Operating Characteristics and Algorithm Performance

AI-Rad Companion Chest X-ray

Version 2.0 – 4th February 2021

SIEMENS Healthineers

Table of contents

Abstract

The following whitepaper details the operating characteristics and bench testing performance of the algorithms within AI-Rad Companion Chest X-ray VA23A and newer versions. It describes the general principles of the software such as its workflow and results. The whitepaper also explains performance metrics like area under receiver operating characteristics curve, sensitivity, specificity etc. and how to interpret these algorithm performance metrics. It also delves into how the device performance can be benchmarked in comparison to average performance of board-certified radiologists. Towards this, a multi-case multi-reader study has been conducted to measure the average reader performance of radiologists for analysis of chest X-rays and compare it with the standalone performance of AI-Rad Companion Chest X-ray VA23A. The target radiographic findings include pulmonary lesions, atelectasis, pneumothorax, consolidation, and pleural effusion. The conclusions of the study are used as a measurable baseline to quantify standard-of-care performance. It has been demonstrated that AI-Rad Companion Chest X-ray VA23A performs comparable to the average performance of radiologists participating in the study with higher area under receiver operating characteristics curve and sensitivity for all the target radiographic findings. With AUC of 95–99 %, the AI algorithms within AI-Rad Companion Chest X-ray demonstrate high accuracy for the detection of the target radiographic findings for use as a diagnostic aid for concurrent reading of Chest X-rays.

Introduction

AI-Rad Companion Chest X-ray is a diagnostic aid for radiologists which identifies and highlights the pre-specified radiographic findings using artificial intelligence algorithms. It is intended to be used by a radiologist concurrently with original images before a final decision is made on a case. The results generated by AI-Rad Companion Chest X-ray summarize the identified radiographic findings with localization information in the format of secondary capture DICOM objects and as machine readable DICOM Structured Reports. AI-Rad Companion Chest X-ray is designed as a clinical extension to the AI-Rad Companion Engine and is deployed through the teamplay digital health platform. It must be used in conjunction with appropriate software such as reporting software, to report the findings and clinical observations. The workflow for clinical integration of AI-Rad Companion Chest X-ray is shown in Figure 1.

AI-Rad Companion Chest X-ray is indicated for adult patients and for use in routine secondary and tertiary care settings including routine check-up, clinical outpatient departments and disease screening. AI-Rad Companion Chest X-ray will operate at a higher sensitivity than an average radiologist as intended for Computer Aided Detection devices and it must be noted that the device could mark occasional false positive or miss findings. AI-Rad Companion Chest X-ray is not designed to detect presence of radiographic findings other than the prespecified list (pulmonary lesions,

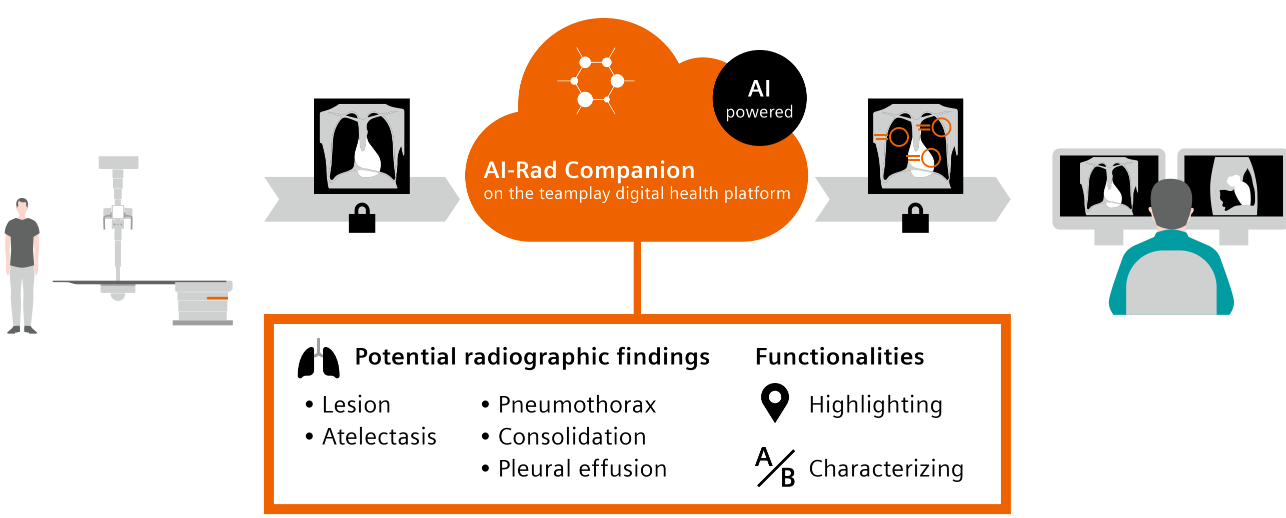
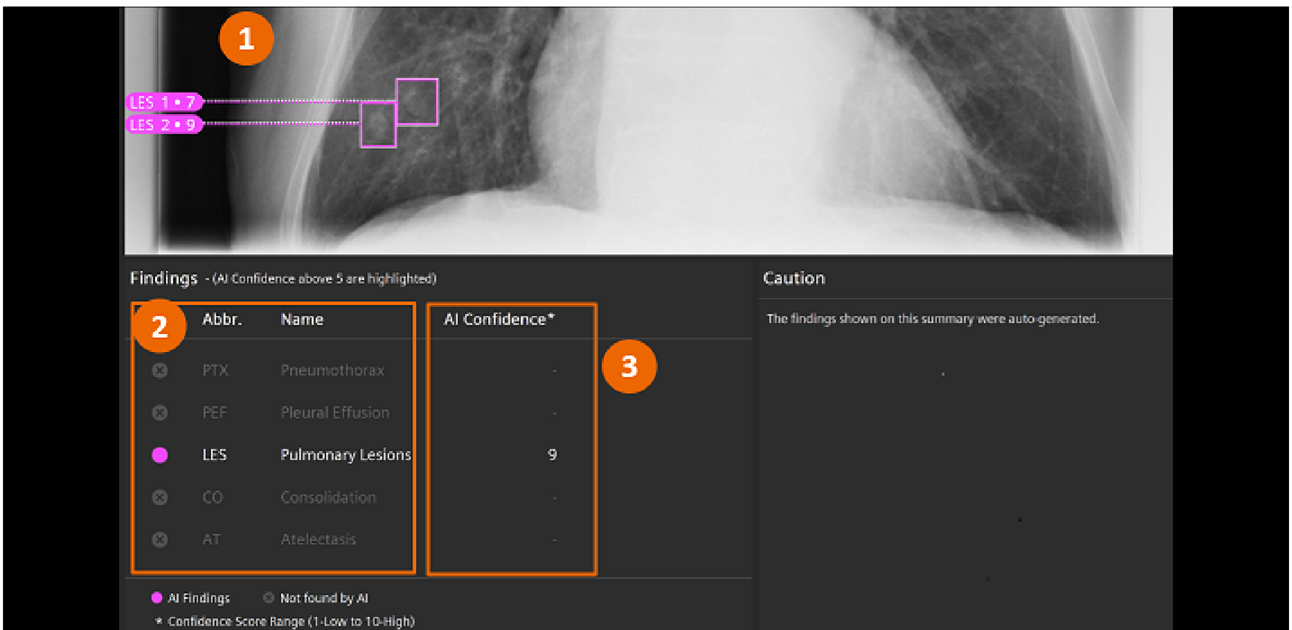


Figure 1: Clinical workflow of AI-Rad Companion Chest X-ray.

consolidation, atelectasis, pneumothorax, and pleural effusion). Radiologists should review original images for all suspected pathologies by following standard clinical procedures. AI-Rad Companion Chest X-ray works best if the Chest X-ray image is acquired in compliance with the practice parameter for performance of chest radiography laid out by American College of Radiology guidelines [1] such as using high-kilovoltage technique (120 to 150 kVp), using anti-scatter technique, etc. American College of Radiology guidelines [1] also recommends quality control of acquired radiographs by a trained technologist prior to archival in the PACS to ensure the anatomy is captured sufficiently and appropriate DICOM tags are filled in. Figure 2 pictorially illustrates the different elements that are generated as a part of the results generated by AI-Rad Companion Chest X-ray. Figure 3 illustrates sample outputs generated by the medical device indicating the suspected findings found, their location, and their corresponding confidence scores (see Figure 2 for an illustration).



1 **annotated image**

2 **overview table**
- Presence of finding indicated by **pink dot**
- Finding types found in image → **white font**
- Not found → **grayed out**

3 **AI Confidence Score**
- **Normalized algorithm confidence:** 1 (low) – 10 (high)
- Shown are only **6–10**
- For findings not found by AI, score is ≤ 5 (Not shown in the image)
- Normalized score to express the **algorithm's certainty for the presence of a finding**
  - **no measure of actionability**
  - **no measure of malignancy**

Figure 2: Description of elements in results generated by AI-Rad Companion Chest X-ray.

The definitions of the radiographic findings detected in AI-Rad Companion Chest X-ray are adapted from RadLex [2] and Fleischner Society Glossary for Thoracic Imaging [3] and are defined as follows:

- **Pulmonary Lesions:** AI-Rad Companion Chest X-ray detects pulmonary lesions including lung nodules and masses. Nodules present as rounded or oval opacities < 3 cm in diameter and masses are defined as pulmonary, pleural, or mediastinal lesions greater than 3 cm in diameter. In a study by Eltorai et al. [4], a vast majority of 88.4 % of US-board certified thoracic radiologists desired AI applications that aided in detection of pulmonary nodules.

- **Pneumothoraces:** AI-Rad Companion Chest X-ray detects radiographic signs suggestive of gas (air) in the pleural space, including both small and large pneumothoraxes[1]. In a study by Eltorai et al. [4], a majority of 56.8 % of US-board certified thoracic radiologists desired AI applications that aided in detection of pneumothoraces.

- **Atelectasis:** AI-Rad Companion Chest X-ray detects increased opacities associated with volume loss which can be accompanied by abnormal displacement of fissures, bronchi, vessels, the diaphragm, the heart, or the mediastinum. It is one of the most common respiratory complication after a surgical procedure.

- **Consolidation:** AI-Rad Companion Chest X-ray detects radiographic signs associated with increased parenchymal attenuation which includes consolidation and ground glass opacity. Consolidation appears as a homogeneous increase in pulmonary parenchymal attenuation that obscures the margins of vessels and airway walls. Ground glass opacity appears as an area of hazy increased lung opacity, usually extensive, within which margins of pulmonary vessels may be indistinct but not obscured.

- **Pleural effusions:** AI-Rad Companion Chest X-ray detects radiographic signs suggestive of fluid in the pleural space between the visceral and parietal pleura.

[1] *The characterization of the size of pneumothorax is based on Baumann, M.H., Strange, C., Heffner, J.E., Light, R., Kirby, T.J., Klein, J., Luketich, J.D., Panacek, E.A. and Sahn, S.A., 2001. Management of spontaneous pneumothorax: an American College of Chest Physicians Delphi consensus statement. Chest, 119(2), pp.590-602.*

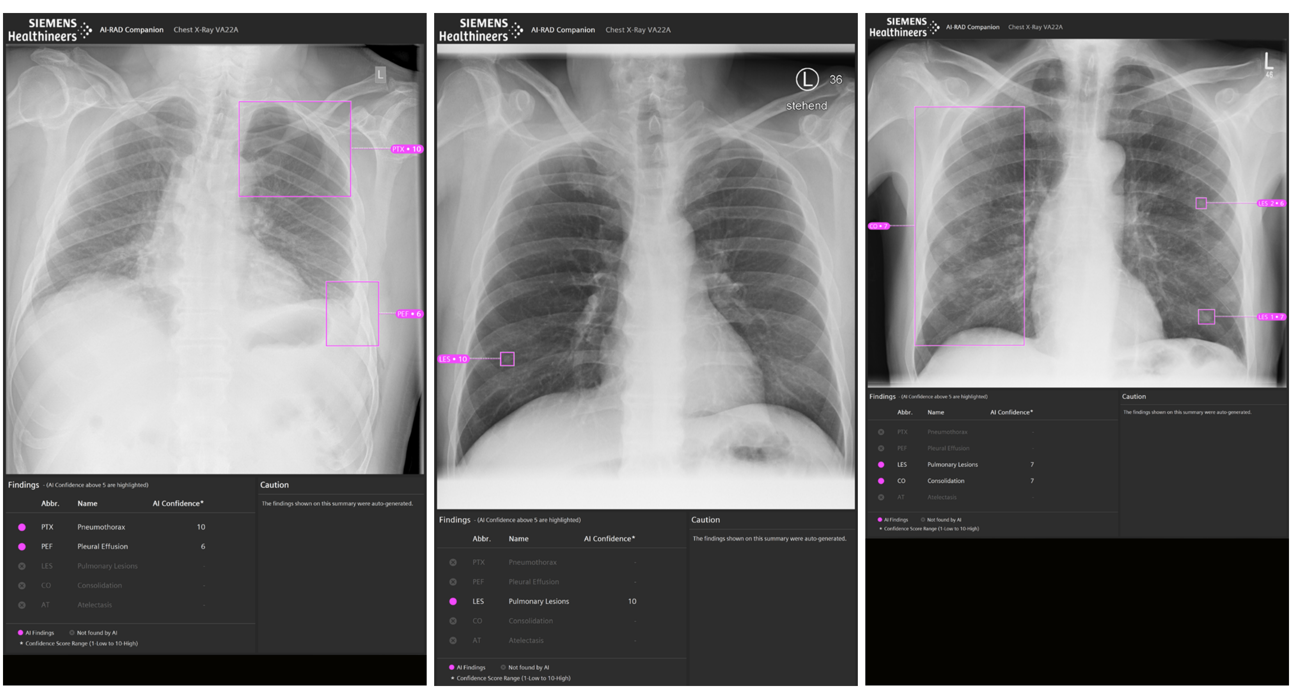


Figure 3: Examples of device output of AI-Rad Companion Chest X-ray with markings indicating Pulmonary Lesions, Atelectasis, Consolidation, Pleural Effusions, and Pneumothorax, along with the corresponding AI-confidence scores. Image courtesy: Medizinische Versorgungszentrum Prof. Dr. Uhlenbrock und Partner in Dortmund, Germany.

Multi-case multi-reader study

Motivation

Reading Chest X-rays is subject to a large inter-reader variability in clinical routine. To better estimate the performance of human readers on the task of identifying radiographic findings scoped in this device, a clinical reader study was conducted where the average reading performance of human readers is estimated. Artificial Intelligence based diagnostic assistant that is intended to improve the diagnostic accuracy of readers should have a performance superior or comparable to the standard of care. Since an objective measurable definition of standard of care is not possible, the average reader performance observed in this study was used to compare and contrast the performance of AI-Rad Companion Chest X-ray. Statistical analyses were performed comparing the means of each group (AI-Rad Companion Chest X-ray vs. radiologists). The null hypothesis of both groups performing equally was tested using a paired t-tests. The goal of the study is purely to form an empirical basis for standalone testing of AI-models and is not intended to replace radiologist review of the original cases. The medical device is intended as a diagnostic aid and does not remove any cases from the radiology worklist.[2]

Dataset description

In total 1019 retrospective patients were selected from three different sites in US and Europe: Princeton Radiology in Princeton NJ, USA, Medizinische Versorgungszentrum Prof. Dr. Uhlenbrock und Partner in Dortmund, Germany, and Ludwigs-Maximilians-Universität, Munich, Germany.

The data set is representative of the target intended population (Mean Age: 59.1 years with 24.7 % of data with < 45 years old, 48.4 % female) and is sampled at random from a larger consecutively collected dataset acquired for product development. The dataset consists of patients from the United States and European Union (Germany). The data set was enriched using parsing of the clinical reports using Natural Language Processing (NLP) for the pre-specified radiographic findings to ensure that enough positive cases will be presented after truthing (at least 30 positive cases). The case selection is based on the initial clinical reports which were automatically processed by in-house NLP technique to reach the target distribution. The distribution of vendors within the dataset is shown in Figure 4. Only the Posterior-anterior (PA) view Chest X-ray images fulfilling a set of pre-defined DICOM image quality gate was included in this study[3]. Additional to the enrichment of positive cases with targeted findings, a control group is randomly selected which only encompasses the normal patients without findings referred in the clinical reports.

[2] Refer to the instructions for use of AI-Rad Companion Chest X-ray for further details on the intended use, indications of use, contraindications, and use scenarios of the device.

[3] Refer to the data sheet of AI-Rad Companion Chest X-ray for further details on DICOM tag prerequisites.

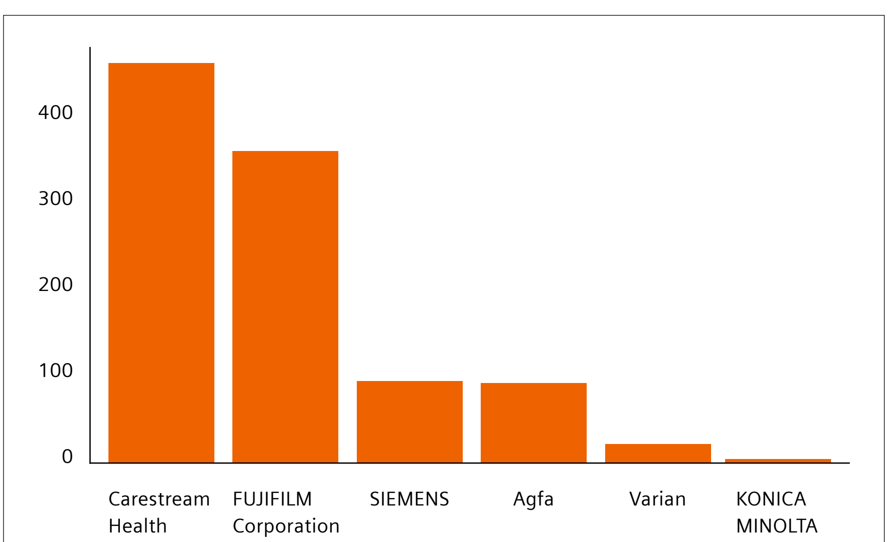


Figure 4: Distribution of vendors within the retrospective cohort in this MRMC study.

Truthing procedures

The ground truth for the test data set was constructed in a consensus reading fashion. Three board-certified radiological experts with each > 7 years of reading experience (two truthers have Fellowship of the Royal College of Radiology), read and marked all the abnormalities they found in the image independently. In case of disagreement, the consensus with regards to the presence of an abnormality is reached by majority voting. At least two experts need to confirm and agree a positive finding in terms of type and location before it is recognized as positive in the ground truth. The distribution of positive and negative cases truthed by consensus reading is given in Table 1. The controls are included in the negatives pool for each of the findings. The truthers also confirmed if the datasets were of diagnostic quality and cases were replaced in case the quality was found to be not sufficient for clinical use.

Reading procedures

In the reading session, seven independent board-certified radiologists (radiology experience ranging from 2 years to 12 years, 5 male / 2 female) were recruited in the reader study. The individual reader iterated each case and identified the findings of which the type and location were indicated and marked. Considering the subjectivity in reading and interpreting chest X-rays, for each finding that the reader identified was ranked with a confidence level scaling from 1 to 10 (1 = lowest confidence to 10 = highest confidence), representing how much confidence was associated to the detection of the finding by the reader. The full dynamic range of the confidence level, which can be translated to the probability of the presence of a finding, ensures a better estimation of the reader performance in terms of AUC (area under the curve) of the receiver operating characteristics curve. The AUC is calculated at a case-level. If there are multiple instances of the same finding, the maximum confidence at the case-level was used. For negatives, zero-imputation was performed.

Results

Performance metrics

The following performance metrics are calculated for each of the findings and are used to measure both reader and AI algorithm performance:

- **Sensitivity:** proportion of positive images (i.e. having a certain radiographic finding) correctly labeled as positive. This is also known as the True Positive Rate and is a measure of what part of positive cases are detected by the reader or the AI algorithm.

- **Specificity:** proportion of negative images (i.e. not having a certain radiographic finding) correctly labeled as negative. False positive rate is defined as 1-specificity and is a measure of how often the device/reader creates a false alarm.

- **Receiver Operating Characteristics (ROC) curve:** this curve is created by plotting the True Positive rate (sensitivity) against the False Positive Rate (1-specificity) at various threshold settings [4].

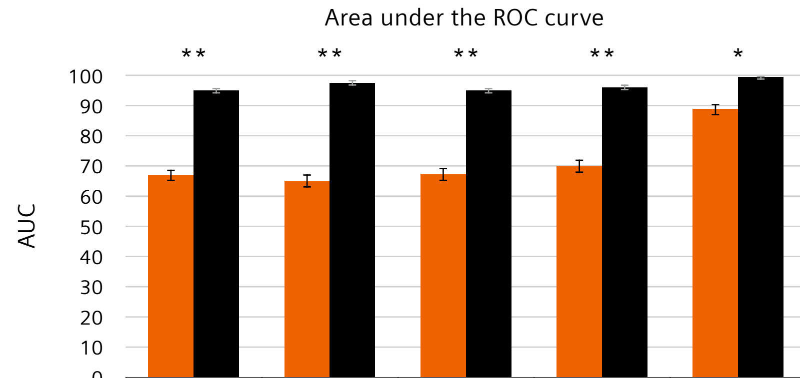
- **Area under the ROC curve (AUC):** this number estimates the probability of correct ranking positive/negative. It is bound between 0 and 1: the closer to 1, the better the system's performance. 1 means a perfect classification of positives vs. negatives. A diagnostic test producing an AUC value between 0.9 to 1.0 is considered 'Excellent' in the traditional academic point system [4].

Results on reader and algorithm performance

The observed average performance of the board-certified readers and the contrasting results of the algorithm performance of AI-Rad Companion Chest X-ray are shown below for each of the target findings in Figures 5–7 for AUC, sensitivity, and specificity respectively. To measure the average reader performance, the area under the ROC curve (AUC), the sensitivity at a confidence of ≥ 5 with the corresponding specificity are reported.

Radiographic Findings	Pulmonary lesions	Atelectasis	Consolidation	Pleural effusion	Pneumothorax
Positive	138	69	55	69	61
Control	306	264	278	253	257

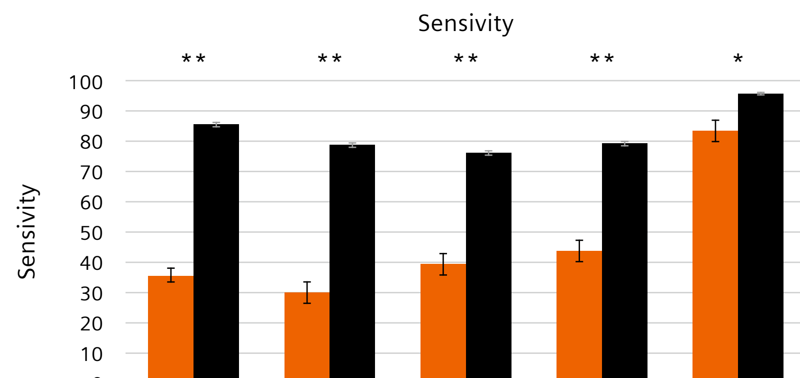
Table 1: Final composition in number of cases within the dataset for bench testing after definition of ground truth.



	LES	PTX	CO	AT	PEF
Readers	67	64.9	67.2	69.9	88.8
AI Algorithm	94.9	97.5	94.9	96	99.4

Figure 5: Performance of Readers and AI algorithm measured in terms of area under the receiver operating characteristics curve.

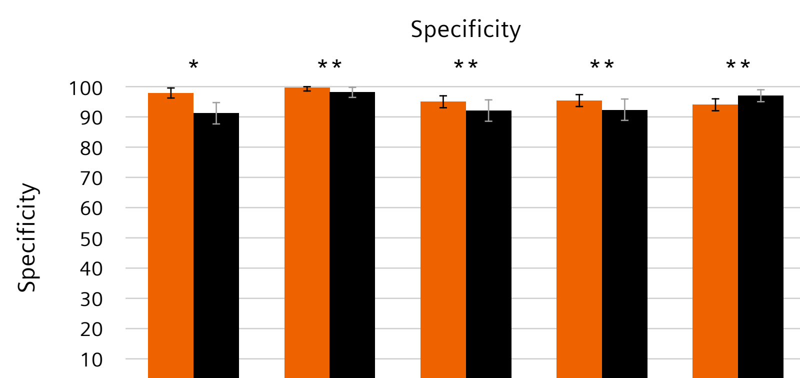
(Notation: *: p-value < 0.05, **: p-value < 0.01 of superiority tests).



	LES	PTX	CO	AT	PEF
Readers	35.5	30	39.4	43.7	83.4
AI Algorithm	85.5	78.7	76.1	79.2	95.7

Figure 6: Performance of Readers and AI Algorithm measured in terms of their sensitivity to detection of each of the radio-graphic findings.

(Notation: *: p-value < 0.05, **: p-value < 0.01 of superiority tests).



	LES	PTX	CO	AT	PEF
Readers	97.9	99.7	95	95.4	94
AI Algorithm	91.2	98.1	92.1	92.3	97

Figure 7: Performance of Readers and AI algorithm measured in terms of specificity in the detection of each of the target radiographic findings.

(Notation: *: p-value < 0.05, **: p-value < 0.01 of non-inferiority tests).

Discussion and conclusions

With AUC of 95–99 %, the AI algorithms within AI-Rad Companion Chest X-ray demonstrate high accuracy for the detection of the target radiographic findings.

Pulmonary Lesions: The detection of pulmonary lesions is considered a challenging task while reading and this is reflected in the performance of the readers in the reader study. Particularly, the case-level sensitivity of readers was observed at 35.5 as subtle pulmonary lesions are easy to miss. In contrast, the sensitivity of AI improved to 85.5 at case-level with a statistically significant margin of 50 points (p-value < 0.01).

Consolidation: The AUC of AI algorithm for detection of consolidation in comparison to readers improved by statistically significant margins by 27.7(p-value < 0.01) and the sensitivity by 36.7 (p-value < 0.01).

Atelectasis: The AUC of AI algorithm for detection in comparison to readers improved by statistically significant margins by 26.2 (p-value < 0.01) and the sensitivity by 35.5 (p-value < 0.01).

Pleural Effusion: We observe a high reader performance of with AUC of 88.8 and high sensitivity of 83.4 for the task of detection of pleural effusion. The AI algorithm demonstrates a higher AUC of 99.4 and sensitivity of 95.7.

Pneumothorax: Subtle and small pneumothoraces are easy to miss while reading chest x-rays. The AI algorithm has higher sensitivity 78.7 than an average reader of 30 with comparable specificity (AI algorithm 97 vs. reader specificity of 99.7).

The intended workflow of AI-Rad Companion Chest X-ray for the clinician is to read the original images and AI results concurrently in his reading workstation. The AI algorithms would mark and identify findings with a higher sensitivity and accuracy than the average of readers participating in this study. This increased sensitivity would help alleviate potentially missed findings and act as additional confirmatory evidence prior to reporting on the image. It must however be noted that AI-Rad Companion Chest X-ray should not be used in lieu of full patient evaluation or solely relied upon to make or confirm a diagnosis. It is not intended to replace the review of the X-ray image by a radiologist.

Revision history

Version	Comments
1.0	Contains methodology, results and conclusions of internal multi-reader multi-case study and bench testing of the algorithms of AI-Rad Companion Chest X-ray VA23 and beyond.
2.0	Revised version including feedback from Customer Relationship Management to include definitions of radiographic findings and findings-specific discussion and conclusions .

References

1 American College of Radiology, 2017. ACR–SPR–STR practice parameter for the performance of chest radiography.
2 RadLex radiology lexicon http://www.radlex.org/
3 Hansell, D.M., Bankier, A.A., MacMahon, H., McLoud, T.C., Muller, N.L. and Remy, J., 2008. Fleischner Society: glossary of terms for thoracic imaging. Radiology, 246(3), pp.697-722.
4 Eltorai, A.E., Bratt, A.K. and Guo, H.H., 2020. Thoracic radiologists’ versus computer scientists’ perspectives on the future of artificial intelligence in radiology. Journal of thoracic imaging, 35(4), pp.255-259.
5 Fawcett, T., 2006. An introduction to ROC analysis. Pattern recognition letters, 27(8), pp.861-874.

AI-Rad Companion Chest X-ray is not commercially available in all countries, and its future availability cannot be ensured.

The information in this document contains general technical descriptions of specifications and options as well as standard and optional features which do not always have to be present in individual cases, and which may not be commercially available in all countries. Due to regulatory reasons their future availability cannot be guaranteed. Please contact your local Siemens organization for further details.

Siemens reserves the right to modify the design, packaging, specifications, and options described herein without prior notice. Please contact your local Siemens sales representative for the most current information.

Note: Any technical data contained in this document may vary within defined tolerances. Original images always lose a certain amount of detail when reproduced.

Siemens Healthineers Headquarters
Siemens Healthcare GmbH
Henkestr. 127
91052 Erlangen, Germany
Phone: +49 9131 84-0
siemens-healthineers.com