BMJ Open

AI Assisted Detection for Chest X-rays (AID-CXR): a Multi-Reader Multi-Case Study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2023-080554
Article Type:	Protocol
Date Submitted by the Author:	04-Oct-2023
Complete List of Authors:	Khan, Farhaan; Oxford University Hospitals NHS Foundation Trust, Radiology Das, Indrajeet; University Hospitals of Leicester NHS Trust Kotnik, Marusa ; Addenbrooke's Hospital Van Beek, Edwin; Edinburgh Imaging, Queen's Medical Research Institute, University of Edinburgh Murchison, John; Royal Infirmary of Edinburgh Ahn, Jong; Lunit Inc Lee, Sang; Lunit Inc Seth, Ambika; Lunit Inc Espinosa Morgado, Abdala Trinidad; Oxford University Hospitals NHS Foundation Trust, Emergency Medicine Research Oxford Fu, Howell; Oxford University Hospitals NHS Foundation Trust, Clinical Radiology Novak, Alex; Oxford University Hospitals NHS Foundation Trust, Emergency Medicine Research Oxford Salik, Nabeeha; RAIQC Ltd Campbell, Alan; University College London Hospitals NHS Foundation Trust, Radiology Shah, Ruchir; Oxford University Hospitals NHS Foundation Trust, Clinical Radiology Gleeson, Fergus; Churchill Hospital Ather, Sarim; Oxford University Hospitals NHS Foundation Trust
Keywords:	Chest imaging < RADIOLOGY & IMAGING, Diagnostic Imaging, RESPIRATORY MEDICINE (see Thoracic Medicine), CARDIOLOGY

AI Assisted Detection for Chest X-rays (AID-CXR): a Multi-Reader Multi-Case Study

ABSTRACT

Introduction: A chest X-ray (CXR) is amongst the most common imaging investigations performed worldwide. Advances in machine learning and computer vision technologies have led to 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 physicians (other than radiologists) who regularly review 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.

Materials and methods: 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' experience will independently review all studies to establish the ground truth reference standard. 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, positive predictive value and negative predictive value), confidence and speed (paired sample t-test).

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 Hospitals information governance teams. The results will be presented at relevant conferences and published in a peer-reviewed journal.

Strengths and limitations of this study:

- This study will evaluate the impact of the AI tool on diagnostic accuracy, speed, and confidence, in its most realistic use-case, as an assistant to healthcare professionals rather than in isolation.
- It includes non-radiologists (EM clinicians and radiographers) among the healthcare professionals that may directly benefit from AI support tools.
- The prevalence of pathologies in the selected scans will be enriched in order to achieve statistical power to detect the impact of AI assistance. Although necessary to facilitate an important evaluation of diagnostic accuracy, this will limit the immediate generalisability of results to real-life clinical performance.

INTRODUCTION

X-rays 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 stand-alone 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 upon 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-ray 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 to the outpatient setting. The poor clinical state of some of these patients, limits their ability to comply with 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, 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 chest x-ray 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 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 accuracy of chest X-ray interpretation by general radiologists, ED physicians, 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 reporting accuracy of healthcare professionals and algorithm performance, e.g. type abnormality, size of abnormality, PA/AP view, mobile/fixed X-ray, 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 employs a fully crossed, multi-reader multi-case (MRMC) design. Patients or the public were not involved in the design, conduct, reporting, or dissemination plans of the study.

Case Selection: 500 CXRs in patients over 18 years of age in the acute hospital setting were retrospectively identified by the clinical and PACS/IT team through a database search of the Computerised Radiology Information System (CRIS) 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 ten 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 subset of images may demonstrate multiple of the above abnormalities.

Diagnoses on each of the 500 CXR images will be confirmed by consensus opinion of two thoracic radiologists.

Inclusion criteria for cases:

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

Exclusion criteria:

- Lateral projections without accompanying anteroposterior (AP) or posteroanterior (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 viewer with a structured reporting template.

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

- emergency medicine (ED)
- adult intensive care (ICU)
- adult general medicine (AGM)
- radiographers (Rad)
- general radiologists

Each specialty group consists of 6 members of ranked seniority. For the physicians this consists of:

- Two 'Juniors' (within 4 years after graduating medical school i.e. F1-ST2 grade)
- Two 'Middle Grades' (between 5-8 years after graduating medical school i.e. 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)

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 5 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. 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).

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 score from the algorithm will be utilised for the ROC analyses, while binary classification results with a predefined operating cut-off will be used for evaluation of sensitivity, specificity, positive predictive value, and negative predictive value.

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

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

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 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 utilisation of the AI algorithm. The impact of the algorithm on the 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%.

MEASURES AND ANALYSES

Outcome measures: Lunit INSIGHT CXR performance: sensitivity, specificity and Area Under Receiver Operating Characteristic Curve (AUROC).

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

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

Reader speed: mean time taken to review a scan, with vs 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 utilized for the ROC analyses, while binary classification results with three different operating cut-offs will be used for evaluation of sensitivity, specificity, positive predictive value, and negative predictive value.

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 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 vs 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)

DATA DE-IDENTIFICATION AND MANAGEMENT

CXR images selected for the study will be anonymised in accordance with the information governance protocol at the participating hospitals. Access to the images will be controlled via the study platform using separate user accounts for each reader.

All study data will be entered into a password-protected and secure database. Individual reader accuracy scores will be anonymised, and the study team will not have access to the identifying link between the participants' personal details and the data. Data about the participants' level of experience and professional group will be retained to allow group comparison.

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 upon request to the corresponding author. Further dissemination strategy will be strongly guided by our 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.

FUNDING STATEMENT AND COMPETING INTERESTS

The study is being funded by a commercial research grant from Lunit Inc.

Jong Seok Ahn, Sang Hyup Lee and Ambika Seth are employees of Lunit Inc.

REFERENCES

1. Jones CM, Buchlak QD, Oakden-Rayner L, Milne M, Seah J, Esmaili N, et al. Chest radiographs and machine learning - Past, present and future. J Med Imaging Radiat Oncol. 2021 Aug;65(5):538–44.

2. Ahmad HK, Milne MR, Buchlak QD, Ektas N, Sanderson G, Chamtie H, et al. Machine Learning Augmented Interpretation of Chest X-rays: A Systematic Review. Diagn Basel Switz. 2023 Feb 15;13(4).

3. van Beek EJR, Ahn JS, Kim MJ, Murchison JT. Validation study of machine-learning chest radiograph software in primary and emergency medicine. Clin Radiol. 2023 Jan;78(1):1–7.

4. Kundu R, Das R, Geem ZW, Han GT, Sarkar R. Pneumonia detection in chest X-ray images using an ensemble of deep learning models. PloS One. 2021;16(9):e0256630.

5. Matsumoto T, Kodera S, Shinohara H, Ieki H, Yamaguchi T, Higashikuni Y, et al. Diagnosing Heart Failure from Chest X-Ray Images Using Deep Learning. Int Heart J. 2020 Jul 30;61(4):781–6.

6. Hillis JM, Bizzo BC, Mercaldo S, Chin JK, Newbury-Chaet I, Digumarthy SR, et al. Evaluation of an Artificial Intelligence Model for Detection of Pneumothorax and Tension Pneumothorax in Chest Radiographs. JAMA Netw Open. 2022 Dec 1;5(12):e2247172.

7. Homayounieh F, Digumarthy S, Ebrahimian S, Rueckel J, Hoppe BF, Sabel BO, et al. An Artificial Intelligence-Based Chest X-ray Model on Human Nodule Detection Accuracy From a Multicenter Study. JAMA Netw Open. 2021 Dec 1;4(12):e2141096.

8. Artificial intelligence for analysing chest X-ray images [Internet]. NICE (The National Institute for Health and Care Excellence); 2022. Available from: 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 Apr 1;135:49–65.

10. Wu JT, Wong KCL, Gur Y, Ansari N, Karargyris A, Sharma A, et al. Comparison of Chest Radiograph Interpretations by Artificial Intelligence Algorithm vs Radiology Residents. JAMA Netw Open. 2020 Oct 1;3(10):e2022779.

11. 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. 2018 Nov;15(11):e1002683.

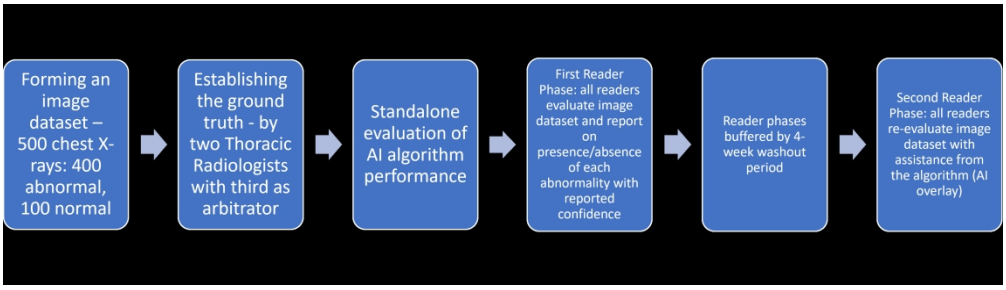


Figure 1: Reader study summary flowchart

823x231mm (130 x 130 DPI)

LUNIT INSIGHT CXR READER STUDY PROTOCOL

SCHEDULE OF WORK

1. STUDY INFORMATION

Study Title: Diagnostic accuracy study comparing the relative efficacy of junior clinicians in the detection of abnormalities on Chest X-Ray with and without a machine-learning algorithm

Study Plan Version: 1

Investigator Name: Dr. Sarim Ather

Institution Name: Oxford University Hospitals

Estimated Number of Subjects/Cases: 30 readers / 500

Type of Subject/Cases: Retrospectively collected AP/PA Chest X-rays

Estimated Study Duration: 12 months

2. STUDY PLAN

2.1. Background

Lunit Insight CXR is a CE marked Artificial Intelligence (AI) solution for chest X-ray (CXR) interpretation. It can detect 10 common abnormalities seen on the CXR including nodules, consolidation, pneumothorax, atelectasis, calcification, cardiomegaly, fibrosis, mediastinal widening, pleural effusion and pneumoperitoneum. In the acute setting, CXRs are often interpreted by junior clinicians with limited exposure to CXR interpretation training. Due to workload pressures, formal radiologist report is not issued until several hours or days later. As a result, junior doctors are required to make decisions on the immediate management of time sensitive pathologies such as pneumothorax and pneumoperitoneum based on their own CXR interpretation.

The study aims to demonstrate that AI can be used the improve the accuracy and the confidence of non-radiologist clinician groups, who are more likely to undertake the first-line of reporting of CXRs in routine clinical practice. This study will explore this aspect of AI-led imaging further, by directly comparing the pooled reporting accuracy of different clinician groups against that of the Lunit INSIGHT CXR tool. This will be achieved through using the Report And Image Quality Control (RAIQC) platform for presenting and recording the reports of the clinicians on a pre-annotated bank of images showing both normal and abnormal CXRs encountered in routine clinical practice.

Study Hypothesis and Aims

1. AI can help improve the diagnostic accuracy of non-radiologists, detecting CXR abnormalities (e.g. Nodule, Consolidation, Pneumothorax, Atelectasis, Calcification, Cardiomegaly, Fibrosis, Mediastinal Widening, Pleural Effusion and Pneumoperitoneum)

2. AI can help improve the diagnostic accuracy of non-radiologist to a similar level to that of radiologists.

3. AI can help improve the confidence of non-radiologists in interpreting CXRs

4. AI can improve the first line clinical decisions of the non-radiologists.

5. AI can reduce the requirement for images to be 2nd reported by radiologists

2.2.1 Objectives

- To validate the Lunit INSIGHT CXR on a retrospective dataset of up to 500 images
- To assess clinician performance with and without the output from the AI algorithm

2.2. Methods

Subject/Case Selection

500 retrospectively selected on patients aged 18 years or older will be collected, annotated and de-identified along with their radiology reports. The cases mix will include 100 normal films along with 40 each with the following 10 abnormalities:

- Lung nodule/mass
- Consolidation
- Pneumothorax
- Atelectasis
- Calcification
- Cardiomegaly
- Fibrosis
- Mediastinal Widening
- Pleural Effusion
- Pneumoperitoneum

A subset of images may demonstrate multiple of the above abnormalities. The diagnosis will be confirmed by a consensus opinion of two thoracic radiologists. Potentially suitable images will be identified by the clinical and/or PACS & IT team by searching the CRIS database in Oxford and Edinburgh transferred to the RAIQC platform via a secure FTP.

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

- Image View (AP/PA)
- System Type (Mobile/Fixed)
- Patient Gender (M/F)
- System Vendor

- Patient Age
- Referral source (ED, inpatient, ICU)

Data collection methods

Imaging data will be fully de-identified in the OUH NHS Foundation Trust, within the Radiology department. This de-identified data will have been transferred to the RAIQC platform via a secure FTP and uploaded to secure servers. RAIQC will have access to the Lunit INSIGHT CXR tool to undertake inference of the collected images.

The data will be reviewed by two thoracic radiologists to establish ground truth reference standard. Each abnormality will be categorized on a 1 to 5 scale, from easy/obvious as 1 to hard/poorly visualized as 5.

Data analysis methods

Standalone evaluation of algorithm performance:

- Performance of Lunit algorithm will be compared with the reference standard. Continuous probability score from the algorithm will be utilised for the ROC analyses, while binary classification results with three different operating cutoffs will be utilized for evaluation of sensitivity, specificity, positive predictive value, and negative predictive value.

Observer performance test:

- For assessment of the value of the algorithm as a second reader, observer performance test will be performed by a reader panel comprised of multiple clinicians from various specialties. The test will include two sessions. In the first session, all readers will review chest X-rays and determine whether there is an abnormality and provide their confidence on a 5-point Likert score. The readers will also be given a precis about the clinical status of the patient.
    - Based on the assessment of this chest X-ray and available clinical information, what further action is required?
        - No further action
        - Image review by a senior colleague or radiologist
        - Further radiological investigation
            - Follow-up CXR
            - CT
            - Ultrasound
            - Other (please state)
        - Initiate treatment
            - Pharmacological intervention
            - Invasive intervention (e.g. chest drain insertion)
            - Other (please state)
        - Refer to another specialty
        - Do you feel that this CXR requires a formal radiology report?

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

- In the second session, all readers will re-evaluate the images with results from the algorithm (AI overlay and AI score). Classification results and heatmaps for localization of abnormality will be provided. The performance (sensitivity, specificity, positive predictive value, and negative predictive value) of readers in two sessions will be compared, to evaluate whether there is any improvement in performance with utilization of algorithm's result.

- Sessions will be buffered by a 4-week washout period per reader, with 4 weeks allocated to undertake each set of 'reads' of the images.

Clinical Readers

Clinical readers will be drawn from the following clinical groups:

- 30 clinical readers in total drawn from 4 clinical specialty groups:
  - Emergency Medicine (ED)
  - Adult Intensive Care (ICU)
  - Adult General Medicine (AGM)
  - Radiographers (Rad)
  - General radiologists

- Each specialty group will consist of 5 members of ranked seniority for the clinicians:
  - Junior (F1 - ST2) x 2
  - Middle Grade (Registrar ST 3-7) x 2
  - Consultant x 2

For the radiographers:

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

- This will be enabled/coordinated through the Thames Valley Emergency medicine Research Network (TAVERN) and will be assigned to an OUH ED Research Fellow who will liaise with TAVERN members at each site to approach clinicians.

- Each reader to read all 500 images on 2 separate sessions (without/with AI overlay), with a washout period in between.

Training

- Before the reader study sessions, a training module will be developed on RAIQC platform to inform the readers about the AI output as well as the definition of the metrics and chest pathologies being recorded.

BMJ Open

AI Assisted Detection for Chest X-rays (AID-CXR): a Multi-Reader Multi-Case Study Protocol

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2023-080554.R1
Article Type:	Protocol
Date Submitted by the Author:	12-Jun-2024
Complete List of Authors:	Khan, Farhaan; Oxford University Hospitals NHS Foundation Trust, Radiology Das, Indrajeet; University Hospitals of Leicester NHS Trust Kotnik, Marusa ; Addenbrooke's Hospital Wing, Louise; Oxford University Hospitals NHS Foundation Trust Van Beek, Edwin; Edinburgh Imaging, Queen's Medical Research Institute, University of Edinburgh Murchison, John; Royal Infirmary of Edinburgh Ahn, Jong; Lunit Inc Lee, Sang; Lunit Inc Seth, Ambika; Lunit Inc Espinosa Morgado, Abdala Trinidad; Oxford University Hospitals NHS Foundation Trust, Emergency Medicine Research Oxford Fu, Howell; Oxford University Hospitals NHS Foundation Trust, Clinical Radiology Novak, Alex; Oxford University Hospitals NHS Foundation Trust, Emergency Medicine Research Oxford Salik, Nabeeha; RAIQC Ltd Campbell, Alan; University College London Hospitals NHS Foundation Trust, Radiology Shah, Ruchir; Oxford University Hospitals NHS Foundation Trust, Clinical Radiology Gleeson, Fergus; Churchill Hospital Ather, Sarim; Oxford University Hospitals NHS Foundation Trust
Primary Subject Heading:	Radiology and imaging
Secondary Subject Heading:	Respiratory medicine, Cardiovascular medicine
Keywords:	Chest imaging < RADIOLOGY & IMAGING, Diagnostic Imaging, RESPIRATORY MEDICINE (see Thoracic Medicine)

AI Assisted Detection for Chest X-rays (AID-CXR): a Multi-Reader Multi-Case Study Protocol

ABSTRACT

Introduction: A chest X-ray (CXR) is the most common imaging investigations performed worldwide. Advances in machine learning and computer vision technologies have led to 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.

Materials and methods: 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' experience will independently review all studies to establish the ground truth reference standard with arbitration from a 3rd 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).

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 Hospitals information governance teams. The results will be presented at relevant conferences and published in a peer-reviewed journal.

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.

- It will be the first UK-based multicentre validation of an AI for CXRs trained on a large data set.

- 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. Although necessary to facilitate an important evaluation of diagnostic accuracy, this will limit the immediate generalisability of results to real-life clinical performance.

  All the readers will read the same cases during the unaided and aided phase of the study creating a risk of recall bias. The unaided phase will be conducted first with a risk that a learning effect may result in an improved performance of readers.

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 stand-alone 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 upon 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-ray 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 to the outpatient setting. The poor clinical state of some of these patients, limits their ability to comply with 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, 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 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 accuracy of chest X-ray interpretation by general radiologists, ED physicians, 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 reporting accuracy of healthcare professionals and algorithm performance, e.g. type abnormality, size of abnormality, PA/AP view, mobile/fixed X-ray, 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 multi-case (MRMC) design. The Oxford Acute Care patients or public were not involved in the design, conduct, reporting, or dissemination plans of the study. The study period is from 31st October 2023 to 31st 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 PACS/IT team through a database search of the Computerised Radiology Information System (CRIS) 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 ten 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 anteroposterior (AP) or posteroanterior (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 viewer with a structured reporting template.

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

- emergency medicine (ED)
- adult intensive care (ICU)
- adult general medicine (AGM)
- radiographers (Rad)
- general radiologists

Each specialty group consists of 6 members of ranked seniority. For the physicians this consists of:

- Two 'Juniors' (within 4 years after graduating medical school i.e. F1-ST2 grade)
- Two 'Middle Grades' (between 5-8 years after graduating medical school i.e. 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 5 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 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 score from the algorithm will be utilised for the ROC analyses, while binary classification results with a predefined operating cut-off will be used for 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. 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 CXR and available clinical information, readers will be asked to select what further action is required from the following:

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

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 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 utilisation of the AI algorithm. The impact of the algorithm on the 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 Area Under Receiver Operating Characteristic Curve (AUROC).
Reader performance: sensitivity, specificity with vs without AI assistance.
Reader confidence: self-reported diagnostic confidence on a 5-point Likert scale, with vs without AI assistance.
Reader speed: mean time taken to review a scan, with vs 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 utilized for the ROC analyses, while binary classification results with three different operating cut-offs will be used for 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 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 vs 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 upon request to the corresponding author. Further dissemination strategy will be strongly guided by our 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.

FIGURE LEGENDS

Figure 1: Reader study summary flowchart

CONTRIBUTORSHIP STATEMENT

SA, AN and FG led the design of the protocol, with contributions from FK, ID, MK, EvB, AM, AE, HF, NS, AC and RS. SA, RS, FK and NS and AC and 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. AE registered the study and coordinates 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.

COMPETING INTERESTS

There are no competing interests to declare. The funders from Lunit Inc. have no input into the study design, analysis, reporting or decision to publish.

FUNDING STATEMENT

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.

DATA DE-IDENTIFICATION AND SHARING STATEMENT

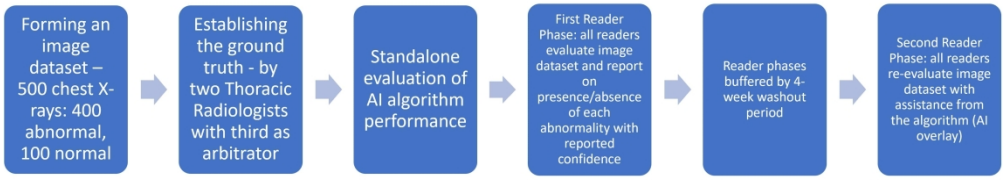
CXR images selected for the study will be anonymised in accordance with the information governance protocol at the participating hospitals. Access to the images will be controlled via the study platform using separate user accounts for each reader.

All study data will be entered into a password-protected and secure database. Individual reader accuracy scores will be anonymised, and the study team will not have access to the identifying link between the participants' personal details and the data. Data about the participants' level of experience and professional group will be retained to allow group comparison.

REFERENCES

1. Jones CM, Buchlak QD, Oakden-Rayner L, Milne M, Seah J, Esmaili N, et al. Chest radiographs and machine learning - Past, present and future. J Med Imaging Radiat Oncol. 2021 Aug;65(5):538–44.
2. Ahmad HK, Milne MR, Buchlak QD, Ektas N, Sanderson G, Chamtie H, et al. Machine Learning Augmented Interpretation of Chest X-rays: A Systematic Review. Diagn Basel Switz. 2023 Feb 15;13(4).
3. van Beek EJR, Ahn JS, Kim MJ, Murchison JT. Validation study of machine-learning chest radiograph software in primary and emergency medicine. Clin Radiol. 2023 Jan;78(1):1–7.
4. Kundu R, Das R, Geem ZW, Han GT, Sarkar R. Pneumonia detection in chest X-ray images using an ensemble of deep learning models. PloS One. 2021;16(9):e0256630.
5. Matsumoto T, Kodera S, Shinohara H, Ieki H, Yamaguchi T, Higashikuni Y, et al. Diagnosing Heart Failure from Chest X-Ray Images Using Deep Learning. Int Heart J. 2020 Jul 30;61(4):781–6.
6. Hillis JM, Bizzo BC, Mercaldo S, Chin JK, Newbury-Chaet I, Digumarthy SR, et al. Evaluation of an Artificial Intelligence Model for Detection of Pneumothorax and Tension Pneumothorax in Chest Radiographs. JAMA Netw Open. 2022 Dec 1;5(12):e2247172.

7. Homayounieh F, Digumarthy S, Ebrahimian S, Rueckel J, Hoppe BF, Sabel BO, et al. An Artificial Intelligence-Based Chest X-ray Model on Human Nodule Detection Accuracy From a Multicenter Study. JAMA Netw Open. 2021 Dec 1;4(12):e2141096.
8. Artificial intelligence for analysing chest X-ray images [Internet]. NICE (The National Institute for Health and Care Excellence); 2022. Available from: 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 Apr 1;135:49–65.
10. Wu JT, Wong KCL, Gur Y, Ansari N, Karargyris A, Sharma A, et al. Comparison of Chest Radiograph Interpretations by Artificial Intelligence Algorithm vs Radiology Residents. JAMA Netw Open. 2020 Oct 1;3(10):e2022779.
11. 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. 2018 Nov;15(11):e1002683.



Reader study summary flowchart

1486x418mm (72 x 72 DPI)

BMJ Open

AI Assisted Detection for Chest X-rays (AID-CXR): a Multi-Reader Multi-Case Study Protocol

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2023-080554.R2
Article Type:	Protocol
Date Submitted by the Author:	15-Oct-2024
Complete List of Authors:	Khan, Farhaan; Oxford University Hospitals NHS Foundation Trust, Radiology Das, Indrajeet; University Hospitals of Leicester NHS Trust Kotnik, Marusa ; Addenbrooke's Hospital Wing, Louise; Oxford University Hospitals NHS Foundation Trust Van Beek, Edwin; Edinburgh Imaging, Queen's Medical Research Institute, University of Edinburgh Murchison, John; Royal Infirmary of Edinburgh Ahn, Jong; Lunit Inc Lee, Sang; Lunit Inc Seth, Ambika; Lunit Inc Espinosa Morgado, Abdala Trinidad; Oxford University Hospitals NHS Foundation Trust, Emergency Medicine Research Oxford Fu, Howell; Oxford University Hospitals NHS Foundation Trust, Clinical Radiology Novak, Alex; Oxford University Hospitals NHS Foundation Trust, Emergency Medicine Research Oxford Salik, Nabeeha; RAIQC Ltd Campbell, Alan; University College London Hospitals NHS Foundation Trust, Radiology Shah, Ruchir; Oxford University Hospitals NHS Foundation Trust, Clinical Radiology Gleeson, Fergus; Churchill Hospital Ather, Sarim; Oxford University Hospitals NHS Foundation Trust
Primary Subject Heading:	Radiology and imaging
Secondary Subject Heading:	Respiratory medicine, Cardiovascular medicine
Keywords:	Chest imaging < RADIOLOGY & IMAGING, Diagnostic Imaging, RESPIRATORY MEDICINE (see Thoracic Medicine)

AI Assisted Detection for Chest X-rays (AID-CXR): a Multi-Reader Multi-Case Study Protocol

ABSTRACT

Introduction: A chest X-ray (CXR) is the most common imaging investigations performed worldwide. Advances in machine learning and computer vision technologies have led to 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' experience will independently review all studies to establish the ground truth reference standard with arbitration from a 3rd 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 31.12.2024.

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 Hospitals information governance teams. The results will be presented at relevant conferences and published in a peer-reviewed journal.

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 firstly during the unaided and then the aided phase of the study, creating a risk of recall bias and learning effect 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 stand-alone 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 upon 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-ray 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 to the outpatient setting. The poor clinical state of some of these patients, limits their ability to comply with 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, 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 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 accuracy of chest X-ray interpretation by general radiologists, ED physicians, 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 reporting accuracy of healthcare professionals and algorithm performance, e.g. type abnormality, size of abnormality, PA/AP view, mobile/fixed X-ray, 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 multi-case (MRMC) design. The Oxford Acute Care patients or public were not involved in the design, conduct, reporting, or dissemination plans of the study. The study period is from 31st March 2024 to 31st 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 PACS/IT team through a database search of the Computerised Radiology Information System (CRIS) 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 ten 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 anteroposterior (AP) or posteroanterior (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 viewer with a structured reporting template.

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

- emergency medicine (ED)
- adult intensive care (ICU)
- adult general medicine (AGM)
- radiographers (Rad)
- general radiologists

Each specialty group consists of 6 members of ranked seniority. For the physicians this consists of:

- Two 'Juniors' (within 4 years after graduating medical school i.e. F1-ST2 grade)
- Two 'Middle Grades' (between 5-8 years after graduating medical school i.e. 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 5 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 score from the algorithm will be utilised for the ROC analyses, while binary classification results with a predefined operating cut-off will be used for 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 as. 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 CXR and available clinical information, readers will be asked to select what further action is required from the following:

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

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 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 utilisation of the AI algorithm. The impact of the algorithm on the 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 Area Under Receiver Operating Characteristic Curve (AUROC).
Reader performance: sensitivity, specificity with vs without AI assistance.
Reader confidence: self-reported diagnostic confidence on a 5-point Likert scale, with vs without AI assistance.
Reader speed: mean time taken to review a scan, with vs 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 utilized for the ROC analyses, while binary classification results with three different operating cut-offs will be used for 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 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 vs 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 upon request to the corresponding

author. Further dissemination strategy will be strongly guided by our 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.

FIGURE LEGENDS

Figure 1: Reader study summary flowchart

CONTRIBUTORSHIP STATEMENT

SA, AN and FG led the design of the protocol, with contributions from FK, ID, MK, EvB, AM, AE, HF, NS, AC and RS. SA, RS, FK and NS and AC and 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. AE registered the study and coordinates 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.

COMPETING INTEREST STATEMENT

There are no competing interests to declare for any of the authors involved. The funders from Lunit Inc. have no input into the study design, analysis, reporting or decision to publish.

FUNDING STATEMENT

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.

DATA DE-IDENTIFICATION AND SHARING STATEMENT

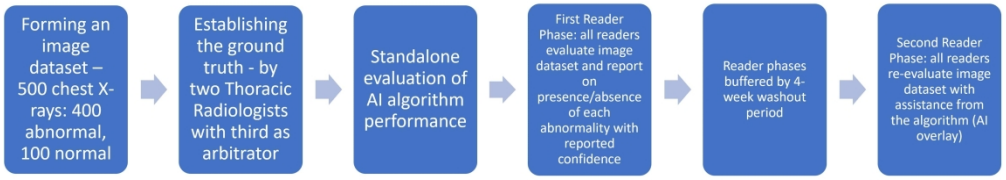
CXR images selected for the study will be anonymised in accordance with the information governance protocol at the participating hospitals. Access to the images will be controlled via the study platform using separate user accounts for each reader.

All study data will be entered into a password-protected and secure database. Individual reader accuracy scores will be anonymised, and the study team will not have access to the identifying link between the participants' personal details and the data. Data about the participants' level of experience and professional group will be retained to allow group comparison.

REFERENCES

1. Jones CM, Buchlak QD, Oakden-Rayner L, Milne M, Seah J, Esmaili N, et al. Chest radiographs and machine learning - Past, present and future. J Med Imaging Radiat Oncol. 2021 Aug;65(5):538–44.
2. Ahmad HK, Milne MR, Buchlak QD, Ektas N, Sanderson G, Chamtie H, et al. Machine Learning Augmented Interpretation of Chest X-rays: A Systematic Review. Diagn Basel Switz. 2023 Feb 15;13(4).
3. van Beek EJR, Ahn JS, Kim MJ, Murchison JT. Validation study of machine-learning chest radiograph software in primary and emergency medicine. Clin Radiol. 2023 Jan;78(1):1–7.
4. Kundu R, Das R, Geem ZW, Han GT, Sarkar R. Pneumonia detection in chest X-ray images using an ensemble of deep learning models. PloS One. 2021;16(9):e0256630.
5. Matsumoto T, Kodera S, Shinohara H, Ieki H, Yamaguchi T, Higashikuni Y, et al. Diagnosing Heart Failure from Chest X-Ray Images Using Deep Learning. Int Heart J. 2020 Jul 30;61(4):781–6.
6. Hillis JM, Bizzo BC, Mercaldo S, Chin JK, Newbury-Chaet I, Digumarthy SR, et al. Evaluation of an Artificial Intelligence Model for Detection of Pneumothorax and Tension Pneumothorax in Chest Radiographs. JAMA Netw Open. 2022 Dec 1;5(12):e2247172.

7. Homayounieh F, Digumarthy S, Ebrahimian S, Rueckel J, Hoppe BF, Sabel BO, et al. An Artificial Intelligence-Based Chest X-ray Model on Human Nodule Detection Accuracy From a Multicenter Study. JAMA Netw Open. 2021 Dec 1;4(12):e2141096.
8. Artificial intelligence for analysing chest X-ray images [Internet]. NICE (The National Institute for Health and Care Excellence); 2022. Available from: 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 Apr 1;135:49–65.
10. Wu JT, Wong KCL, Gur Y, Ansari N, Karargyris A, Sharma A, et al. Comparison of Chest Radiograph Interpretations by Artificial Intelligence Algorithm vs Radiology Residents. JAMA Netw Open. 2020 Oct 1;3(10):e2022779.
11. 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. 2018 Nov;15(11):e1002683.



Reader study summary flowchart

1486x418mm (72 x 72 DPI)