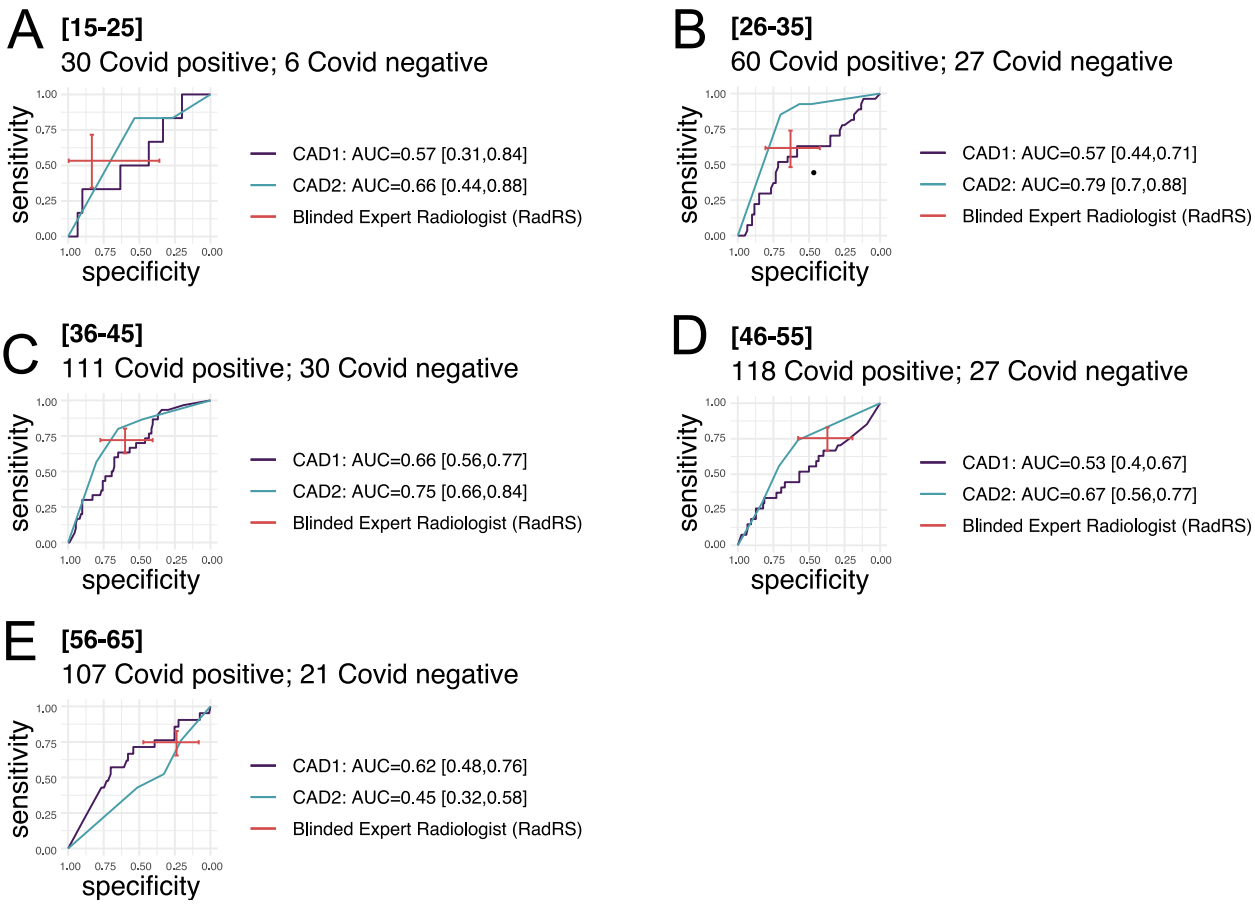
1. IHME. COVID-19 Estimates Downloads [Available from: http://www.healthdata.org/covid/data-downloads.
2. Sachs JD, Karim SSA, Aknin L, Allen J, Brosbol K, Colombo F, et al. The Lancet Commission on lessons for the future from the COVID-19 pandemic. Lancet. 2022;400(10359):1224-80.
3. Collaborators C-EM. Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of COVID-19-related mortality, 2020-21. Lancet. 2022;399(10334):1513-36.
4. One year on, new data show global impact of COVID-19 on TB epidemic is worse than expected [press release]. Geneva: World Health Organization, March 18, 2021 2021.
5. Qin ZZ, Barrett R, Del Mar Castro M, Zaidi S, Codlin AJ, Creswell J, et al. Early user experience and lessons learned using ultra-portable digital X-ray with computer-aided detection (DXR-CAD) products: A qualitative study from the perspective of healthcare providers. PLoS One. 2023;18(2):e0277843.
6. Colman J, Zamfir G, Sheehan F, Berrill M, Saikia S, Saltissi F. Chest radiograph characteristics in COVID-19 infection and their association with survival. Eur J Radiol Open. 2021;8:100360.
7. Jensen CM, Costa JC, Norgaard JC, Zucco AG, Neesgaard B, Niemann CU, et al. Chest x-ray imaging score is associated with severity of COVID-19 pneumonia: the MBrixia score. Sci Rep. 2022;12(1):21019.

8. Frija G, Blazic I, Frush DP, Hierath M, Kawooya M, Donoso-Bach L, et al. How to improve access to medical imaging in low- and middle-income countries ? EClinicalMedicine. 2021;38:101034.
9. J LG, Abraham B, M SS, Nair MS. A computer-aided diagnosis system for the classification of COVID-19 and non-COVID-19 pneumonia on chest X-ray images by integrating CNN with sparse autoencoder and feed forward neural network. Comput Biol Med. 2022;141:105134.
10. Stop TB Partnership F. AI4hlth resource centre on computer-aided detection products forthe diagnosis of tuberculosis [Available from: https://www.ai4hlth.org/.
11. Zouch W, Sagga D, Echtioui A, Khemakhem R, Ghorbel M, Mhiri C, et al. Detection of COVID-19 from CT and Chest X-ray Images Using Deep Learning Models. Ann Biomed Eng. 2022;50(7):825-35.
12. World Health Organization. Use of chest imaging in COVID-19: a rapid advice guide, 11 June 2020 Geneva, CH: World Health Organization; 2020 [Available from: https://www.who.int/publications/i/item/use-of-chest-imaging-in-covid-19.
13. Health GoMMo. National Digital Health Strategy 2020-2025 2020 [Available from: https://www.healthdatacollaborative.org/fileadmin/uploads/hdc/Documents/Country_documents/Malawi/Malawi_Digital_Health_Strategy__20-25.pdf.
14. Republic of Zambia Ministry of Health. Digital Health Strategy 2022-2026 2022 [cited 2024 April 22]. Available from: https://www.moh.gov.zm/wp-content/uploads/filebase/guidelines/presentations/Digital-Health/Digital-Health-strategy-final.pdf.
15. Chow SC JS, H Wang, Y Lokhnygina. Sample size calculations in clinical research, third edition. Statistical Theory and Related Fields. 2017;1:265-6.
16. Gelaw SM, Kik SV, Ruhwald M, Ongarello S, Egzertegegne TS, Gorbacheva O, et al. Diagnostic accuracy of three computer-aided detection systems for detecting pulmonary tuberculosis on chest radiography when used for screening: Analysis of an international, multicenter migrants screening study. PLOS Glob Public Health. 2023;3(7):e0000402.
17. The Global Fund. List of SARS-CoV-2 Diagnostic test kits and equipment eligible for procurement according to Board Decision on Additional Support for Country Responses to COVID-19 2022 [Available from: https://www.theglobalfund.org/media/9629/covid19_diagnosticproducts_list_en.pdf.
18. Hare SS, Tavare AN, Dattani V, Musaddaq B, Beal I, Cleverley J, et al. Validation of the British Society of Thoracic Imaging guidelines for COVID-19 chest radiograph reporting. Clin Radiol. 2020;75(9):710 e9- e14.
19. Qin ZZ, Naheyan T, Ruhwald M, Denkinger CM, Gelaw S, Nash M, et al. A new resource on artificial intelligence powered computer automated detection software products for tuberculosis programmes and implementers. Tuberculosis (Edinb). 2021;127:102049.
20. Hwang EJ, Jeong WG, David PM, Arentz M, Ruhwald M, Yoon SH. AI for Detection of Tuberculosis: Implications for Global Health. Radiol Artif Intell. 2024;6(2):e230327.
21. Codlin AJ, Dao TP, Vo LNQ, Forse RJ, Van Truong V, Dang HM, et al. Independent evaluation of 12 artificial intelligence solutions for the detection of tuberculosis. Sci Rep. 2021;11(1):23895.

22. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.
23. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008;61(Pt 1):29-48.
24. Lee JE, Hwang M, Kim YH, Chung MJ, Sim BH, Jeong WG, et al. SARS-CoV-2 Variants Infection in Relationship to Imaging-based Pneumonia and Clinical Outcomes. Radiology. 2023;306(3):e221795.
25. Group WIF. Clinical evaluation of AI for Health. 2023 March.
26. Organization WH. Regulatory considerations on artificial intelligence for Health. 2023.

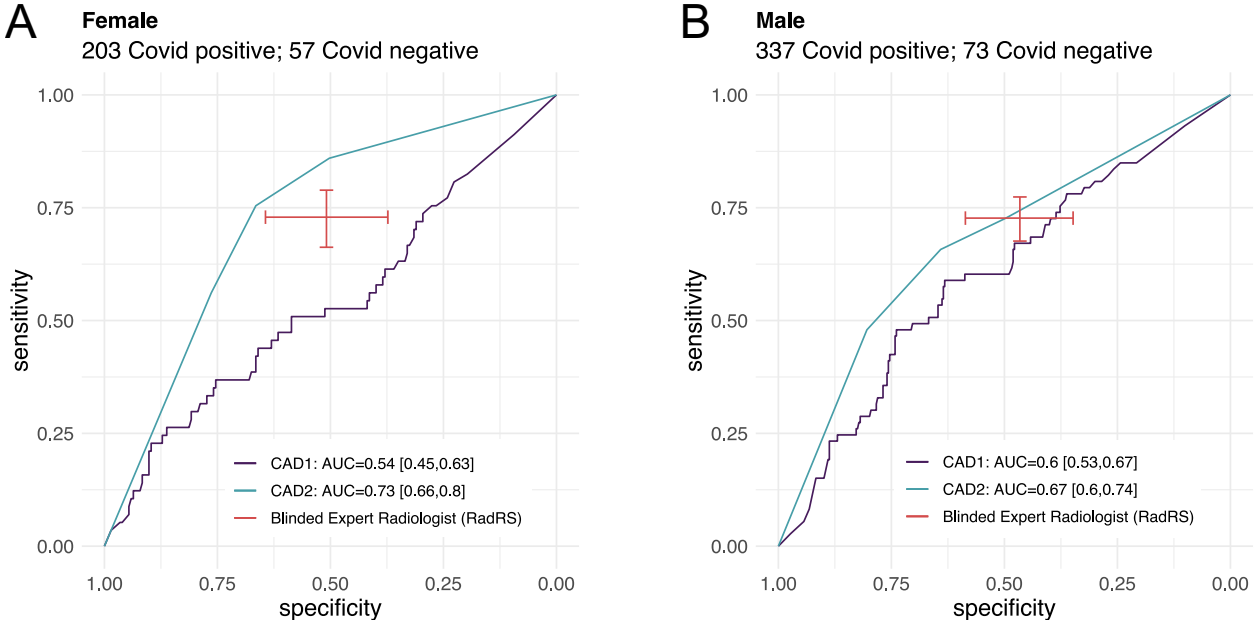
Supporting information

Appendix figure 1



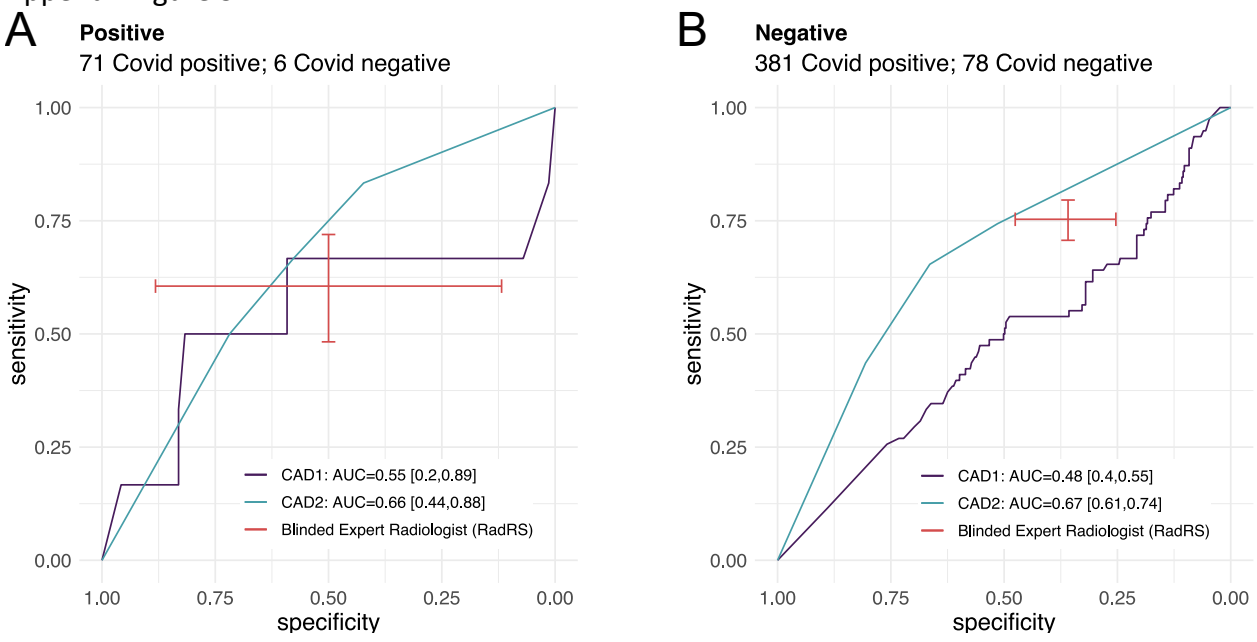
Appendix figure 1: subgroup analysis of performance of radiologist and software CAD1 and CAD2 at identifying COVID-19 in comparison to baseline molecular testing grouped by age.

Appendix figure 2



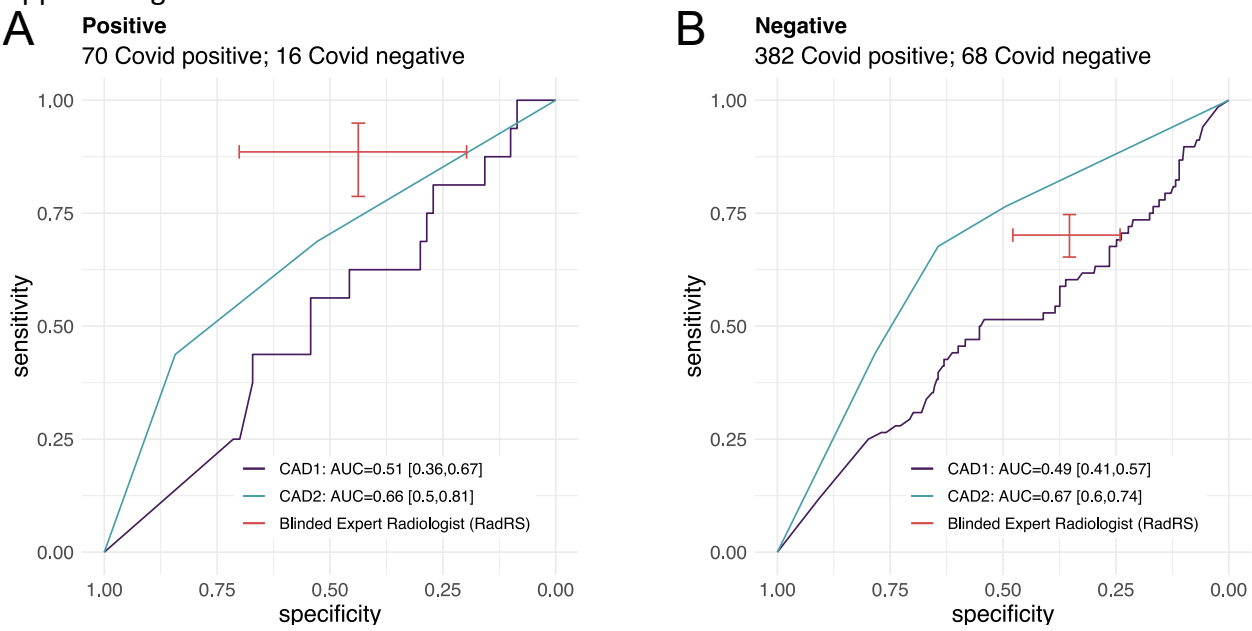
Appendix figure 2: subgroup analysis of performance of radiologist and software CAD1 and CAD2 at identifying COVID-19 in comparison to baseline molecular testing grouped by gender.

Appendix figure 3



Appendix figure 3: subgroup analysis of performance of radiologist and software CAD1 and CAD2 at identifying COVID-19 in comparison to baseline molecular testing grouped by HIV status.

Appendix figure 4



Appendix figure 4: subgroup analysis of performance of radiologist and software CAD1 and CAD2 at identifying COVID-19 in comparison to baseline molecular testing grouped by diabetes status.

Appendix table 1

	Radiologist assessment	
	Negative	**Positive**
CAD1		
Negative	95 (45%)	151 (33%)
Positive	116 (55%)	309 (67%
CAD2		
Negative	162 (77%)	119 (26%)
Positive	49 (23%)	341 (74%)

Appendix table 1a: Agreement of CAD software with radiologists, at manufacturer suggested thresholds for COVID-19

Software	N	Agreement	AC1
CAD1	**671**	**60.00%**	**0.28**
CAD2	**671**	**75.00%**	**0.53**

Appendix table 1b: Percentage agreement of CAD software with radiologists, at manufacturer suggested thresholds for COVID-19 and the calculated agreement coefficient. AC1: Gwet's Agreement Coefficient.