Open access Protocol

BMJ Open

# AI-assisted detection for chest X-rays (AID-CXR): a multi-reader multi-case study protocol

Farhaan Khan ,[1] Indrajeet Das,[2] Marusa Kotnik,[3] Louise Wing,[1] Edwin Van Beek,[4] John Murchison,[5] Jong Seok Ahn,[6] Sang Hyup Lee,[6] Ambika Seth,[6] Abdala Trinidad Espinosa Morgado ,[7] Howell Fu ,[1] Alex Novak ,[7] Nabeeha Salik,[8] Alan Campbell,[9] Ruchir Shah,[1] Fergus Gleeson,[10] Sarim Ather [1]

**To cite:** Khan F, Das I, Kotnik M, *et al*. AI-assisted detection for chest X-rays (AID-CXR): a multi-reader multi-case study protocol. *BMJ Open* 2024;**14**:e080554. doi:10.1136/bmjopen-2023-080554

► Prepublication history for this paper is available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2023-080554).

Received 04 October 2023
Accepted 25 October 2024

For numbered affiliations see end of article.

**Correspondence to**
Dr Farhaan Khan;
farhaan.a.khan@outlook.com

## ABSTRACT

**Introduction** A chest X-ray (CXR) is the most common imaging investigation performed worldwide. Advances in machine learning and computer vision technologies have led to the development of several artificial intelligence (AI) tools to detect abnormalities on CXRs, which may expand diagnostic support to a wider field of health professionals. There is a paucity of evidence on the impact of AI algorithms in assisting healthcare professionals (other than radiologists) who regularly review CXR images in their daily practice.

**Aims** To assess the utility of an AI-based CXR interpretation tool in assisting the diagnostic accuracy, speed and confidence of a varied group of healthcare professionals.

**Methods and analysis** The study will be conducted using 500 retrospectively collected inpatient and emergency department CXRs from two UK hospital trusts. Two fellowship-trained thoracic radiologists with at least 5 years of experience will independently review all studies to establish the ground truth reference standard with arbitration from a third senior radiologist in case of disagreement. The Lunit INSIGHT CXR tool (Seoul, Republic of Korea) will be applied and compared against the reference standard. Area under the receiver operating characteristic curve (AUROC) will be calculated for 10 abnormal findings: pulmonary nodules/mass, consolidation, pneumothorax, atelectasis, calcification, cardiomegaly, fibrosis, mediastinal widening, pleural effusion and pneumoperitoneum. Performance testing will be carried out with readers from various clinical professional groups with and without the assistance of Lunit INSIGHT CXR to evaluate the utility of the algorithm in improving reader accuracy (sensitivity, specificity, AUROC), confidence and speed (paired sample t-test). The study is currently ongoing with a planned end date of 31 December 2024.

**Ethics and dissemination** The study has been approved by the UK Healthcare Research Authority. The use of anonymised retrospective CXRs has been authorised by Oxford University Hospital's information governance teams. The results will be presented at relevant conferences and published in a peer-reviewed journal.

**Trial registration number** Protocol ID 310995-B (awaiting approval), ClinicalTrials.gov

### STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ This study will evaluate the impact of the artificial intelligence (AI) tool on diagnostic accuracy, speed and confidence, in its most realistic use-case, as an assistant to healthcare professionals rather than in isolation.

⇒ The study includes a relatively large number of readers (30) and the participants include a variety of non-radiologists (emergency medicine clinicians and radiographers) among the healthcare professionals that may benefit from AI assistance.

⇒ The prevalence of pathologies in the selected scans will be enriched in order to achieve statistical power to detect the impact of AI assistance, however, this will limit the immediate generalisability of results to real-life clinical performance.

⇒ All the readers will read the same cases first during the unaided and then the aided phase of the study, creating a risk of recall bias and learning effects that may result in improved reader performance.

## INTRODUCTION

Plain X-ray radiographs are the most common first-line imaging investigation in the diagnostic pathway of chest disease. In recent years, several Artificial Intelligence (AI) tools have become available to aid chest X-ray (CXR) reporting and have shown promise in identifying critical findings, mapping their location for clinician review and flagging abnormal scans for urgent attention.[1 2] The tools have demonstrated comparable sensitivity and specificity to radiologists in detecting important pulmonary pathologies such as nodules, consolidation and fibrosis.[2–7]

Current AI solutions are primarily designed as decision support tools rather than standalone diagnostic devices, and clinicians are likely to retain responsibility for accurate interpretations and diagnoses for the foreseeable future.[1 8] However, the tools provide added benefit by way of improved reader accuracy and confidence, thereby limiting errors of misinterpretation and subsequent patient mismanagement or harm.[9]

Increasing numbers of published studies evaluate the performance of AI tools against radiologists, and the impact of AI assistance in improving the accuracy of radiologists in chest X-ray interpretation.[6 7 10] However, there is relatively little research evaluating the impact of AI assistance on other healthcare professionals, such as emergency and general medicine physicians, who regularly interpret and act on CXR findings, particularly in the acute setting where a formal radiologist report may not be available until several hours or days later. Validating AI algorithms within the geographic setting in which they are intended to be used is also an important step in development as variable patient populations and imaging practices can impact performance.[11]

In this study, we aim to evaluate the impact of one such tool (Lunit INSIGHT CXR) that can detect and localise ten common abnormalities on chest X-rays namely: pulmonary nodules, consolidation, pneumothorax, atelectasis, calcification, cardiomegaly, fibrosis, mediastinal widening, pleural effusion and pneumoperitoneum.

The study will focus on CXRs from emergency department patients and hospital inpatients. This is a particularly challenging cohort of patients as they are often acutely unwell and demonstrate a high prevalence of, often multiple, abnormalities compared with the outpatient setting. The poor clinical state of some of these patients limits their ability to comply with the radiographer's instructions resulting in an increased number of technically suboptimal radiographs than in the outpatient setting. As a result, the radiographs are often acquired using anteroposterior or supine projection or using mobile imaging systems. There are also other confounding factors such as the presence of vascular lines, feeding tubes and external leads which make interpretation more challenging.

## STUDY AIMS AND HYPOTHESES

We aim to assess the impact of the INSIGHT CXR tool (Lunit Inc., Seoul, Republic of Korea) on the reporting accuracy, speed and confidence of a range of healthcare professionals of different seniority including radiologists, radiographers, and emergency and general physicians. We will also assess the impact of the tool on the clinical decision-making of the physicians reviewing the CXRs.

We hypothesise that the AI tool can improve the diagnostic accuracy and confidence of junior radiologists and non-radiologists in detecting common pathologies on CXRs to a degree akin to senior radiologists. Two key benefits arising from this are an improvement in timely, first-line clinical decision-making by less experienced clinicians and potentially a reduction in the need for a second review of these films by radiologists, thus alleviating their workload. Specifically, we aim to

1. Validate the accuracy of Lunit INSIGHT CXR in detecting pulmonary nodules, consolidation, pneumothorax, atelectasis, calcification, cardiomegaly, fibrosis, mediastinal widening, pleural effusion and pneumoperitoneum on a retrospective dataset of 500 inpatient and emergency department chest X-ray images (primary).
2. Determine the effect on the accuracy of chest X-ray interpretation by general radiologists, emergency department (ED) physicians, intensive care unit (ICU) physicians, general medicine physicians and radiographers for the above abnormalities, with the assistance of Lunit INSIGHT CXR (primary).
3. Measure the time taken by the above healthcare to evaluate images, and their diagnostic confidence therein, with and without input from the AI tool (secondary).
4. Explore which imaging factors influence the reporting accuracy of healthcare professionals and algorithm performance, for example, type abnormality, size of abnormality, posteroanterior (PA)/anteroposterior (AP) view, mobile/fixed X-ray and presence of multiple abnormalities (secondary).
5. Explore the utility of the AI tool in changing the course of reporting workflow and clinical management (secondary).

## METHODS

### Study design

The study will employ a fully crossed, multi-reader multicase design. The Oxford Acute Care patients or the public were not involved in the design, conduct, reporting or dissemination plans of the study. The study period is from 31 March 2024 to 31 December 2024.

### Case selection

500 CXRs in patients over 18 years of age in the acute hospital setting will be retrospectively identified by the clinical and picture archiving and communication system (PACS)/information technology (IT) team through a database search of the Computerised Radiology Information System at two large UK teaching hospitals (Oxford University Hospitals NHS Foundation Trust and NHS Lothian Health Board). CXR images will be extracted and de-identified along with their associated formal radiology reports. The case mix will include 100 normal CXR films along with at least 40 from each of the following 10 abnormalities:

1. Lung nodule/mass
2. Consolidation
3. Pneumothorax
4. Atelectasis
5. Calcification
6. Cardiomegaly
7. Fibrosis
8. Mediastinal widening
9. Pleural effusion
10. Pneumoperitoneum

A random sampling approach will be taken to ensure that the cases represent the natural spectrum of disease severity. A subset of images may demonstrate multiple of the above abnormalities.

Inclusion criteria for cases:

- Individuals undergoing CXR in the hospital setting (inpatient or ED).
- Age ≥ 18 years.

Exclusion criteria:

- Lateral projections without accompanying AP or PA views.

## Setting

Cases will be selected from the following hospital sites:

- Oxford University Hospitals NHS Foundation Trust
- NHS Lothian Health Board

The reads will be performed using a web-based image viewing platform (www.raiqc.com) which combines a DICOM (Digital Imaging and Communications in Medicine) viewer with a structured reporting template.

## Reader selection

30 readers will be selected from the following five clinical specialty groups:

- ED
- Adult ICU
- Adult general medicine (AGM)
- Radiographers (Rad)
- General radiologists

Each specialty group consists of six members of ranked seniority. For the physicians this consists of:

- Two 'Juniors' (within 4 years after graduating medical school that is, F1-ST2 grade).
- Two 'Middle Grades' (between 5 and 8 years after graduating medical school that is, Registrar ST3-6 grade).
- Two consultants.

For the radiographers, this consists of:

- Two 'Junior/Newly qualified radiographers' (up to 18 months experience post qualification).
- Two 'Mid-experience radiographers' (approx. 3 years' experience).
- Two 'Reporting radiographers' (5+ years' experience).

Five additional readers, one from each clinical specialty group, will be selected as a control group. They will perform unaided reads in both phases and their results will be used to assess for any improvement due to learning effects.

Inclusion criteria:

- General radiologists/radiographers/physicians who review CXRs as part of their routine clinical practice.

Exclusion criteria:

- Thoracic radiologists.
- Non-radiology physicians with previous formal postgraduate CXR reporting training.
- Non-radiology physicians with a previous career in radiology, respiratory medicine or thoracic surgery to registrar or consultant level.

## Reader training

Prior to commencing each session of the study, the readers will be asked to review five practice cases to familiarise themselves with the use of the study platform as well as the output of the Lunit INSIGHT CXR tool.

## Ground truthing

Two consultant thoracic radiologists will independently review the images to establish the 'ground truth' findings on the CXRs. Where a consensus is reached, it will serve as the reference standard. In the case of disagreement, a third senior thoracic radiologist's opinion (>20 years' experience) will undertake arbitration. The arbitration will be done at a finding level and the arbitrator will only review the findings where there is a disagreement between the initial ground truthers. The ground truthers will be asked to mark the location of the abnormality with a region of interest. A difficulty score will be assigned to each abnormality by the ground truthers using a 5-point Likert scale (1 being easy/obvious to 5 being hard/poorly visualised).

Where a contemporaneous chest CT scan is available (scan performed within 2 weeks of the CXR), an analysis will be performed using the results of the CT scan as the reference standard.

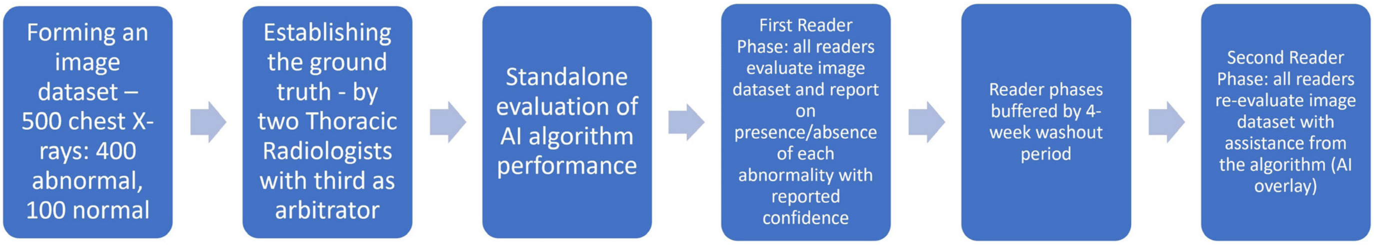
## Performance of AI algorithm

First, a standalone evaluation of the Lunit INSIGHT CXR algorithm will be performed comparing it to the reference standard. Continuous probability scores from the algorithm will be used for the ROC analyses, while binary classification results with a predefined operating cut-off will be used for the evaluation of sensitivity and specificity.

## Performance of readers with and without AI assistance

To assess the value of the algorithm as a second reader, observer performance testing will be carried out by a reader panel composed of multiple clinical staff from various specialities (see section above on reader selection). The study will include two sessions (with and without AI overlay), with all 30 readers reviewing all 500 CXR cases each time separated by a washout period of 4 weeks to mitigate recall bias. The cases will be randomised between the two reads and for every reader. This is summarised in figure 1.

In the first session, readers blinded to the ground truth and without AI assistance will review the CXRs and provide an opinion on the presence or absence of the abnormalities listed above. For each case, the ground-truthers and the readers will be asked to select all the possible options that an abnormality could be categorised. Where a case is deemed to have a positive finding, the readers will be asked to click on the image to indicate the abnormality location. The time taken for each read will be automatically recorded. They will also provide a confidence level in their diagnosis on a 5-point Likert scale. A precis regarding the patient's clinical status will be given to the readers. Based on the assessment of the


Forming an image dataset – 500 chest X-rays: 400 abnormal, 100 normal

Establishing the ground truth - by two Thoracic Radiologists with third as arbitrator

Standalone evaluation of AI algorithm performance

First Reader Phase: all readers evaluate image dataset and report on presence/absence of each abnormality with reported confidence

Reader phases buffered by 4-week washout period

Second Reader Phase: all readers re-evaluate image dataset with assistance from the algorithm (AI overlay)

**Figure 1** Reader study summary flowchart. AI, artificial intelligence.

CXR and available clinical information, readers will be asked to select what further action is required from the following:

- No further action/discharge.
- Image review by a senior colleague or radiologist.
- Further radiological investigation (if yes then select from the options below)
  - Follow-up CXR
  - CT
  - Ultrasound
  - Other (please state)
- Initiate treatment (if yes then select from the options below)
  - Pharmacological intervention
  - Invasive intervention (eg, chest drain insertion)
  - Other (please state)
- Refer to another specialty

Readers will also be asked 'Do you feel that this CXR requires a formal radiologist report?' with the following options:

- Yes
- No
- N/A (I'm a radiologist)

In the second session, all readers will re-evaluate the CXR cases in a randomised order, remaining blinded to the ground truth. Alongside the original CXR image, they will also be provided with the output of the Lunit INSIGHT CXR algorithm. The output will include classification results and heat maps overlaid on any abnormality identified by the algorithm. The performance (sensitivity, specificity), speed and confidence of readers between the two sessions will be compared, to evaluate whether there is any improvement in performance with the utilisation of the AI algorithm. The impact of the algorithm on clinical management decisions will be evaluated by comparing the variability of the decisions between junior and senior readers.

Readers will also complete surveys about the perceived algorithm usability and utility after completing the second session of the study.

The two sessions will be buffered by a 4-week washout period per reader, with 3 weeks allocated to undertake each set of 'reads' of the 500 CXR images.

### Sample size and power calculation

The sample size was calculated using the 'Multi-Reader Sample Size Sample Size Program for Diagnostic Studies' to estimate power for the number of readers cases in our study (https://perception.lab.uiowa.edu/power-sample-size-estimation). Parameter values for the error variances and the covariances were taken from a previous multi-reader, multi-case study on detecting pneumothoraces. 30 readers, reading 500 cases yields 85% power to detect a difference in accuracy of 10% with a type 1 error of 5%.

### Patient and public involvement

None. Patients or the public were not involved in the design, conduct, reporting or dissemination plans of the study.

## MEASURES AND ANALYSES

### Outcome measures

Lunit INSIGHT CXR performance: sensitivity, specificity and AUROC.

Reader performance: sensitivity, specificity with versus without AI assistance.

Reader confidence: self-reported diagnostic confidence on a 5-point Likert scale, with versus without AI assistance.

Reader speed: mean time taken to review a scan, with versus without AI assistance.

### Statistical analyses

The performance of the Lunit INSIGHT CXR algorithm will be compared with the ground truth. Continuous probability scores from the algorithm will be used for the ROC analyses, while binary classification results with three different operating cut-offs will be used for the evaluation of sensitivity and specificity.

Reader performance (sensitivity, specificity), reader confidence and reader speed (paired sample t-test) with and without AI assistance will be compared. The main analysis will consider the pooled performance of all professional groups across all cases. Subgroup analyses will be performed comparing:

- Professional groups (general radiologist vs ED clinician vs ICU clinician vs AGM clinician vs radiographer).
- Senior versus mid-experience vs junior.
- Pathological finding.
- Difficulty of abnormality as determined by ground truthers.

Results from the qualitative reader survey about actioning the image will be collated and used to explore the perceived utility and usability of the algorithm.

Additional data to be provided on a per-image basis for statistical sub-analyses includes:

- Image view (AP/PA).
- System type (Mobile/Fixed).
- Patient gender (M/F).
- System vendor.
- Patient age.
- Referral source (ED, inpatient, ICU).

## Ethics and dissemination

The study has been approved by the UK Health Research Authority (IRAS ID: 310995). The use of anonymised retrospective CXR images has been authorised by the Caldicott Guardian and information governance team at Oxford University Hospitals NHS Foundation Trust and NHS Lothian Health Board. Readers will provide written informed consent and will be able to withdraw at any time.

The results of the study will be presented at relevant conferences and published in peer-reviewed journals. The detailed study protocol will be freely available on request to the corresponding author. Further dissemination strategy will be strongly guided by our patient and public involvement and engagement (PPIE) activities. This will be based on co-productions between patient partners and academics and will involve media pieces (mainstream and social media) as well as communication through charity partners.

**Author affiliations**
[1]Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[2]University Hospitals of Leicester NHS Trust, Leicester, UK
[3]Addenbrooke's Hospital, Cambridge, UK
[4]Edinburgh Imaging, Queen's Medical Research Institute, University of Edinburgh, Edinburgh, UK
[5]Royal Infirmary of Edinburgh, Edinburgh, UK
[6]Lunit Inc, Gangnam-gu, Seoul, Korea (the Republic of)
[7]Emergency Medicine Research Oxford, Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[8]RAIQC Ltd, Oxford, UK
[9]Radiology, University College London Hospitals NHS Foundation Trust, London, UK
[10]Churchill Hospital, Oxford, UK

**X** Alan Campbell @DrAlanCampbell

**Contributors** SA, AN and FG led the design of the protocol, with contributions from FK, ID, MK, EVB, JM, AEM, HF, NS, AC and RS. SA, RS, FK, NS and AC reviewed the image dataset. ID and MK are the primary ground-truthers, with arbitration from LW. NS manages the online reading platform and will be aiding in data collection and management. AEM registered the study and coordinated reader recruitment and data collection. JSA, SHL and AS informed the study team of the workings of the algorithm and how to interpret them. They will be involved in processing the CXRs for AI analysis. RS leads the statistical analysis plan. SA performed the simulations estimating statistical power for the study. FK, HF, AN, and SA wrote the manuscript. SA is the guarantor.

**Funding** The study is being funded by a commercial research grant amounting to £5,500.00 from Lunit Inc. Jong Seok Ahn, Sang Hyup Lee and Ambika Seth are employees of Lunit Inc.

**Competing interests** None declared.

**Patient and public involvement** Patients and/or the public were not involved in the design, conduct, reporting or dissemination plans of this research.

**Patient consent for publication** Not applicable.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**ORCID iDs**
Farhaan Khan http://orcid.org/0000-0003-2308-414X
Abdala Trinidad Espinosa Morgado http://orcid.org/0000-0003-0967-3554
Howell Fu http://orcid.org/0000-0002-0582-9518
Alex Novak http://orcid.org/0000-0002-5880-8235
Sarim Ather http://orcid.org/0000-0001-9614-5033

## REFERENCES

1 Jones CM, Buchlak QD, Oakden-Rayner L, *et al*. Chest radiographs and machine learning - Past, present and future. *J Med Imaging Radiat Oncol* 2021;65:538–44.
2 Ahmad HK, Milne MR, Buchlak QD, *et al*. Machine Learning Augmented Interpretation of Chest X-rays: A Systematic Review. *Diagnostics (Basel)* 2023;13:743.
3 van Beek EJR, Ahn JS, Kim MJ, *et al*. Validation study of machine-learning chest radiograph software in primary and emergency medicine. *Clin Radiol* 2023;78:1–7.
4 Kundu R, Das R, Geem ZW, *et al*. Pneumonia detection in chest X-ray images using an ensemble of deep learning models. *PLoS One* 2021;16:e0256630.
5 Matsumoto T, Kodera S, Shinohara H, *et al*. Diagnosing Heart Failure from Chest X-Ray Images Using Deep Learning. *Int Heart J* 2020;61:781–6.
6 Hillis JM, Bizzo BC, Mercaldo S, *et al*. Evaluation of an Artificial Intelligence Model for Detection of Pneumothorax and Tension Pneumothorax in Chest Radiographs. *JAMA Netw Open* 2022;5:e2247172.
7 Homayounieh F, Digumarthy S, Ebrahimian S, *et al*. An Artificial Intelligence-Based Chest X-ray Model on Human Nodule Detection Accuracy From a Multicenter Study. *JAMA Netw Open* 2021;4:e2141096.
8 Artificial intelligence for analysing chest x-ray images. NICE; 2022. Available: https://www.nice.org.uk/advice/mib292
9 Wilson C. X-ray misinterpretation in urgent care: where does it occur, why does it occur, and does it matter? *N Z Med J* 2022;135:49–65.
10 Wu JT, Wong KCL, Gur Y, *et al*. Comparison of Chest Radiograph Interpretations by Artificial Intelligence Algorithm vs Radiology Residents. *JAMA Netw Open* 2020;3:e2022779.
11 Zech JR, Badgeley MA, Liu M, *et al*. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med* 2018;15:e1002683.