

Protocol

AI-Assisted Chest X-Ray Interpretation in Resource-Limited Settings: LuAna Stepped-Wedge Trial Protocol

Maria Carolina Bueno da Silva¹, MD; Paula Bresciani M de Andrade¹, PhD; Henrique Min Ho Lee¹, MD; Pedro Vinicius Alves Silva¹, BSc; Ana Cristina Ferreira¹, MD; Cintia Pereira Kuss¹, MSc; Maria Gabriela de Almeida Rodrigues¹, MSc; Guilherme Alberto Sousa Ribeiro¹, PhD; Thiago Felliipe Ortiz Camargo¹, MSc; William Yang Chen Fan¹, MD; Pedro Vieira Santana Netto¹, MD; Giovanna de Souza Mendes¹, BSN; Gilberto Szarf¹, MD, PhD; Rafael Maffei Loureiro¹, MD; Ary Serpa Neto^{2,3,4,5}, MD, PhD; Joselisa Péres Queiroz de Paiva¹, MSc, PhD; Jaqueline Driemeyer Correia Horvath¹, PhD

¹Department of Radiology, Hospital Israelita Albert Einstein, São Paulo, São Paulo, Brazil

²Department of Critical Care Medicine, Hospital Israelita Albert Einstein, São Paulo, São Paulo, Brazil

³Australian and New Zealand Intensive Care Research Centre (ANZIC-RC), School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia

⁴Department of Intensive Care, Austin Hospital, Melbourne, Australia

⁵Department of Critical Care, Melbourne Medical School, University of Melbourne, Austin Hospital, Melbourne, Australia

Corresponding Author:

Maria Carolina Bueno da Silva, MD
Department of Radiology
Hospital Israelita Albert Einstein
Av. Albert Einstein 627, Bldg. D, 4th Floor, 627/701
São Paulo, São Paulo 05652-900
Brazil
Phone: 55 11998314571
Email: maria.cbsilva@einstein.br

Abstract

Background: Artificial intelligence (AI) has the potential to transform chest radiography interpretation by enhancing diagnostic accuracy, identifying subtle findings, reducing errors, and helping prioritize patient care. Although chest radiography remains a cost-effective and widely used imaging tool, its effectiveness is limited by overlapping anatomy and variability in clinical expertise. Integrating AI can help overcome some of these challenges, especially in resource-constrained settings. However, robust validation in real-world clinical contexts is essential before widespread implementation. This study protocol evaluates whether AI assistance improves general practitioners' ability to detect radiographic findings on chest radiography in adults with respiratory complaints or those undergoing treatment for respiratory diseases compared with unaided interpretation. Potential benefits include increased diagnostic safety, higher physician confidence, more efficient workflows, and expanded access to expert support in underserved areas.

Objective: This study aims to evaluate whether AI assistance enhances physicians' ability to detect key radiographic abnormalities, including consolidation or pulmonary opacity, pneumothorax, atelectasis, pleural effusion, and cardiomegaly. The primary outcome is the difference in physicians' diagnostic accuracy (per examination) when assisted by the AI tool compared with usual practice, using expert radiologist consensus as the reference value.

Methods: This study is a protocol for a multicenter, stepped-wedge, cluster-randomized clinical trial following the CONSORT-AI (Consolidated Standards of Reporting Trials-Artificial Intelligence) extension and SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence) guidelines. The intervention involves the diagnostic support solution for chest radiography, Lung Analysis (LuAna), an AI-powered chest X-ray interpretation tool developed in partnership with the Brazilian Ministry of Health. Across 9 cities in Brazil, clusters will transition monthly from unaided chest X-ray interpretation by general practitioners to AI-assisted interpretation, with performance benchmarked against thoracic radiologists. The stepped-wedge design ensures that all clusters receive the intervention, reflecting real-world coordination, enhancing acceptability, improving statistical power, and strengthening causal inference through repeated measures. Diagnostic performance will be compared with a reference standard established by thoracic radiologists.

Results: This project was funded in October 2024 (following ethics approval by the institutional review board). Data collection commenced in January 2026 and is projected to be completed by September 2026, marking the end of the trial period. As of November 2025, 3 centers were fully prepared for enrollment initiation. The LuAna clinical trial is currently ongoing, with data analysis (including statistical analyses) forecasted to be finalized by November 2026. Results are expected to be published by January 2027.

Conclusions: This intervention is expected to enhance clinical decision-making by supporting earlier treatment initiation and more appropriate diagnostic pathways for patients with respiratory symptoms while maintaining a favorable safety profile and high physician usability. Findings from this trial will provide real-world evidence on the clinical utility of AI-assisted chest radiography. If effective, LuAna may leverage its scalability and equity advantages to become a replicable model for integrating AI into routine imaging workflows worldwide, especially in regions with limited access to specialist care.

Trial Registration: ClinicalTrials.gov NCT06686251; <https://clinicaltrials.gov/study/NCT06686251>

International Registered Report Identifier (IRRID): PRR1-10.2196/88626

JMIR Res Protoc 2026;15:e88626; doi: [10.2196/88626](https://doi.org/10.2196/88626)

Keywords: chest X-ray; artificial intelligence; pulmonary abnormalities; tuberculosis; cluster-randomized clinical trial; stepped wedge; protocol

Introduction

Artificial intelligence (AI) is increasingly recognized as a transformative tool in global health care. In high-income countries, AI is already enhancing diagnostic accuracy, supporting clinical decision-making, accelerating research, and improving health system management [1]. However, most low- and middle-income countries (LMICs) have yet to benefit from these advances [2]. In LMICs, AI holds significant potential, enabling better care, strengthening disease surveillance, expanding telemedicine, and improving medical imaging interpretation [2]. Integrating AI in these settings could accelerate progress toward the United Nations Sustainable Development Goals [3] and promote fair access to advanced health care technologies [4].

Chest radiography is the most commonly performed imaging test worldwide and remains vital for assessing cardiothoracic conditions [5,6]. It accounts for a significant portion of the billions of radiologic examinations conducted globally each year [7]. Chest X-ray (CXR) is still a cornerstone for diagnosing pulmonary diseases, particularly where advanced imaging is unavailable. It is accessible and cost-effective [8], but its interpretation can be hindered by overlapping anatomical structures, technical limitations, and variable reader expertise [9,10]. These limitations are more pronounced in LMICs, where shortages of trained specialists contribute to diagnostic delays and errors [11].

Given its high volume and clinical relevance, CXR imaging has become a key target for developing deep-learning tools to improve diagnostic accuracy and workflow efficiency [8,12]. AI offers promising solutions by automating image analysis and supporting frontline physicians in clinical decision-making. Advances in convolutional neural networks have enabled the development of AI tools with diagnostic accuracy comparable to radiologists for key findings, such as pneumonic consolidation, pneumothorax, and pulmonary nodules [4,6,13]. A recent study shows that an AI system using deep convolutional neural networks could reliably analyze chest radiographs to identify features

of COVID-19 pneumonia, such as ground-glass opacities and consolidation, offering a faster, cheaper alternative to computed tomography (CT) scans [14]. Though accuracy decreased for non-COVID-19 abnormalities poorly represented in the training data, it remained useful for screening and confirming typical cases, with suspicious cases requiring a CT scan. Ongoing improvements and diverse data will boost its performance, making AI-assisted CXR evaluation practical and cost-effective. However, 2 persistent challenges remain: (1) integrating AI into real-world clinical workflows and (2) providing effective support for nonradiologist physicians in resource-limited environments.

Rigorous validation is essential before implementing AI tools. This process requires not only comparisons with expert diagnoses to establish accuracy and reliability, but also prospective evaluation in real-world clinical settings to confirm safety, usability, and cost-effectiveness. Despite being resource-intensive, such efforts have already shown tangible benefits: in lung cancer pathways, adjunct AI for CXR interpretation can reduce missed cancers and is expected to be cost-saving compared with radiologist-only reporting [15,16].

In tuberculosis (TB) control, economic models from Pakistan indicate that AI-based computer-aided CXR triage can reduce diagnostic costs by approximately 20% to 40% compared with standard smear- or GeneXpert-based algorithms, primarily by limiting unnecessary confirmatory testing while maintaining case detection [17,18]. Cost-effectiveness is strongest in moderate- to high-prevalence settings and high-throughput programs, where economies of scale reduce per-screen costs. Similarly, evidence from Zambia shows that comprehensive radiography-based case-finding interventions can improve detection and achieve favorable incremental cost-effectiveness compared with passive case detection [19]. Overall, economic value is driven by prevalence, screening volume, and test cost differentials, with added system-level benefits including earlier diagnosis and more efficient resource allocation [15-19].

Despite the promising retrospective performance of AI models for CXR interpretation, their real-world effectiveness remains uncertain when deployed across large, heterogeneous health systems. Models trained and validated in limited or single-center datasets frequently experience distribution shifts that degrade performance when applied to different populations, imaging equipment, and clinical workflows [20]. A continental-scale public health system, such as Brazil's Unified Health System (Sistema Único de Saúde [SUS]), encompasses substantial geographic, demographic, and operational heterogeneity, making it an ideal yet challenging environment to evaluate clinical utility and equity of AI tools. Therefore, prospective, multicenter implementation trials that measure the impact of AI on clinician performance and patient management across diverse, real-world settings are essential to establish generalizability, robustness, and fairness prior to widescale adoption [20,21].

The LuAna (Lung Analysis) trial will evaluate whether AI-assisted CXR interpretation improves the detection of radiographic findings in adults with respiratory symptoms or undergoing treatment for respiratory diseases compared with physician-only interpretation within Brazil's Unified Health System. We hypothesize that AI assistance will enhance diagnostic accuracy, expedite triage, reduce delays, improve cost-effectiveness, and facilitate scalability. If successful, LuAna could become a model for equitable AI integration into diagnostic workflows in LMICs and beyond.

This study aims to evaluate the effectiveness of AI-assisted CXR interpretation in improving diagnostic accuracy, clinical efficiency, and patient outcomes within Brazil's Unified Health System (SUS). Using a stepped-wedge cluster-randomized design, this protocol aims to investigate whether AI-based decision-support tools enhance general practitioners' ability to detect predefined radiographic findings in adults presenting with respiratory symptoms or

with pulmonary disease compared with unaided interpretation. The trial further seeks to determine AI's potential to strengthen diagnostic performance in resource-limited settings and inform the implementation of scalable, equitable, and efficient health care.

Methods

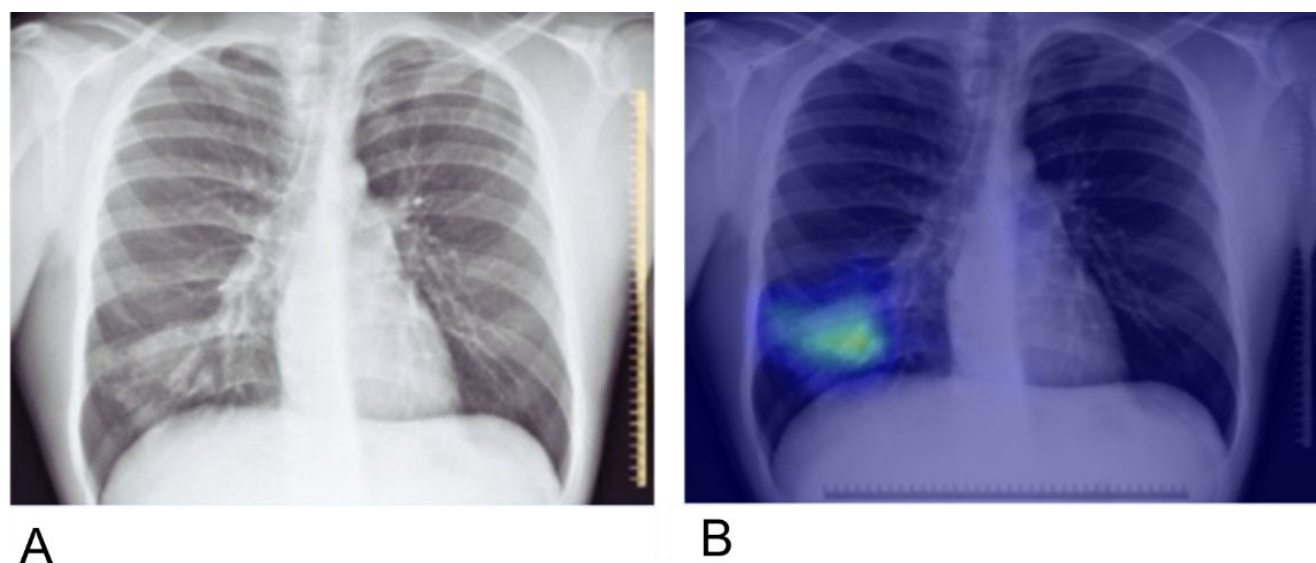
Algorithm Development

Overview

Between 2019 and 2024, 3 independent deep-learning models were developed and internally validated [22]. The "Lung Abnormality" model classified CXRs as either normal or abnormal. The "Radiological Findings" model categorized patterns into predefined classes, including consolidation, pulmonary opacity, pneumothorax, atelectasis, pleural effusion, and cardiomegaly. Finally, the "Tuberculosis" model predicted whether a CXR was suggestive or not suggestive of pulmonary TB. The models were trained on CXRs from adults aged 18 years or older, including both male and female patients. The dataset comprised anteroposterior or posteroanterior views sourced from public databases and from multiple partner health care centers across Brazil, including referral clinics for TB care.

Model predictions were given as probabilities ranging from 0 to 1, with a threshold of 0.5: values of 0.5 or higher were classified as positive, and values below 0.5 as negative. The exceptions are pneumothorax (0.8), consolidation (0.6), and TB (0.4), which had different thresholds determined by the GHOST (generalized threshold shifting) method [23]. Heatmaps highlight regions of interest, contributing to the algorithm's output (Figure 1).

Figure 1. Chest X-ray (CXR) processed by the radiological findings model. (A) Original CXR image showing consolidation in the right lower lung field. (B) Artificial intelligence-generated output with heatmap highlighting the region with the highest predicted probability of consolidation.



Internal Algorithm Validation

Internal validation proved robust performance, with the 3 models showing consistent outcomes across the validation datasets (Table 1). Overall metrics for the models were as follows: the “Pulmonary Abnormality” model achieved an area under the receiver operating characteristic curve (AUC) of 0.82, with an accuracy of 0.82, a sensitivity of 0.82, and a specificity of 0.84. The “Radiological Findings” model

reached an overall AUC of 0.82, with per-class performance ranging from 0.77 (atelectasis) to 0.90 (pneumothorax). The “Tuberculosis” model achieved an AUC of 0.81, with an accuracy of 0.81, a sensitivity of 0.83, and a specificity of 0.78. These findings suggest that the models could be suitable for eventual clinical use, though additional testing in prospective real-world clinical settings is still necessary.

Table 1. Internal validation dataset values.

Model	AUC ^a	Sensitivity	Specificity	Accuracy
Pulmonary abnormality	0.82	0.82	0.84	0.82
Tuberculosis	0.81	0.83	0.78	0.81
Radiological findings				
Consolidation	0.78	0.84	0.57	0.59
Lung opacity	0.78	0.89	0.50	0.64
Atelectasis	0.77	0.88	0.49	0.57
Cardiomegaly	0.84	0.86	0.67	0.70
Pneumothorax	0.90	0.81	0.85	0.83
Pleural effusion	0.90	0.92	0.70	0.78

^aAUC: area under the receiver operating characteristic curve.

External Algorithm Validation

An external retrospective validation of 1092 anonymized CXRs, including 586 double-read by thoracic radiologists at Hospital Israelita Albert Einstein, demonstrated variability in model performance. The Pulmonary Abnormality and

Tuberculosis models achieved AUC values of 0.89 and 0.84, respectively, whereas the Radiological Findings model showed AUC values ranging from 0.811 to 0.96, as shown in Table 2.

Table 2. External validation dataset values.

Model	AUC ^a	Sensitivity	Specificity	Accuracy
Pulmonary abnormality	0.89	0.82	0.84	0.82
Tuberculosis	0.84	0.83	0.71	0.77
Radiological findings				
Consolidation	0.92	0.95	0.70	0.74
Lung opacity	0.91	0.90	0.72	0.80
Atelectasis	0.81	0.55	0.87	0.81
Cardiomegaly	0.95	0.84	0.90	0.90
Pneumothorax	0.96	1.0	0.79	0.80
Pleural effusion	0.95	0.80	0.92	0.90

^aAUC: area under the receiver operating characteristic curve.

In the prospective phase of the Radiological Findings model, which included 30 real-time cases, no statistically significant differences were observed. All physicians were provided with standardized CARPL (Clinical AI Research Platform for Labeling) training to promote uniformity in image labeling [24].

Therefore, the current gap between internal and external performance per class of the Radiological Findings model underscores the need for multi-institutional datasets and rigorous real-world testing before clinical adoption. The AI models will be locked for the duration of the trial; no updates to weights or thresholds will be implemented once recruitment starts. Overall, this study underscores both the potential and limitations of AI in radiology, reinforcing that

limited real-world evidence warrants caution and that robust validation frameworks are as essential as algorithm development [25-27].

Study Design: Stepped-Wedge Randomized Design

The stepped-wedge cluster-randomized design was chosen for this trial because of its clear methodological benefits. Unlike parallel designs, all clusters eventually receive intervention, enhancing acceptability in clinical settings where prior evidence suggests potential benefit [28,29]. The phased implementation also mirrors real-world logistical constraints while allowing for both within- and between-cluster comparisons, thereby improving statistical efficiency

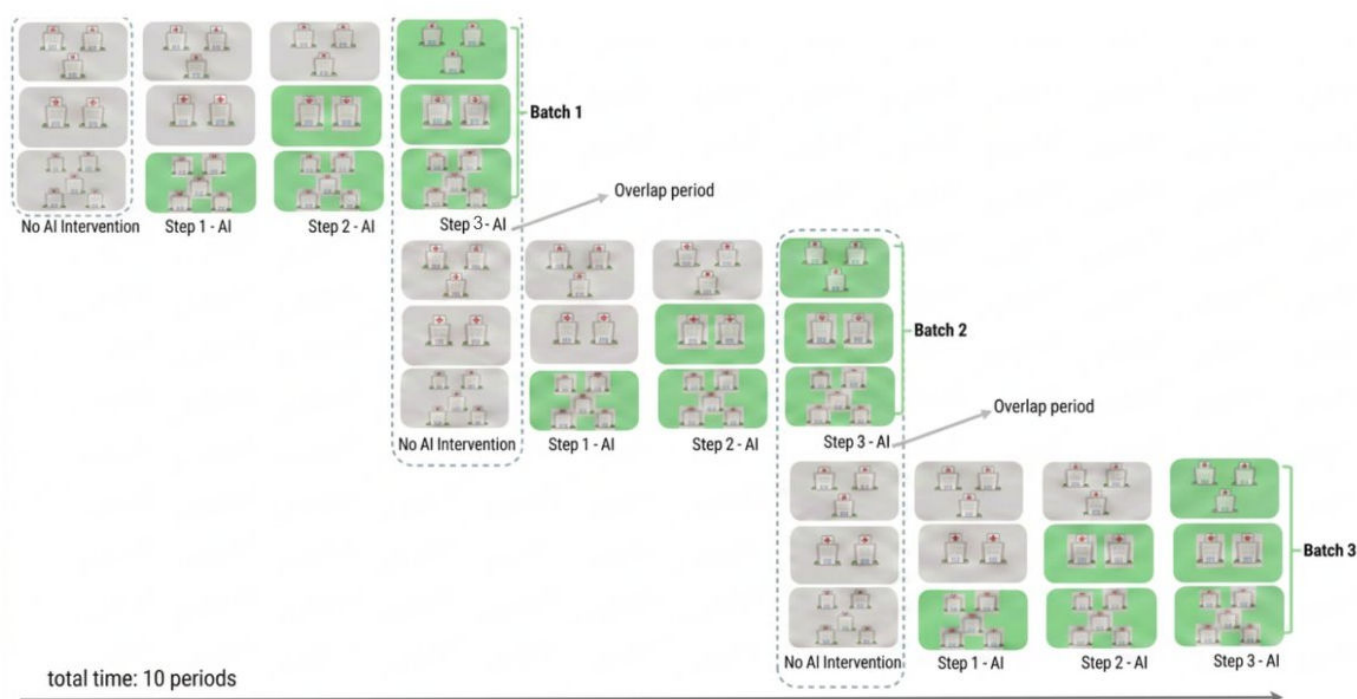
and power [28,30]. Additionally, repeated measurements over time enable adjustment for secular trends, strengthening causal inference [28-31]. These characteristics make the stepped-wedge design especially suitable for pragmatic evaluations of health care interventions.

This nationwide, multicenter, stepped-wedge cluster-randomized trial (SW-CRT) follows the CONSORT-AI (Consolidated Standards of Reporting Trials-Artificial Intelligence) and SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence) guidelines [32,33]. It will evaluate whether AI can improve CXR interpretation in routine clinical care across Brazil. Cities will act as clusters, comprising public, private, and philanthropic health care units to ensure broad socioeconomic and geographic representation [28,30,31,34].

All clusters will begin in the control condition (standard physician interpretation) and will cross over to the intervention (AI-assisted physician interpretation) at randomly assigned monthly intervals. Potential contamination across clusters will be minimized by selecting independent clinical units with nonoverlapping staff and workflows.

A batched stepped-wedge approach allows staggered initiation, accommodating operational constraints while minimizing contamination or learning effects. Once a cluster transitions, it remains in the intervention arm (Figure 2). Simple urn-based randomization without replacement will occur independently in each batch. Participating centers within cities must have infrastructure for CXR interpretation (negatoscope or digital screen), reliable internet access, and commitment to study protocols.

Figure 2. Study design and implementation flow of the LuAna (Lung Analysis) trial. The LuAna trial uses a batched stepped-wedge cluster-randomized design, where all centers begin in the control phase (gray) with physician-only interpretation. At randomly assigned sequences, clusters are grouped into batches and sequentially cross over to the intervention phase (green). Once a cluster transitions, it remains in the intervention phase until trial completion. At the site level, patient inclusion and physician interpretation occur continuously. During control periods, only the physician's interpretation is recorded. During intervention periods, physicians first provide their interpretation, after which AI support (LuAna) is introduced. This design ensures staggered, overlapping transitions across centers while maintaining consistent trial procedures. AI: artificial intelligence.



All participating clusters will undergo a site initiation visit (SIV) prior to enrollment. All physician participants will complete the informed consent process and sign the informed consent form. During the SIV, local investigators will receive standardized training on study procedures, data entry, image capture protocols, and use of the study platform (LuAna). Training will include hands-on instruction to ensure the uniform application of inclusion criteria, proper anonymization of patient data, and completion of interpretation forms. Following the SIV, sites will begin patient recruitment and data collection. This study was registered on ClinicalTrials.gov (NCT06686251).

Radiograph Eligibility Criteria

Inclusion Criteria

Eligible chest radiographs must meet all of the following criteria: it should (1) be of individuals aged 18 years or older; (2) be obtained either during the presence of respiratory symptoms or while the patient is undergoing treatment for a respiratory disease; (3) be any radiograph without a preexisting report at the time of inclusion; (4) be acquired on any X-ray equipment, provided that at least 1 frontal chest view is included; and (5) be performed either during the acute episode of the respiratory condition or throughout the postacute follow-up period, until its completion.

Exclusion Criteria

CXRs performed for peripherally inserted central catheter line positioning, preoperative risk assessment, lung cancer screening, or trauma-related evaluation will be excluded. In addition, radiographs printed on plain paper and examinations with technical quality below the minimum acceptable standard will also be excluded.

Physicians

Radiograph interpretation will be performed by general practitioners or other nonradiologist physicians employed or otherwise engaged at participating trial sites. No restrictions will be imposed regarding age, sex, years since graduation, or previous training in imaging interpretation. All participating physicians will receive training using the study platform and must provide informed consent before contributing to interpretations. The training procedure emphasizes that AI outputs are advisory only; clinical responsibility remains fully with the interpreting physician, and AI suggestions never override clinical judgment. Afterward, they will complete a brief questionnaire assessing their baseline confidence in CXR interpretation, perceived challenges in reading CXRs, and initial impressions regarding the usability of the study platform (Multimedia Appendix 1).

The LuAna Platform

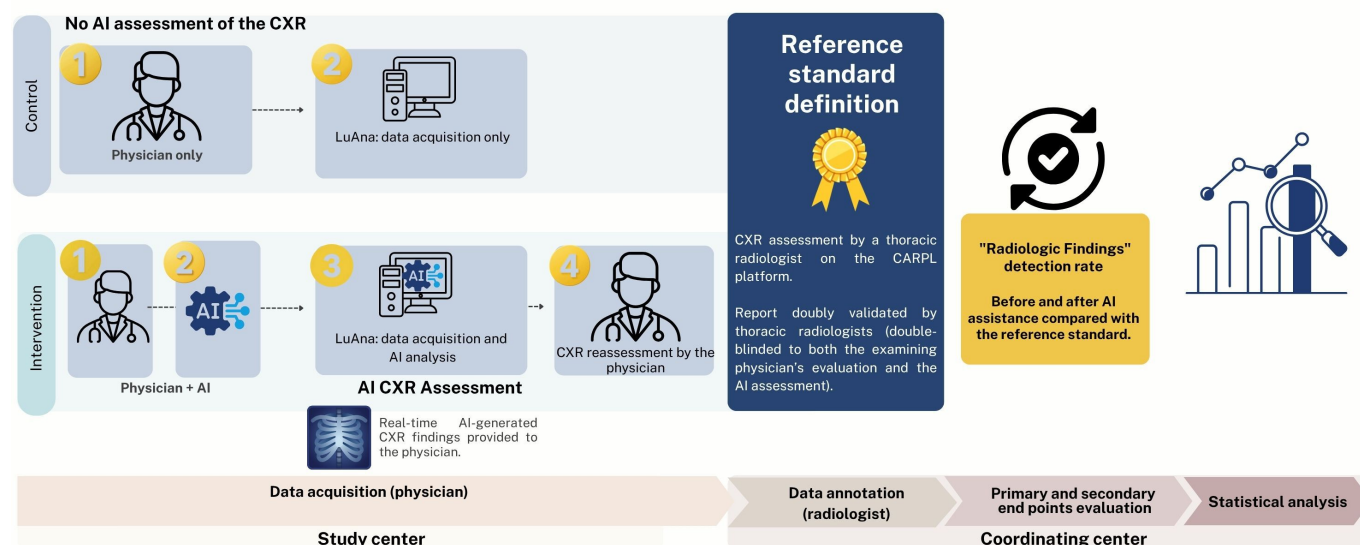
LuAna is a cloud-based AI platform for CXR interpretation developed in partnership with the Brazilian Ministry of Health. The platform receives anonymized CXR with the accompanying physician report and generates 3 deep-learning outputs: (1) classification of images as normal or abnormal, (2) classification of specific radiologic findings (consolidation or pulmonary opacity, pneumothorax, atelectasis, pleural effusion, and cardiomegaly), and (3) probability scores for radiographic features suggestive of pulmonary TB. Operating fully in the cloud, LuAna provides decision support without requiring modifications to current imaging systems or clinical workflows.

Data Collection

Anonymized data will be gathered using the LuAna platform (accessible on desktop or mobile), which records CXRs and physician interpretations (Figure 3). Physicians will access the platform using secure login credentials. All data will be automatically logged to enable auditing, reproducibility, and safety monitoring.

Figure 3. Workflow for chest X-ray (CXR) interpretation in control and intervention groups. Top panel (control group—without AI): CXRs from patients with respiratory complaints are interpreted solely by the physician (1), without access to AI support (2). Bottom panel (intervention group—with AI): physicians first perform an initial assessment of the CXR (1) and register their findings via the LuAna mobile app (2). After this, the AI model processes the image and returns a score and a heatmap highlighting abnormal regions (3). The physician can then review and, if needed, revise their interpretation based on the AI output (4). Both pre- and post-AI interpretations are recorded for analysis. For both groups, CXR assessments are compared against a reference standard, defined as thoracic radiologist evaluation performed on the CARPL, with reports independently double-validated by thoracic radiologists blinded to both the physician assessment and AI output. The primary outcome is the comparison of radiologic findings detection rates before and after AI assistance relative to the reference standard. The workflow includes data acquisition at the study centers, followed by radiologist annotation, primary and secondary end point evaluation, and statistical analysis at the coordinating center. AI: artificial intelligence; CARPL: Clinical Artificial Intelligence Research Platform for Labeling.

LuAna Trial



Only demographic variables, including age group, sex, and race, will be collected in aggregated form to characterize the study population and allow for fairness analyses, ensuring equitable and unbiased AI performance. CXRs will be captured either by photographing printed films positioned on

a lit negatoscope, by photographing digital images displayed on a picture archiving and communication system monitor, or by direct screen capture, all following strict protocols to guarantee full anonymization. Physicians will capture only the thoracic image, ensuring that all patient identifiers

are systematically removed and never visible. A patient ID entered solely to return the radiologist's report to the local health unit will be encrypted and remain inaccessible to the research team. Nondiagnostic images or studies failing predefined quality-control criteria will be excluded from analysis.

Physicians will complete a brief questionnaire of their first CXR interpretation based on the following predefined questions: (1) Is the CXR normal or abnormal? (2) Are the specified radiological findings present or absent? and (3) Does the CXR show features suggestive of pulmonary TB? In the intervention group, after the physician's clinical management, AI output (including probability scores and heatmaps) will be made available. Physicians will then document their final interpretation and decision-making process and subsequently record any changes made after reviewing the AI results. All data will be securely stored in encrypted, deidentified cloud-based repositories.

Reference Standard Generation, Report Return, and Critical Findings

The radiologists are responsible for producing the ground-truth (reference standard) for all CXR images during the study. The first step in ground-truth generation is to assign 2 thoracic radiologists to independently interpret the CXR images to set up the reference standard, classifying each finding as "normal" or "abnormal." The correct classification matches the reference standard. For cases classified as *abnormal*, the presence or absence of the relevant radiological

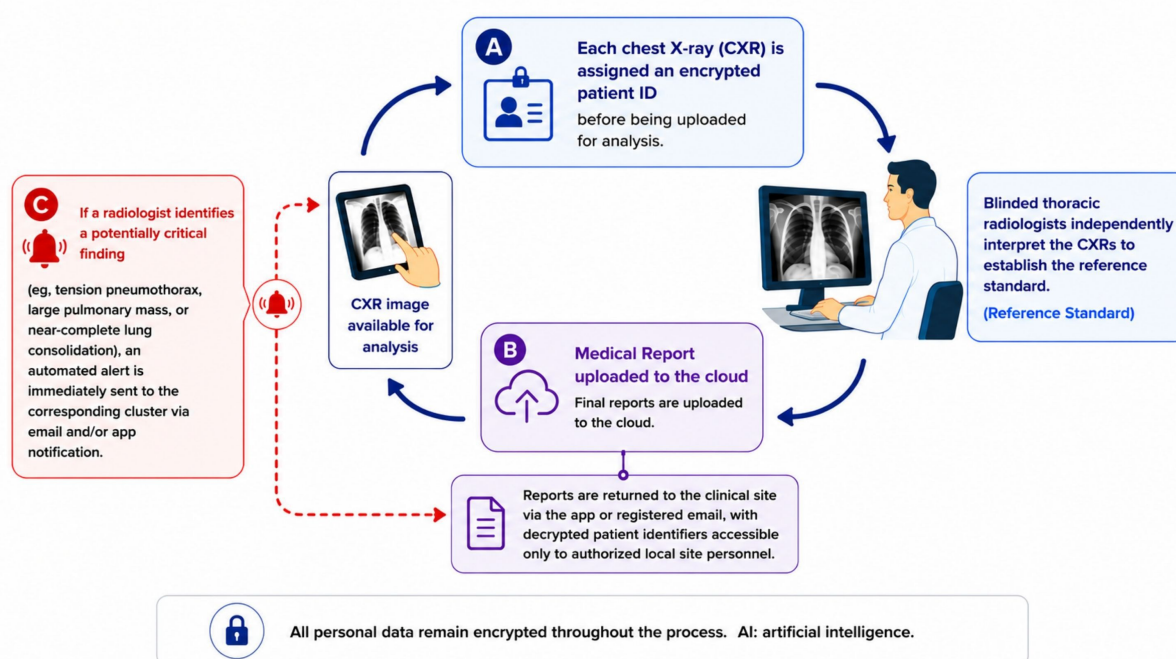
findings (consolidation, pulmonary opacity, pneumothorax, atelectasis, pleural effusion, and cardiomegaly) will be assessed, as well as the presence or absence of changes suggestive of TB. Based on these evaluations by the 2 thoracic radiologists, the ground truth is established. The radiologists will be blinded to both the physicians' interpretations and the AI outputs.

A chest radiograph will be labeled as suggestive of pulmonary TB if at least 1 radiographic feature considered compatible with pulmonary TB is present, as determined by the adjudicated thoracic radiologist reference standard. Examples include cavitation, upper-lobe-predominant consolidation or opacities, fibronodular changes or scarring with upper lobe volume loss, or a miliary pattern. If none of these features is present, the examination will be labeled as "not suggestive of pulmonary TB."

For physician performance analyses, an examination will be considered correctly classified when the physician's binary classification (normal vs abnormal; suggestive vs not suggestive of pulmonary TB; presence or absence of specific radiographic patterns) matches the adjudicated thoracic radiologist reference standard. In cases of any disagreement between the 2 experts in any of the evaluated items, a third radiologist, more senior, will adjudicate.

All findings will be registered in the CARPL, with abnormalities annotated to allow direct comparison with AI outputs. This will serve as the reference value for model validation and performance analysis (Figure 4).

Figure 4. Workflow for reference standard generation, report return, and critical findings alerts. (A) Each chest X-ray (CXR) is assigned an encrypted patient ID before being uploaded for analysis. (B) Blinded thoracic radiologists independently interpret the images to establish the reference standard. Final reports are uploaded to the cloud and returned to the clinical site via the app or registered email, with decrypted patient identifiers accessible only to the site. (C) If a radiologist identifies a potentially critical finding, such as tension pneumothorax, a large pulmonary mass, or near-complete lung consolidation, an automated alert is immediately sent to the cluster via email or app notification. All personal data remain encrypted throughout the process. AI: artificial intelligence.



Patient identifiers will be used solely to return reports of critical imaging findings to clinical sites, when present. These identifiers are cryptographically isolated from the research dataset and are never accessible to the research team (Figure 4). If the radiologist identifies a potentially critical finding, such as tension pneumothorax, a large obstructive mass, or near-complete lung consolidation, an alert will be sent to the originating site via email and/or phone (Figure 4). Patient data will remain encrypted; only the automated system will link the alert to the corresponding patients, ensuring privacy and prompt clinical follow-up. An institutional lifetime license ensures consistent labeling with the same software version throughout the study.

At the end of the study, all participating physicians and radiologists will be invited to complete a standardized questionnaire assessing the usability of the LuAna platform and its integration into routine clinical workflows. Responses will be used to address implementation challenges and opportunities for improvement (Multimedia Appendix 1).

Data and Safety Monitoring Board

An independent Data and Safety Monitoring Board, comprising external experts in clinical trials, radiology, statistics, and ethics with no conflicts of interest, will oversee the study in accordance with international guidelines and good clinical practice. The board will be responsible for safeguarding patient safety and ensuring study integrity.

Statistical Analysis

The sample size was calculated to detect a minimum difference of 10% in the primary outcome—a binary indicator of whether the physician correctly labeled a CXR—between the AI-assisted and control groups. The calculation accounted for the stepped-wedge cluster-randomized design with batched implementation using the following assumptions: 9 clusters randomized in 3 batches to cross over from control to intervention, 10 time periods, a baseline detection rate of 67% in the control group, an expected increase to 77% in the intervention group, and an intracluster correlation coefficient ranging from 0.001 to 0.003.

The design effect was adjusted for the correlation structure induced by repeated measures within clusters over time using methods described by Hussey and Hughes [28] and extended to account for batched randomization, as described by Hemming et al [35,36]. The power calculation targeted 80% power at a 2-sided significance level of 5% and accounted for potential variability in cluster sizes. The final average cluster size was 37, and the overall sample size was inflated by 10% to account for anticipated exclusions due to image quality issues or missing data (dropout). This approach ensures adequate power to detect a meaningful effect of the intervention while maintaining the integrity of the stepped-wedge design with staggered implementation across batches.

Primary analysis will follow the modified intention-to-treat principle. All eligible CXRs (the unit of analysis in this study) will be analyzed according to the study group (control

or intervention), to which their corresponding cluster was allocated at the time of image acquisition, regardless of physician adherence to the intervention or exposure to AI output. Images of poor quality, as defined by our radiologists, will be excluded postrandomization.

The primary outcome analysis will use a generalized linear mixed model with a binomial distribution and logit link function to estimate the effect of the AI-assisted intervention on detection probability. The model will include fixed effects for the intervention group (AI-assisted vs control), calendar time (to adjust for secular trends), time since intervention rollout (to account for learning or adaptation effects), and batch effects. In addition, random intercepts for clusters will be specified to account for intracluster correlation.

Results will be reported as adjusted odds ratios with 95% CIs and 2-sided *P* values. Model fit and assumptions will be evaluated using the Residual Diagnostics for Hierarchical (Multilevel/Mixed) Regression Models (DHARMa) [37]. In case of convergence issues or model instability, generalized estimating equations with robust SEs will be used as an alternative to estimate population-averaged effects. Independent models will also be fitted for each radiological finding, as well as for the pulmonary abnormality and TB algorithm.

Model performance metrics will be reported with 95% CIs [38,39]. Responses to questions regarding the mobile interface will be summarized as frequencies. Chi-square test and ordinal logistical regression will be used to investigate associations between these responses and physician characteristics (eg, years in practice and biological sex). Additionally, generalized linear mixed models will be fitted with the algorithm's output as the dependent variable, adjusted for image characteristics, to investigate their effect on AI classification.

Missing data are expected to be minimal due to app-enforced completeness checks. Incomplete records will be described, and sensitivity analyses using multiple imputation will be performed if the proportion of missing data exceeds 10%. All analyses will be performed in R (R Foundation for Statistical Computing), and code scripts will be version-controlled and archived to ensure reproducibility.

Ethical Considerations

This study was reviewed and approved by the local Ethics Committee of Hospital Israelita Albert Einstein (CEP Einstein), São Paulo, Brazil, registered under IRB00005041, in accordance with applicable national and international regulations (CAAE number 82494524.0.1001.0071). A waiver of written informed consent for patients was granted, as the study is strictly observational and involves no intervention in routine clinical care, no modification of clinical management, and no access to patients' medical records. Only chest radiographs obtained as part of standard care were used, and all images were fully anonymized prior to storage and analysis, with no directly identifiable information included in the research dataset. However, written informed

consent was obtained from the physicians responsible for enrolling patients in the LuAna trial, as they are considered direct research participants.

All data are handled under strict confidentiality procedures. Chest radiograph images were collected via the LuAna app and stored in an anonymized format on a secure cloud-based server.

The analytical database contains exclusively deidentified data. Access to identifiable information at the study sites, when applicable, is restricted to authorized local investigators. Study data will be securely stored at the hosting institution for a minimum of 5 years, in compliance with institutional and regulatory requirements.

Posttrial access pathways for the tool will be discussed with the Brazilian Ministry of Health to ensure continuity if the intervention proves beneficial. Technical details of the algorithm, including its development, training, validation, and performance characteristics, will be described in accordance with the CLAIM (Checklist for Artificial Intelligence in Medical Imaging) reporting guidelines [40], complementing the CONSORT-AI and SPIRIT-AI standards applied to the clinical trial components [32,33].

Outcomes

Primary Outcome

The primary outcome is expected to show that AI support improves physicians’ ability to detect radiographic

abnormalities, including consolidation or pulmonary opacity, pneumothorax, atelectasis, pleural effusion, and cardiomegaly. Physicians are anticipated to be more likely to correctly identify at least 1 of these 5 findings when assisted by AI than when interpreting CXRs unaided. This outcome is defined as a patient-level binary indicator of whether the physician correctly identifies any of the 5 prespecified findings compared with the thoracic radiologist’s reference standard.

Secondary Outcomes

The secondary outcomes are expected to demonstrate that AI assistance increases detection rates for each of the 5 radiographic findings, as well as overall pulmonary abnormality and TB-related patterns, relative to unaided physician interpretation. Improvements in abnormality localization accuracy are also anticipated, both for the AI outputs alone and for physician interpretations after reviewing AI results. Diagnostic performance compared with the reference standard is expected to improve, with higher sensitivity, specificity, AUC, positive predictive value, and negative predictive value, reflecting greater reliability in identifying clinically relevant radiological abnormalities. The secondary outcomes are depicted in [Textbox 1](#).

Textbox 1. Secondary objectives and outcomes.

Secondary objectives
• Measure the effect of artificial intelligence (AI) on physicians’ detection rates for each of the 5 radiologic findings independently, pulmonary abnormality, and features suggestive of pulmonary tuberculosis • Evaluate AI algorithm performance in a real-world setting • Evaluate physicians’ experiences regarding usability of the mobile interface • Assess the impact of external influencing factors on AI algorithm predictions • Evaluate physicians’ learning related to interpretation of radiologic findings on chest X-rays (CXRs) before and after the intervention
Secondary outcomes
• Whether the physician correctly identifies each of the 5 radiologic findings, pulmonary abnormality, and features suggestive of pulmonary tuberculosis on chest radiographs • Comparison between AI algorithm outputs for CXRs and thoracic radiologist reference standard classifications to calculate sensitivity, specificity, area under the receiver operating characteristic curve, positive predictive value, and negative predictive value • Responses to questions regarding the mobile app interface, responsiveness, and functionality measured using a 7-point Likert scale • Associations between AI algorithm outputs for CXRs and CXR characteristics, including photo quality, presence of thoracic devices, and technical artifacts, as labeled by radiologists • Change in physician interpretation performance over time and according to the duration of participation in the study

Results

The LuAna trial received institutional ethics board approval and was formally activated in October 2024. Data collection began in January 2026 with a projected end date of September 2026. At interim review (November 2025), 3 of the 13 participating centers had completed all site activation

requirements and were prepared to initiate enrollment. The trial is currently ongoing with active data collection. Data analysis is scheduled for completion by November 2026, and results dissemination is expected by January 2027.

Discussion

Principal Findings

This trial is a pivotal effort to evaluate how AI can enhance diagnostic accuracy, clinical workflows, and physician performance in interpreting chest radiographs.

By using a batched stepped-wedge cluster-randomized design across a nationally representative sample of health care settings, we aim to rigorously assess the real-world utility of LuAna, an AI-powered platform for CXR interpretation developed in the public sector to assist general practitioners.

Despite ongoing debate about the advantages of SW-CRTs, including concerns about equipoise and delayed access to interventions [41], this design offers distinct benefits for implementation research. Governed by the same ethical principles as all clinical research [42], SW-CRTs are particularly valuable when evaluating innovations that are expected to be beneficial and for which withholding access may be contentious [42]. In the SUS context, this approach ensures that all centers ultimately gain access to AI support, promotes fairness and acceptability, and provides a pragmatic pathway to sustainable integration of the intervention into routine care.

The central goal is to determine whether LuAna improves clinicians' ability to detect radiographic abnormalities, particularly in environments with limited access to radiology expertise. Comparisons against adjudicated, double-blinded radiologist interpretations will enable the quantification of diagnostic agreement, while stratified analyses by site type and physician characteristics will offer insights into contextual factors that may influence AI performance.

Importantly, the trial also investigates the clinical impact beyond diagnostic accuracy. By evaluating patient management decisions and turnaround times, we aim to capture how AI affects treatment pathways and efficiency in routine care. Usability assessments will provide critical feedback for refining and scaling the platform, thereby ensuring seamless integration into diverse clinical workflows across both mobile and desktop interfaces.

Robustness will be evaluated through sensitivity analyses on image quality, artifacts, and medical devices, thereby validating the model's generalizability under real-world conditions. Additionally, longitudinal analysis of physician interpretations will explore whether repeated AI exposure improves independent performance, highlighting AI's potential educational role.

Beyond its methodological rigor, the LuAna trial advances an area in which robust evidence remains scarce: the real-world, prospective clinical evaluation of AI tools. While algorithmic performance is well documented in retrospective studies, few investigations have examined how AI behaves once deployed within the operational complexity of large, heterogeneous public health systems [40,43,44]. This gap

has been repeatedly highlighted by international frameworks, including the DECIDE-AI (Developmental and Exploratory Clinical Investigation of Decision Support Systems Driven-Artificial Intelligence) guidelines, which emphasize the need for pragmatic, early-stage clinical evaluation before widescale deployment. Indeed, studies show positive primary outcomes for AI in clinical practice, but concerns remain due to mostly single-center trials, limited demographic data, and inconsistent reports on operational efficiency, which affect generalizability and practical use [45,46].

Evaluating AI-based decision-support systems involves several challenges: understanding their nature as complex interventions; considering user variability and the biases it introduces; integrating human factors into the collaboration between clinicians and AI; assessing both real patients and their data; dealing with continuous system changes; preventing the reproduction of health inequalities; ensuring that results generalize to different contexts; and guaranteeing reproducibility in the real world. By implementing LuAna across a nationally diverse range of health care settings, this study not only provides valuable empirical data on AI performance in real-world situations but also offers a unique opportunity to examine issues related to fairness [47-49].

Variability in patient demographics, imaging equipment, staffing, and workflow conditions across Brazil's public health system creates a natural stress test for equity: the trial can assess whether AI-assisted interpretation benefits all subpopulations and settings equally or whether performance disparities emerge [50,51]. In doing so, it addresses 2 critical gaps in the global AI literature—real-world clinical validation and algorithmic fairness—offering evidence that is seldom available yet essential for responsible, scalable, and equitable integration of AI into routine care.

Grounded in operational realities and health care equity, this study contributes to a broader movement toward responsible, evidence-based AI integration in clinical practice. It aligns with international recommendations for validation frameworks and addresses a global need for tools that are both technically robust and contextually appropriate for low- and middle-income health systems.

Limitations

Although the stepped-wedge design enhances real-world applicability by including diverse health care settings, the findings may still be influenced by context-specific factors, such as infrastructure, physician training, and patient demographics. Variation in image acquisition, whether from printed films, computer displays, or mobile capture, may introduce heterogeneity that affects both physician interpretation and algorithm performance. In addition, while the trial evaluates diagnostic accuracy, turnaround times, usability, and model robustness, it does not directly measure patient-level outcomes such as morbidity or mortality. Equity analyses are further limited by the narrow range of demographic variables collected, which may not fully capture socioeconomic or clinical determinants of performance.

Finally, the lack of detailed clinical data, including comorbidities, treatment history, and laboratory findings, restricts the ability to contextualize radiographic interpretations and fully assess the AI integration into patient care pathways.

Conclusion

Future work should prioritize expanding datasets to include a broader spectrum of acquisition standards, additional radiographic projections (eg, lateral views), and complementary imaging modalities such as CT. Scaling to additional centers across Brazil—capturing greater geographic, demographic, and service-level diversity—will further strengthen dataset representativeness and enhance algorithm robustness. The adoption of standardized annotation protocols will support higher-quality data curation and reduce uncertainty in the reference standard. In parallel,

strategies such as transfer learning and active learning should be explored to optimize performance in data-limited environments.

Ultimately, LuAna seeks to demonstrate that AI-powered diagnostic support can be safely and effectively integrated into public health care, empowering clinicians and improving outcomes for underserved populations. This study will lay the groundwork for broader AI adoption in medical imaging, grounded in evidence, equity, and clinical impact.

If proven effective, LuAna could help bridge diagnostic gaps by supporting nonspecialist clinicians, accelerating decision-making, and strengthening health care delivery in underserved regions, offering a scalable, evidence-based model for equitable AI integration in medical imaging.

Acknowledgments

The authors are grateful to the Proadi-Sistema Único de Saúde Banco de Imagens team for their support and contributions to this project. The authors also thank Natali Lourenço Nascimento and all collaborators and team members involved in the project for their support and dedication.

Microsoft Copilot was used in a limited capacity to refine wording and improve clarity. No scientific content, analyses, or interpretations were generated using artificial intelligence (AI), and all AI-assisted edits were fully reviewed and approved by the authors.

The authors declare the use of generative AI (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing, summarizing text, and translation. The GAI tool used was Microsoft 365 Copilot.

Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Declaration submitted by Paula Bresciani M de Andrade.

Funding

This project is funded through the Support Program for Institutional Development of the Unified Health System (Proadi-SUS; NUP: 25000.156740/2023-25), a partnership between Hospital Israelita Albert Einstein and the Brazilian Ministry of Health. The study sponsor provided financial support for the trial but had no role in the study design, data collection, data management, data analysis, data interpretation, manuscript preparation, or the decision to submit the results for publication. The investigators retain full academic independence and ultimate authority over all aspects of the study.

Data Availability

All data generated or analyzed during this protocol study are included in this published article and [Multimedia Appendix 1](#).

Authors' Contributions

MCBdS is the primary author of the protocol, with assistance in research and protocol writing. MCBdS conceived the study design with the researchers ACF, ASN, CPK, GASR, GdSM, HMHL, JDCH, JPQdP, MGdAR, PBMdA, PVAS, RML, TFOC, and WYCF. ASN contributed to the design of the stepped-wedge protocol and the statistical analysis and participated in the protocol review. GS participated in the manuscript review. PBMdA prepared the manuscript. PVAS assisted with statistical analysis. All authors discussed the results, commented on the manuscript, and reviewed it. JDCH is the main author of the protocol, responsible for project administration and supervision. JPQdP supervised the project and critically reviewed the protocol. PVSN is the coordinator of this study.

Conflicts of Interest

The authors are participants in the Brazilian Ministry of Health through the Support Program for Institutional Development of the Unified Health System (Proadi-SUS) Banco de Imagens project, a collaboration between the Brazilian Ministry of Health and Hospital Israelita Albert Einstein aimed at enhancing public health in Brazil. LuAna is currently a research-use-only system and not a certified medical device.

Multimedia Appendix 1

Application evaluation questionnaire for physicians

[\[PDF File \(Adobe File\), 471 KB-Multimedia Appendix 1\]](#)

Checklist 1

SPIRIT Checklist

[\[DOCX File \(Microsoft Word File\), 32 KB-Checklist 1\]](#)**References**

1. Artificial intelligence for health: opportunities, risks, and governance. World Health Organization. 2024. URL: <https://www.who.int/news-room/events/detail/2024/05/30/default-calendar/artificial-intelligence-for-health-opportunities-risks-and-governance> [Accessed 2026-05-19]
2. Akbarialiabad H, Sadeghian N, Haghighat S, et al. The utility of generative AI in advancing global health. *NEJM AI*. Feb 27, 2025;2(3). [doi: [10.1056/AIp2400875](https://doi.org/10.1056/AIp2400875)]
3. The 17 Goals. United Nations Sustainable Development Goals (SDGs). URL: <https://sdgs.un.org/goals> [Accessed 2026-05-19]
4. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv*. Preprint posted online on Nov 14, 2017. [doi: [10.48550/arXiv.1711.05225](https://doi.org/10.48550/arXiv.1711.05225)]
5. Hwang EJ, Nam JG, Lim WH, et al. Deep learning for chest radiograph diagnosis in the emergency department. *Radiology*. Dec 2019;293(3):573-580. [doi: [10.1148/radiol.2019191225](https://doi.org/10.1148/radiol.2019191225)] [Medline: [31638490](https://pubmed.ncbi.nlm.nih.gov/31638490/)]
6. Sierra JFT, Herrera JAC, Sutachan MAT, Zafra MJJ, Rincón EHH. Innovations in artificial intelligence to minimize diagnostic error - a comparison with human interpretation of chest radiographs in the clinical context: a scoping review. *Health Technol*. Nov 2025;15(6):1021-1036. [doi: [10.1007/s12553-025-00999-z](https://doi.org/10.1007/s12553-025-00999-z)]
7. Mettler FA Jr, Bhargavan M, Faulkner K, et al. Radiologic and nuclear medicine studies in the United States and worldwide: frequency, radiation dose, and comparison with other radiation sources—1950-2007. *Radiology*. Nov 2009;253(2):520-531. [doi: [10.1148/radiol.2532082010](https://doi.org/10.1148/radiol.2532082010)] [Medline: [19789227](https://pubmed.ncbi.nlm.nih.gov/19789227/)]
8. Çallı E, Sogancioglu E, van Ginneken B, van Leeuwen KG, Murphy K. Deep learning for chest X-ray analysis: a survey. *Med Image Anal*. Aug 2021;72:102125. [doi: [10.1016/j.media.2021.102125](https://doi.org/10.1016/j.media.2021.102125)] [Medline: [34171622](https://pubmed.ncbi.nlm.nih.gov/34171622/)]
9. Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *Radiographics*. Oct 2015;35(6):1668-1676. [doi: [10.1148/rg.2015150023](https://doi.org/10.1148/rg.2015150023)] [Medline: [26466178](https://pubmed.ncbi.nlm.nih.gov/26466178/)]
10. Brady AP. Error and discrepancy in radiology: inevitable or avoidable? *Insights Imaging*. Feb 2017;8(1):171-182. [doi: [10.1007/s13244-016-0534-1](https://doi.org/10.1007/s13244-016-0534-1)] [Medline: [27928712](https://pubmed.ncbi.nlm.nih.gov/27928712/)]
11. Frija G, Blažić I, Frush DP, et al. How to improve access to medical imaging in low- and middle-income countries? *EClinicalMedicine*. Aug 2021;38:101034. [doi: [10.1016/j.eclinm.2021.101034](https://doi.org/10.1016/j.eclinm.2021.101034)] [Medline: [34337368](https://pubmed.ncbi.nlm.nih.gov/34337368/)]
12. 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*. Oct 18, 2019;9(1):15000. [doi: [10.1038/s41598-019-51503-3](https://doi.org/10.1038/s41598-019-51503-3)] [Medline: [31628424](https://pubmed.ncbi.nlm.nih.gov/31628424/)]
13. Hwang EJ, Park S, Jin KN, et al. Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Netw Open*. Mar 1, 2019;2(3):e191095. [doi: [10.1001/jamanetworkopen.2019.1095](https://doi.org/10.1001/jamanetworkopen.2019.1095)] [Medline: [30901052](https://pubmed.ncbi.nlm.nih.gov/30901052/)]
14. Sasaki K, Garcia-Manero G, Nigo M, et al. Artificial intelligence assessment of chest radiographs for COVID-19. *Clin Lymphoma Myeloma Leuk*. May 2025;25(5):319-327. [doi: [10.1016/j.clml.2024.11.013](https://doi.org/10.1016/j.clml.2024.11.013)] [Medline: [39710565](https://pubmed.ncbi.nlm.nih.gov/39710565/)]
15. Colquitt J, Jordan M, Court R, et al. Artificial intelligence software for analysing chest X-ray images to identify suspected lung cancer: an evidence synthesis early value assessment. *Health Technol Assess*. Aug 2024;28(50):1-75. [doi: [10.3310/LKRT4721](https://doi.org/10.3310/LKRT4721)] [Medline: [39254229](https://pubmed.ncbi.nlm.nih.gov/39254229/)]
16. Tam MDBS, Dyer T, Dissez G, et al. Augmenting lung cancer diagnosis on chest radiographs: positioning artificial intelligence to improve radiologist performance. *Clin Radiol*. Aug 2021;76(8):607-614. [doi: [10.1016/j.crad.2021.03.021](https://doi.org/10.1016/j.crad.2021.03.021)] [Medline: [33993997](https://pubmed.ncbi.nlm.nih.gov/33993997/)]
17. Nsengiyumva NP, Hussain H, Oxlade O, et al. Triage of persons with tuberculosis symptoms using artificial intelligence-based chest radiograph interpretation: a cost-effectiveness analysis. *Open Forum Infect Dis*. Dec 2021;8(12):ofab567. [doi: [10.1093/ofid/ofab567](https://doi.org/10.1093/ofid/ofab567)] [Medline: [34917694](https://pubmed.ncbi.nlm.nih.gov/34917694/)]
18. Bashir S, Kik SV, Ruhwald M, et al. Economic analysis of different throughput scenarios and implementation strategies of computer-aided detection software as a screening and triage test for pulmonary TB. *PLoS One*. 2022;17(12):e0277393. [doi: [10.1371/journal.pone.0277393](https://doi.org/10.1371/journal.pone.0277393)] [Medline: [36584194](https://pubmed.ncbi.nlm.nih.gov/36584194/)]
19. Jo Y, Kagujje M, Johnson K, et al. Costs and cost-effectiveness of a comprehensive tuberculosis case finding strategy in Zambia. *PLoS One*. 2021;16(9):e0256531. [doi: [10.1371/journal.pone.0256531](https://doi.org/10.1371/journal.pone.0256531)] [Medline: [34499668](https://pubmed.ncbi.nlm.nih.gov/34499668/)]

20. Kim C, Yang Z, Park SH, et al. Multicentre external validation of a commercial artificial intelligence software to analyse chest radiographs in health screening environments with low disease prevalence. *Eur Radiol*. May 2023;33(5):3501-3509. [doi: [10.1007/s00330-022-09315-z](https://doi.org/10.1007/s00330-022-09315-z)] [Medline: [36624227](https://pubmed.ncbi.nlm.nih.gov/36624227/)]
21. Yang J, Dung NT, Thach PN, et al. Generalizability assessment of AI models across hospitals in a low-middle and high income country. *Nat Commun*. Sep 27, 2024;15(1):8270. [doi: [10.1038/s41467-024-52618-6](https://doi.org/10.1038/s41467-024-52618-6)]
22. de Camargo TFO, Ribeiro GAS, da Silva MCB, et al. Clinical validation of an artificial intelligence algorithm for classifying tuberculosis and pulmonary findings in chest radiographs. *Front Artif Intell*. 2025;8:1512910. [doi: [10.3389/frai.2025.1512910](https://doi.org/10.3389/frai.2025.1512910)] [Medline: [39991462](https://pubmed.ncbi.nlm.nih.gov/39991462/)]
23. Esposito C, Landrum GA, Schneider N, Stiefl N, Riniker S. GHOST: adjusting the decision threshold to handle imbalanced data in machine learning. *J Chem Inf Model*. Jun 28, 2021;61(6):2623-2640. [doi: [10.1021/acs.jcim.1c00160](https://doi.org/10.1021/acs.jcim.1c00160)] [Medline: [34100609](https://pubmed.ncbi.nlm.nih.gov/34100609/)]
24. CARPL. URL: <https://carpl.ai/> [Accessed 2026-05-20]
25. Lawrence R, Dodsworth E, Massou E, et al. Artificial intelligence for diagnostics in radiology practice: a rapid systematic scoping review. *EClinicalMedicine*. May 2025;83:103228. [doi: [10.1016/j.eclinm.2025.103228](https://doi.org/10.1016/j.eclinm.2025.103228)] [Medline: [40474995](https://pubmed.ncbi.nlm.nih.gov/40474995/)]
26. Korfiatis P, Kline TL, Meyer HM, et al. Implementing artificial intelligence algorithms in the radiology workflow: challenges and considerations. *Mayo Clin Proc Digit Health*. Dec 2024;3(1):100188. [doi: [10.1016/j.mcpdig.2024.100188](https://doi.org/10.1016/j.mcpdig.2024.100188)] [Medline: [40207002](https://pubmed.ncbi.nlm.nih.gov/40207002/)]
27. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. Dec 2017;42:60-88. [doi: [10.1016/j.media.2017.07.005](https://doi.org/10.1016/j.media.2017.07.005)] [Medline: [28778026](https://pubmed.ncbi.nlm.nih.gov/28778026/)]
28. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials*. Feb 2007;28(2):182-191. [doi: [10.1016/j.cct.2006.05.007](https://doi.org/10.1016/j.cct.2006.05.007)] [Medline: [16829207](https://pubmed.ncbi.nlm.nih.gov/16829207/)]
29. Dylla L, Douin DJ, Anderson EL, et al. A multicenter cluster randomized, stepped wedge implementation trial for targeted normoxia in critically ill trauma patients: study protocol and statistical analysis plan for the Strategy to Avoid Excessive Oxygen (SAVE-O2) trial. *Trials*. Nov 8, 2021;22(1):784. [doi: [10.1186/s13063-021-05688-6](https://doi.org/10.1186/s13063-021-05688-6)] [Medline: [34749762](https://pubmed.ncbi.nlm.nih.gov/34749762/)]
30. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ*. Feb 6, 2015;350:h391. [doi: [10.1136/bmj.h391](https://doi.org/10.1136/bmj.h391)] [Medline: [25662947](https://pubmed.ncbi.nlm.nih.gov/25662947/)]
31. Mdege ND, Man MS, Taylor Nee Brown CA, Torgerson DJ. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *J Clin Epidemiol*. Sep 2011;64(9):936-948. [doi: [10.1016/j.jclinepi.2010.12.003](https://doi.org/10.1016/j.jclinepi.2010.12.003)] [Medline: [21411284](https://pubmed.ncbi.nlm.nih.gov/21411284/)]
32. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. Sep 2020;26(9):1364-1374. [doi: [10.1038/s41591-020-1034-x](https://doi.org/10.1038/s41591-020-1034-x)] [Medline: [32908283](https://pubmed.ncbi.nlm.nih.gov/32908283/)]
33. Ibrahim H, Liu X, Rivera SC, et al. Reporting guidelines for clinical trials of artificial intelligence interventions: the SPIRIT-AI and CONSORT-AI guidelines. *Trials*. Jan 6, 2021;22(1):11. [doi: [10.1186/s13063-020-04951-6](https://doi.org/10.1186/s13063-020-04951-6)] [Medline: [33407780](https://pubmed.ncbi.nlm.nih.gov/33407780/)]
34. Hemming K, Taljaard M. Reflection on modern methods: when is a stepped-wedge cluster randomized trial a good study design choice? *Int J Epidemiol*. Jun 1, 2020;49(3):1043-1052. [doi: [10.1093/ije/dyaa077](https://doi.org/10.1093/ije/dyaa077)] [Medline: [32386407](https://pubmed.ncbi.nlm.nih.gov/32386407/)]
35. Kasza J, Bowden R, Hooper R, Forbes AB. The batched stepped wedge design: a design robust to delays in cluster recruitment. *Stat Med*. Aug 15, 2022;41(18):3627-3641. [doi: [10.1002/sim.9438](https://doi.org/10.1002/sim.9438)] [Medline: [35596691](https://pubmed.ncbi.nlm.nih.gov/35596691/)]
36. Hemming K, Kasza J, Hooper R, Forbes A, Taljaard M. A tutorial on sample size calculation for multiple-period cluster randomized parallel, cross-over and stepped-wedge trials using the Shiny CRT calculator. *Int J Epidemiol*. Jun 1, 2020;49(3):979-995. [doi: [10.1093/ije/dyz237](https://doi.org/10.1093/ije/dyz237)] [Medline: [32087011](https://pubmed.ncbi.nlm.nih.gov/32087011/)]
37. Hartig F. DHARMA: residual diagnostics for hierarchical (multi-level/mixed) regression models. The Comprehensive R Archive Network (CRAN). URL: <https://cran.r-project.org/web/packages/DHARMA/vignettes/DHARMA.html> [Accessed 2026-05-19]
38. Boyd K, Eng KH, Page CD. Area under the precision-recall curve: point estimates and confidence intervals. In: *Machine Learning and Knowledge Discovery in Databases*. Springer; 2013:451-466. [doi: [10.1007/978-3-642-40994-3_29](https://doi.org/10.1007/978-3-642-40994-3_29)]
39. Saito T, Rehmsmeier M. Precrec: fast and accurate precision-recall and ROC curve calculations in R. *Bioinformatics*. Jan 1, 2017;33(1):145-147. [doi: [10.1093/bioinformatics/btw570](https://doi.org/10.1093/bioinformatics/btw570)] [Medline: [27591081](https://pubmed.ncbi.nlm.nih.gov/27591081/)]
40. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. *Radiol Artif Intell*. Jul 2024;6(4):e240300. [doi: [10.1148/ryai.240300](https://doi.org/10.1148/ryai.240300)] [Medline: [38809149](https://pubmed.ncbi.nlm.nih.gov/38809149/)]
41. Prost A, Binik A, Abubakar I, et al. Logistic, ethical, and political dimensions of stepped wedge trials: critical review and case studies. *Trials*. Aug 17, 2015;16(1):351. [doi: [10.1186/s13063-015-0837-4](https://doi.org/10.1186/s13063-015-0837-4)] [Medline: [26278521](https://pubmed.ncbi.nlm.nih.gov/26278521/)]

42. Joag K, Ambrosio G, Kestler E, Weijer C, Hemming K, Van der Graaf R. Ethical issues in the design and conduct of stepped-wedge cluster randomized trials in low-resource settings. *Trials*. Dec 19, 2019;20(Suppl 2):703. [doi: [10.1186/s13063-019-3842-1](https://doi.org/10.1186/s13063-019-3842-1)] [Medline: [31852547](https://pubmed.ncbi.nlm.nih.gov/31852547/)]
43. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med*. May 2022;28(5):924-933. [doi: [10.1038/s41591-022-01772-9](https://doi.org/10.1038/s41591-022-01772-9)] [Medline: [35585198](https://pubmed.ncbi.nlm.nih.gov/35585198/)]
44. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*. Mar 25, 2020;368:m689. [doi: [10.1136/bmj.m689](https://doi.org/10.1136/bmj.m689)] [Medline: [32213531](https://pubmed.ncbi.nlm.nih.gov/32213531/)]
45. Han R, Acosta JN, Shakeri Z, Ioannidis JPA, Topol EJ, Rajpurkar P. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *Lancet Digit Health*. May 2024;6(5):e367-e373. [doi: [10.1016/S2589-7500\(24\)00047-5](https://doi.org/10.1016/S2589-7500(24)00047-5)] [Medline: [38670745](https://pubmed.ncbi.nlm.nih.gov/38670745/)]
46. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. Oct 29, 2019;17(1):195. [doi: [10.1186/s12916-019-1426-2](https://doi.org/10.1186/s12916-019-1426-2)] [Medline: [31665002](https://pubmed.ncbi.nlm.nih.gov/31665002/)]
47. Drukker K, Chen W, Gichoya J, et al. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. *J Med Imaging (Bellingham)*. Nov 2023;10(6):061104. [doi: [10.1117/1.JMI.10.6.061104](https://doi.org/10.1117/1.JMI.10.6.061104)] [Medline: [37125409](https://pubmed.ncbi.nlm.nih.gov/37125409/)]
48. Xu Z, Li J, Yao Q, Li H, Zhao M, Zhou SK. Addressing fairness issues in deep learning-based medical image analysis: a systematic review. *NPJ Digit Med*. Oct 17, 2024;7(1):286. [doi: [10.1038/s41746-024-01276-5](https://doi.org/10.1038/s41746-024-01276-5)] [Medline: [39420149](https://pubmed.ncbi.nlm.nih.gov/39420149/)]
49. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. Mar 2019;28(3):231-237. [doi: [10.1136/bmjqs-2018-008370](https://doi.org/10.1136/bmjqs-2018-008370)] [Medline: [30636200](https://pubmed.ncbi.nlm.nih.gov/30636200/)]
50. Willemink MJ, Koszek WA, Hardell C, et al. Preparing medical imaging data for machine learning. *Radiology*. Apr 2020;295(1):4-15. [doi: [10.1148/radiol.2020192224](https://doi.org/10.1148/radiol.2020192224)] [Medline: [32068507](https://pubmed.ncbi.nlm.nih.gov/32068507/)]
51. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*. Nov 2018;15(11):e1002683. [doi: [10.1371/journal.pmed.1002683](https://doi.org/10.1371/journal.pmed.1002683)] [Medline: [30399157](https://pubmed.ncbi.nlm.nih.gov/30399157/)]

Abbreviations

AI: artificial intelligence
AUC: area under the receiver operating characteristic curve
CARPL: Collaborative Artificial Intelligence Annotation Platform
CLAIM: Checklist for Artificial Intelligence in Medical Imaging
CONSORT-AI: Consolidated Standards of Reporting Trials-Artificial Intelligence
CT: computed tomography
CXR: chest X-ray
DECIDE-AI: Developmental and Exploratory Clinical Investigation of Decision Support Systems Driven-Artificial Intelligence
DHARMA: Residual Diagnostics for Hierarchical (Multilevel/Mixed) Regression Models
GHOST: generalized threshold shifting
LMICs: low- and middle-income countries
LuAna: Lung Analysis
SIV: site initiation visit
SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence
SUS: Sistema Único de Saúde
SW-CRT: stepped-wedge cluster-randomized trial
TB: tuberculosis

Edited by Javad Sarvestan; peer-reviewed by George Kafentzis; submitted 28.Nov.2025; final revised version received 09.Mar.2026; accepted 10.Mar.2026; published 13.Jul.2026

Please cite as:

da Silva MCB, de Andrade PBM, Lee HMH, Silva PVA, Ferreira AC, Kuss CP, Rodrigues MGdA, Ribeiro GAS, Camargo TFO, Fan WYC, Netto PVS, Mendes GdS, Szarf G, Loureiro RM, Serpa Neto A, Paiva JPQ, Horvath JDC
 AI-Assisted Chest X-Ray Interpretation in Resource-Limited Settings: LuAna Stepped-Wedge Trial Protocol
JMIR Res Protoc 2026;15:e88626
 URL: <https://www.researchprotocols.org/2026/1/e88626>

doi: [10.2196/88626](https://doi.org/10.2196/88626)

© Maria Carolina Bueno da Silva, Paula Bresciani M de Andrade, Henrique Min Ho Lee, Pedro Vinicius Alves Silva, Ana Cristina Ferreira, Cintia Pereira Kuss, Maria Gabriela de Almeida Rodrigues, Guilherme Alberto Sousa Ribeiro, Thiago Fellipe Ortiz Camargo, William Yang Chen Fan, Pedro Vieira Santana Netto, Giovanna de Souza Mendes, Gilberto Szarf, Rafael Maffei Loureiro, Ary Serpa Neto, Joselisa Péres Queiroz de Paiva, Jaqueline Driemeyer Correia Horvath. Originally published in JMIR Research Protocols (<https://www.researchprotocols.org>), 13.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Research Protocols, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.research-protocols.org>, as well as this copyright and license information must be included.