# Independent evaluation of the accuracy of 5 artificial intelligence software for detecting lung nodules on chest X-rays

Kirill Arzamasov[1,2]^, Yuriy Vasilev[1,3]^, Maria Zelenova[1]^, Lev Pestrenin[1]^, Yulia Busygina[1]^, Tatiana Bobrovskaya[1]^, Sergey Chetverikov[1]^, David Shikhmuradov[1]^, Andrey Pankratov[1], Yury Kirpichev[1]^, Valentin Sinitsyn[1,4]^, Irina Son[5]^, Olga Omelyanskaya[1]^

[1]State Budget-Funded Health Care Institution of the City of Moscow "Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department", Moscow, Russian Federation; [2]MIREA – Russian Technological University, Moscow, Russian Federation; [3]Federal State Budgetary Institution "National Medical and Surgical Center named after N.I. Pirogov" of the Ministry of Health of the Russian Federation, Moscow, Russian Federation; [4]Lomonosov Moscow State University, Moscow, Russian Federation; [5]Federal State Budgetary Educational Institution of Further Professional Education "Russian Medical Academy of Continuous Professional Education" of the Ministry of Healthcare of the Russian Federation, Moscow, Russian Federation

*Contributions:* (I) Conception and design: Y Vasilev, V Sinitsyn, I Son, K Arzamasov; (II) Administrative support: O Omelyanskaya; (III) Provision of study materials or patients: D Shikhmuradov, A Pankratov; (IV) Collection and assembly of data: T Bobrovskaya, Y Busygina, S Chetverikov; (V) Data analysis and interpretation: M Zelenova, L Pestrenin, Y Kirpichev; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

*Correspondence to:* Kirill Arzamasov, MD, PhD. State Budget-Funded Health Care Institution of the City of Moscow "Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department", Petrovka str., 24, 127051, Moscow, Russian Federation; MIREA – Russian Technological University, Vernadsky Avenue, 78, 119454, Moscow, Russian Federation. Email: ArzamasovKM@zdrav.mos.ru.

**Background:** The integration of artificial intelligence (AI) into medicine is growing, with some experts predicting its standalone use soon. However, skepticism remains due to limited positive outcomes from independent validations. This research evaluates AI software's effectiveness in analyzing chest X-rays (CXR) to identify lung nodules, a possible lung cancer indicator.

**Methods:** This retrospective study analyzed 7,670,212 record pairs from radiological exams conducted between 2020 and 2022 during the Moscow Computer Vision Experiment, focusing on CXR and computed tomography (CT) scans. All images were acquired during clinical routine. The final dataset comprised 100 CXR images (50 with lung nodules, 50 without), selected consecutively and based on inclusion and exclusion criteria, to evaluate the performance of all five AI-based solutions, participating in the Moscow Computer Vision Experiment and analyzing CXR. The evaluation was performed in 3 stages. In the first stage, the probability of a nodule in the lung obtained from AI services was compared with the Ground Truth (1— there is a nodule, 0—there is no nodule). In the second stage, 3 radiologists evaluated the segmentation of nodules performed by the AI services (1—nodule correctly segmented, 0—nodule incorrectly segmented or not segmented at all). In the third stage, the same radiologists additionally evaluated the classification of the nodules (1—nodule correctly segmented and classified, 0—all other cases). The results obtained in stages 2 and 3 were compared with Ground Truth, which was common to all three stages. For each stage, diagnostic accuracy metrics were calculated for each AI service.

**Results:** Three software solutions (Celsus, Lunit INSIGHT CXR, and qXR) demonstrated diagnostic

---

^ ORCID: Kirill Arzamasov, 0000-0001-7786-0349; Yuriy Vasilev, 0000-0002-5283-5961; Maria Zelenova, 0000-0001-7458-5396; Lev Pestrenin, 0000-0002-1786-4329; Yulia Busygina, 0000-0002-4775-258X; Tatiana Bobrovskaya, 0000-0002-2746-7554; Sergey Chetverikov, 0000-0002-3097-8881; David Shikhmuradov, 0000-0003-1597-5786; Yury Kirpichev, 0000-0002-9583-5187; Valentin Sinitsyn, 0000-0002-5649-2193; Irina Son, 0000-0001-9309-2853; Olga Omelyanskaya, 0000-0002-0245-4431.

metrics that matched or surpassed the vendor specifications, and achieved the highest area under the receiver operating characteristic curve (AUC) of 0.956 [95% confidence interval (CI): 0.918 to 0.994]. However, when evaluated by three radiologists for accurate nodule segmentation and classification, all solutions performed below the vendor-declared metrics, with the highest AUC reaching 0.812 (95% CI: 0.744 to 0.879). Meanwhile, all AI services demonstrated 100% specificity at stages 2 and 3 of the study.

**Conclusions:** To ensure the reliability and applicability of AI-based software, it is crucial to validate performance metrics using high-quality datasets and engage radiologists in the evaluation process. Developers are recommended to improve the accuracy of the underlying models before allowing the standalone use of the software for lung nodule detection. The dataset created during the study may be accessed at https://mosmed.ai/datasets/mosmeddatargogksnalichiemiotsutstviemlegochnihuzlovtipvii/.

**Keywords:** Chest X-ray (CXR); artificial intelligence (AI); lung nodules; radiology; computer vision

# Introduction

Machine learning (ML), artificial neural networks (ANNs), and deep learning (DL) are all components of artificial intelligence (AI) that have seen a surge in interest and application in recent times. ML is the process of using algorithms to automate decision-making by employing models that are not manually coded but are instead trained on datasets. ANNs, a subset of ML, are designed to mimic the brain's structure and functionality. DL, meanwhile, utilizes a network of interconnected neurons across multiple layers, facilitating the analysis and processing of extensive and intricate data. In the medical field, these technologies are being integrated to enhance the speed and effectiveness of disease diagnosis and treatment (1).

Recently, radiologists have investigated how to use AI in medical imaging (2-6). AI-based analysis of chest X-rays (CXR) has the potential to assist in the diagnosis and triage of patients with lung cancer, tuberculosis, pneumonia, and other diseases. Although various researchers develop solutions that may be introduced to clinical settings, several companies have already made their software available to end users. These companies often claim that their software has high diagnostic accuracy metrics, but in most cases, independent evaluations either do not occur or do not validate the companies' claimed metrics.

For example, in a study by van Leeuwen KG *et al.*, it was shown that for 64 out of 100 studied CE-labelled AI products, there was insufficient peer-reviewed evidence on their efficacy. Only 18/100 AI products have demonstrated potential clinical impact (7).

In a systematic review conducted by Kelly *et al.* showed that in 77 studies for which external validation was performed and direct comparison was possible, AI-based software performance decreased on average by 6% when externally validated (range of increase from 4% to a decrease of 44%) (8).

Approaches such as the evaluation commercial AI solutions in radiology (the ECLAIR guidelines) have been proposed to critically evaluate AI-based solutions before purchase. In particular, they suggest paying attention to issues related to performance and validation of AI-based software (9).

Therefore, we decided to make a small contribution to independent evaluations of commercial AI software.

The aim of our work was to provide an independent evaluation of five commercially available AI-based solutions for chest radiography and to assess their applicability for the diagnosis of lung nodules. We present this article in accordance with the STARD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-24-160/rc).

# Methods

## *The computer vision experiment and choice of software*

In 2019 the Moscow government began Moscow Computer Vision Experiment which aims to investigate if AI solutions can be introduced into routine clinical practice (10). The ongoing experiment (2023) aims to support radiologists' decision-making. During the Experiment, different vendors

**Table 1** A detailed description of the metrics presented by the developers of the studied software (declared within the Moscow Computer Vision Experiment)

| Software name (vendor name if different) | Diagnostic accuracy metrics | | | |
|---|---|---|---|---|
| | AUC | Sensitivity | Specificity | Accuracy |
| qXR (Qure.ai, represented in Russia by LLC "Chestnaya meditsina") | 0.920 | 0.900 | 0.820 | 0.850 |
| Celsus (LLC "Medical Screening Systems") | 0.920 | 0.900 | 0.860 | 0.860 |
| Program for automated analysis of digital fluorograms (LLC "FtizisBioMed") | 0.950 | 0.900 | 0.980 | 0.940 |
| Care Mentor AI | 0.930 | 0.860 | 0.920 | 0.910 |
| Lunit INSIGHT CXR (Lunit, represented in Russia by R-Pharm) | 0.920 | 0.790 | 0.950 | N/A |

AUC, area under the receiver operating characteristic curve; AI, artificial intelligence; CXR, chest X-ray; N/A, not available.

**Table 2** Details of the AI-based software

| AI-based software | Architectures | Dataset | Source |
|---|---|---|---|
| qXR | CNN | Training: 3.5 million CXR images. Testing: 213,459 X-rays randomly selected from a set of 3.5 million X-rays used. Validation: 13,426 independent images | (12) |
| Celsus | Mask-RCNN, DenseNet-121 and some other models | Training: 29 thousand CXR images. Testing: 4 thousand CXR images | N/A |
| Program for automated analysis of digital fluorograms | N/A | N/A | N/A |
| Care Mentor AI | Inception-V3 and ResNet-50 | 276,840 frontal X-ray lung images | (13) |
| Lunit INSIGHT CXR | N/A | 54,221 normal chest radiographs and 35,613 chest radiographs from patients with major thoracic diseases | (14) |

AI, artificial intelligence; CXR, chest X-ray; N/A, not available.

of AI-based software presented their solutions to the Center for Diagnostics and Telemedicine, which approved the inclusion of the best-performing solutions in the Unified Radiological Information Service (URIS). For this study, we chose the software from the participants of Moscow Computer Vision Experiment. We realise that there are now many other AI services, including non-commercial ones with open access. However, our choice was driven primarily by the need for healthcare system in the safe delivery of medical care. In the Moscow Experiment, AI services are used by radiologists to make decisions, so the omission of pathology by AI services is highly undesirable. Identifying nodules in the lungs is a demanding task, and their omission by AI services undermines confidence in new technologies for both doctors and patients. That is why we have chosen only those AI services that are available to our radiologists.

Seven solutions proved capable of analyzing CXR data, but two of the vendors declined public publication of their results. Therefore, five AI-based solutions were included

in the further analysis. Many software contributing to the Experiment is not typically mentioned in the AI-based solution reviews. Thus, a current study might update the readers on some novel AI-based software and its performance. The details on AI-based software regulations in Russia can be found elsewhere (11).

### AI-based software solutions

The following five solutions were chosen for the present study: qXR, Celsus, Program for automated analysis of digital fluorograms, Care Mentor AI, and Lunit INSIGHT CXR. The software was accepted to participate in the Moscow Experiment if they claimed the area under the receiver operating characteristic curve (AUC) greater than 0.810. A detailed description of the solutions and their diagnostic metrics (presented by the vendors) is shown in *Tables 1-3*.

All 5 AI services work in medical organizations and are

**Table 3** Five AI-based solutions chosen for evaluation and comparison

| Name | Link | Description | References (including usage examples) |
|---|---|---|---|
| qXR | https://app.qure.ai/landing-page | The software helps detecting findings across lungs, pleura, mediastinum, bones, diaphragm, and heart on chest X-ray in <1 min. It is able to differentiate normal X-ray studies and flag radiological signs of such conditions as TB, lung cancer and heart failure | (15-17) |
| Celsus | https://lk.celsus.ai/demo?lang=eng | The solution reduces the analysis time and improves the interpretation accuracy for fluorography and radiography images | (18) |
| Program for automated analysis of digital fluorograms | http://www.ftizisbiomed.ru/ | Analyzes digital fluorographic images and identifies pathological foci | (19) |
| Care Mentor AI | http://carementor.ru/ | Interprets the results of radiological examinations (X-ray, CT, MRI, and mammography) in order to optimize the detection of various pathological conditions at an early stage | (13,20-26) |
| Lunit INSIGHT CXR | https://insight.lunit.io/cxr/login | Computer-Assisted Detection Software that serves as a concurrent/second reading aid for the physicians. Its capabilities include detection, localization, identification, and reporting of suspicious abnormal radiologic findings | (27-29) |

AI, artificial intelligence; TB, tuberculosis; CT, computed tomography; MRI, magnetic resonance imaging; CXR, chest X-ray.

therefore registered in Russia as medical devices (National Registration by Roszdravnadzor). Without a registration certificate, a medical device cannot be sold, used or imported. In general, National Medical Registration by Roszdravnadzor is an analogue of CE-mark and FDA certification. Among the 5 services we evaluated, Lunit INSIGHT CXR and qXR have CE and FDA approval, Celsus has CE approval.

All AI services processed each image for less than 1 minute, including the time of study transfer, processing, and response. There were no statistical differences in image processing time between the AI services.

### *The dataset*

This is a retrospective study. To assess the diagnostic metrics of the software, we prospectively created a dataset of CXR images. The images were chosen from all radiological examinations acquired within the Moscow Computer Vision Experiment from 2020 to 2022. The final dataset consisted of 100 diagnostic images of CXR, 50 of which were with the signs of lung cancer and other 50 were with no specific features found (30). To select images with pathological findings, we analyzed 7,670,212 pairs

of records that included both CXR and chest computed tomography (CT) scans. Images (both CXR and chest CT scans) were acquired during clinical routine. It is to note that 100 images represented a target dataset size, since the number of images that can be analyzed within the rational time frames is limited. The rationale of the sample size may be found in our previous work (31).

The selection of studies was carried out consecutively. We started with selecting the records where CXR was done within 14 days prior to CT (92,436 pairs of images). Subsequently, using keywords indicating the presence of lung lesions and negative keywords, we obtained 8,503 pairs of images. The keywords (in Russian) represented correct and incorrect spellings of the following words: "lesion", "focus" (for CXR records) and "lump", "tumor", "neoplasia", "mass" (for CT records). Negative keywords consisted of a long list of organs and pathologies not related to any lung pathology. The next step involved a thorough review of the textual reports attached to the records. The descriptions were processed if they contained the following words: "oncology, consultation, CT, peripheral lesion, focal changes, signs of single foci, multiple foci, solid nodules, neoplasms" and otherwise excluded if they contained "lung resection, tuberculoma, pleural effusion, pneumonia, cyst,
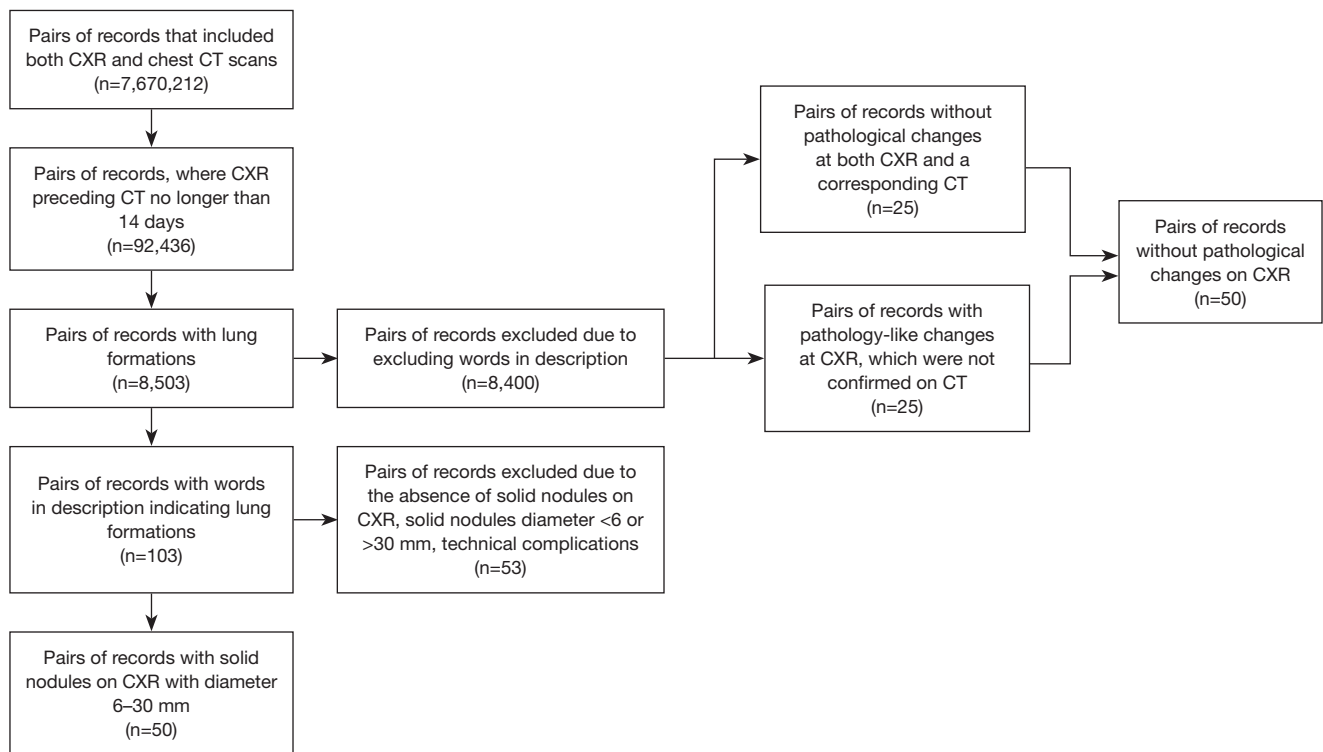
**Figure 1** Systematic selection of images with and without lung nodules. CXR, chest X-ray; CT, computed tomography.

fibrosis, post-tuberculous changes, pleuritis, not detected, normal, central lesion of the lung". The resulting pairs of CXR and CT files consisted of 103 pairs of images. Afterward, the images were reviewed by an expert radiologist and selected if a solid nodule was observed on CXR, and excluded if an opacity of less than 6 mm and greater than 30 mm was present. During the final stage, the records with technical issues were excluded from the study, so the final dataset consisted of 50 CXR images with lung nodules. Next, 50 images without pathological changes were selected from the CXR records. The selection consisted of two steps: first, 25 images with no pathological changes in both CXR and a corresponding CT were selected. Then, we picked 25 images where CXR showed pathological changes with no CT confirmation. The resulting dataset consisted of 100 carefully processed CXR images [for the details on the sample sufficiency, see (31,32)], where 50 images highlighted lung nodules and the other 50 images had no findings. Due to technical reasons, some of the 5 solutions failed to process all 100 images. Therefore, the image processing results (Appendix 1) contain 95 analyzed images. The dataset may be accessed at https://mosmed.ai/datasets/mosmeddatargogksna lichiemiotsutstviemlegochnihuzlovtipvii/.

Study demographics was as follows: 51 females, 47 males, and 2 cases with missing gender data. The median age was 58 years (minimum: 18 years, maximum: 87 years). All CXR and CT included in the study were performed in Moscow between November 16, 2017, and April 6, 2022. The detailed process of image selection is described in Materials and Methods. We selected the records where CXR was done within 14 days prior to CT. The participant flow is presented in *Figure 1*.

There were no adverse events associated with the performance of the index test or reference standard.

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013), and was approved by the Independent Ethics Committee of the Moscow Regional Office of the Russian Society of Radiologists and Radiographists (approval number 2, protocol code 2/2020 and date of approval 20.02.2020). The clinical trial number is NCT04489992. Informed consent form was signed during the clinical routine.

### Interpretation and statistical analysis

The scores for the AI-based software were calculated as

**Table 4** Study stages

| Features of the stages | Stage 1 | Stage 2 | Stage 3 |
|---|---|---|---|
| Object of assessment | CXR | Labeling of nodules on CXR by AI-based software | Labeling and classification of nodules on CXR by AI-based software |
| Who or what performed the assessment | AI-based software | 3 radiologists | 3 radiologists |
| Assessment process | AI-based software assessed the images and returned a Kafka message with a probability of unspecified pathology | Radiologists checked for nodules in the exact places where the AI-based software put the markups | Radiologists checked for nodules in the exact places where the AI-based software put the markups, and checked whether the AI-based software classified the marked-up opacity as a nodule and not otherwise |
| Assessment result of one CXR image | Probability of pathology provided by AI-based software | Binary estimate of a region with a lung nodule markup made by AI-based software | Binary estimate of a lung nodule markup and labeling made by AI-based software |
| Assessment result presentation scale | Probability of pathology presented as an integer number (from 0 to 100) | Categorical values:<br>• For 50 CXR with nodules: FN—nodule is absent (the actual nodule on CXR is not marked-up); TP—nodule is present (the actual nodule on CXR is marked-up)<br>• For 50 CXR without pathological changes: TN—nodule is absent (no markup on CXR); FP—nodule is present (the markup is present on CXR) | Categorical values:<br>• For 50 CXR with nodules: FN—nodule is absent (the actual nodule on CXR is not marked-up or marked-up but not labeled as a nodule); TP—nodule is present (the actual nodule on CXR is marked-up and labeled as a nodule)<br>• For 50 CXR without pathological changes: TN—nodule is absent (no markup and label on CXR); FP—nodule is present (markup and label of nodule are present on CXR) |
| Approach to choosing a cut-off value | Modified. The cut-off value corresponded to the maximum value of the Youden index | Unchanged. The cut-off value was pre-set by the developer, since the AI-based software marked-up only the nodules whose probability exceeded the pre-set cut-off value | Unchanged. The cut-off value was pre-set by the developer, since the AI-based software marked-up and labeled only the nodules whose probability exceeded the pre-set cut-off value |
| Ground truth | Each study was initially assigned one of the two classes: 1—a nodule on CXR with a diameter of 9–30 mm, confirmed on CT; 0—a CXR image without pathology, with the corresponding CT containing no findings | | |

CXR, chest X-ray; AI, artificial intelligence; FN, false negative; TP, true positive; TN, true negative; FP, false positive; CT, computed tomography.
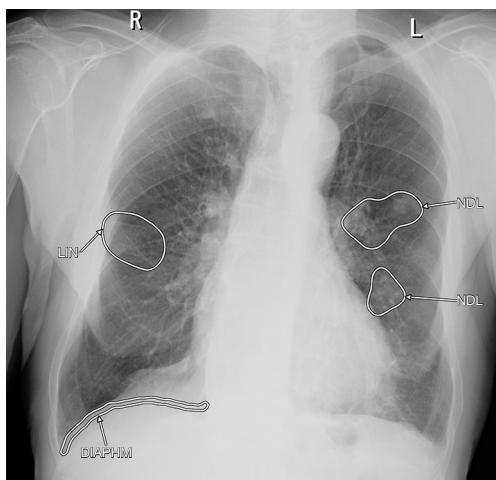
**Figure 2** An image with the markup made by one of the AI-based solutions. R, right side; L, left side; LIN, linear opacification; NDL, nodular opacification; DIAPHM, diaphragm disorder; AI, artificial intelligence.

follows (*Table 4*).

(I) Stage 1. AI-based software had access only to medical images. The scores of the AI-based solutions were calculated as follows. Once the images had been processed, the software returned the probability of pathology (a score from 0 to 100) in a Kafka message and on images containing the corresponding markups [in a Digital Imaging and Communications in Medicine (DICOM) format]. Kafka message is a text message in JavaScript Object Notation (JSON) format containing study processing details. Kafka messages are used to facilitate the interaction between AI-based software and the radiological information system. The probability of pathology in a study is an indicator important for triaging studies in the radiologist's worklist. In this case, the response threshold can be redefined depending on the clinical task. For each AI-based solution, using the probability estimates from AI, we calculated the AUC and determined the point that corresponds to the maximum value of the Youden index. Next, the point was set as a cutoff value for each model. Using these cut-off values, the image scores were tagged as "No finding" when the initial score was below the cut-off value; the "Pathological changes" tag was set when the initial score was above the cut-off value.

(II) Stage 2. The radiologists accessed the marked-up images (*Figure 2*) along with the findings and impressions from DICOM structured report (SR) (*Figure 3*). Every expert independently evaluated the DICOM images that the AI-based software returned with the marked pathological loci. At this stage, the experts reviewed the markup (if the nodule was exactly where the software placed the markup). If the opinions of at least 2 out of three experts matched, the corresponding answer [true positive (TP), true negative (TN), false positive (FP), and false negative (FN)] was put in the final evaluation table.

(III) Stage 3. At this stage, the experts evaluated whether the software recognized the marked-up lesion as a lung nodule. If the opinions of at least 2 out of three experts matched, the corresponding answer (TP, TN, FP, and FN) was put in the final evaluation table.

A schematic illustration of the three stages of our study is presented in Figure S1.

Diagnostic metrics (AUC, specificity, sensitivity, accuracy) were calculated using a Web tool for receiver operating characteristic analysis (https://roc-analysis.mosmed.ai/, accessed on the 1st of June, 2023) (33). McNemar's test was used for paired comparisons of the sensitivity and specificity of the software. Three radiologists manually studied the images that contained the annotation provided by the AI-based solutions to ensure they included correct lung nodule markup and labels.

### *The experts*

The experts were represented by three radiology specialists with working experience of greater than 5 years and sub-specialization in CXR. All doctors received the corresponding medical diplomas from the medical universities with residency in radiology. The experts had access to AI-processed CXRs as well as CT and clinical information. At stages 2 and 3 of the present study, every expert independently evaluated the DICOM images that the AI-based software returned with the marked pathological loci. Then, the independent assessments were collected and in case at least two radiologists presented the same estimate, were set as final estimates.

### Results

The results of the study are presented in *Table 5*.

| >>>Code meaning | Finding |
|---|---|
| >>Text value | Lungs: Linear opacification detected with a probability of 0.62, nodular opacification detected with a probability of 0.66<br>Pathology of the pleura and diaphragm: Pathology of the diaphragm was detected with a probability of 0.62<br>Cardiac changes: No cardiac changes<br>Other changes: No other changes found<br>Final probability of pathology: 0.66 |
| >>Relationship type | CONTAINS |
| >>Value type | TEXT |
| >>Concept name code sequence | [This is a sequence] |
| >>>Code value | 209001 |
| >>>Coding scheme designator | 99PMP |
| >>>Code meaning | Impression |
| >>Text value | Probability of pathology: 0.66 |

**Figure 3** Example of a DICOM SR containing findings and impressions with the probability of pathology scores provided by one of the AI-based software. DICOM, Digital Imaging and Communications in Medicine; SR, structured report; AI, artificial intelligence.

### Stage 1: detection of pathologies overall

As shown in *Table 5*, the AUC values obtained at the first stage of our study were consistent with the developer's claims for three AI-based software: qXR, Celsus, and Lunit INSIGHT CXR. The remaining two software (Programme for automated analysis of digital fluorograms and Care Mentor AI) showed AUCs that were statistically significantly lower than those stated by the developers.

The sensitivity indicators showed a similar pattern. The only exception was Lunit INSIGHT CXR, which demonstrated a statistically significantly higher sensitivity than was declared by the developers. On the contrary, the specificity of this AI-software turned out to be significantly lower than the declared one.

It is also interesting that the specificity of Celsus, which demonstrated sensitivity not inferior to that declared by the developer, was higher than declared.

The data were further used to calculate the number of TP, TN, FP and FN interpretations (Table S1).

### Stage 2: lung nodules segmentation

In the second stage of the study, we analyzed the abilities of 5 software solutions to segment nodules on chest radiographs. Three radiologists reviewed all the processed images to see if the AI correctly segmented the nodules. The segmentation was considered correct if the model segmented the nodule or the region where it was located. Segmenting more than 1/3 of the lung area was considered an incorrect segmentation.

At the second stage of our study, we found that the AUC of 4 out of 5 services was significantly lower than the developers' claims. Only Celsus demonstrated the AUC corresponding to the one declared by the developer (*Table 5*).

The sensitivity of all AI-based software was significantly lower than that stated by the developers. In practice, this manifested itself in the fact that the software either did not segment lung nodules at all (*Figure 4*) or segmented more than 1/3 of the lung field in which the nodule was located (*Figure 5*).

In contrast to sensitivity, the specificity of all 5 AI services at the second stage of the study was significantly higher than that stated by the developers. In fact, this meant that none of the AI services segmented non-existent nodules on all 50 CXRs without lung nodules.

The data were further used to calculate the number of TP, TN, FP and FN interpretations (Table S2).

### Stage 3: lung nodules segmentation and classification

In the third stage of the study, we analysed the ability of AI services to correctly segment and classify lung nodules. In other words, to correctly solve the task that radiologists perform in their routine clinical practice.

Evaluations of the software's ability to segment nodules were obtained in the second stage of the study. In the third stage, three radiologists further reviewed all images for correct classification performed by the AI models. The classification was considered correct if the model identified the detected nodule as a nodule (1—"Correct Classification") and incorrect if the model misclassified the detected nodule as another pathology (0—"Incorrect Classification"). The data were further used to calculate the

**Table 5** Comparison of AUC, sensitivity, specificity, and accuracy declared by the developers of AI-based software and obtained during the three stages of the experiment for the five models

| Diagnostic accuracy metrics | AI-based software | Declared | Obtained | | |
| --- | --- | --- | --- | --- | --- |
| | | | Stage 1 (detection of pathologies overall) | Stage 2 (lung nodules segmentation) | Stage 3 (lung nodules segmentation and classification) |
| AUC (95% CI) | qXR | 0.920 | 0.921 (0.862 to 0.980) | 0.823* (0.754 to 0.889) | 0.792* (0.721 to 0.860) |
| | Celsus | 0.920 | 0.956 (0.918 to 0.994) | 0.885 (0.824 to 0.945) | 0.812* (0.744 to 0.879) |
| | Program for automated analysis of digital fluorograms | 0.950 | 0.858* (0.790 to 0.925) | 0.844* (0.775 to 0.910) | 0.688* (0.619 to 0.753) |
| | Care Mentor AI | 0.930 | 0.810* (0.723 to 0.897) | 0.708* (0.640 to 0.773) | 0.667* (0.599 to 0.734) |
| | Lunit INSIGHT CXR | 0.920 | 0.932 (0.887 to 0.977) | 0.787* (0.720 to 0.854) | 0.787* (0.720 to 0.854) |
| Sensitivity (95% CI) | qXR | 0.900 | 0.854 (0.750 to 0.954) | 0.646* (0.510 to 0.781) | 0.583* (0.444 to 0.723) |
| | Celsus | 0.900 | 0.875 (0.781 to 0.970) | 0.770* (0.652 to 0.890) | 0.625* (0.488 to 0.762) |
| | Program for automated analysis of digital fluorograms | 0.900 | 0.750* (0.630 to 0.872) | 0.690* (0.556 to 0.819) | 0.375* (0.238 to 0.512) |
| | Care Mentor AI | 0.860 | 0.604* (0.466 to 0.740) | 0.417* (0.277 to 0.556) | 0.333* (0.200 to 0.467) |
| | Lunit INSIGHT CXR | 0.790 | 0.920** (0.840 to 0.990) | 0.574* (0.433 to 0.716) | 0.574* (0.433 to 0.716) |
| Specificity (95% CI) | qXR | 0.820 | 0.830 (0.722 to 0.937) | 1.0** (1.0 to 1.0) | 1.0** (1.0 to 1.0) |
| | Celsus | 0.860 | 0.960** (0.900 to 1.0) | 1.0** (1.0 to 1.0) | 1.0** (1.0 to 1.0) |
| | Program for automated analysis of digital fluorograms | 0.980 | 0.960 (0.900 to 1.0) | 1.0** (1.0 to 1.0) | 1.0** (1.0 to 1.0) |
| | Care Mentor AI | 0.920 | 0.910 (0.835 to 0.990) | 1.0** (1.0 to 1.0) | 1.0** (1.0 to 1.0) |
| | Lunit INSIGHT CXR | 0.950 | 0.810* (0.700 to 0.920) | 1.0** (1.0 to 1.0) | 1.0** (1.0 to 1.0) |
| Accuracy (95% CI) | qXR | 0.850 | 0.880 (0.820 to 0.950) | 0.820 (0.744 to 0.898) | 0.789 (0.707 to 0.871) |
| | Celsus | 0.860 | 0.916 (0.860 to 0.972) | 0.884 (0.820 to 0.950) | 0.810 (0.732 to 0.889) |
| | Program for automated analysis of digital fluorograms | 0.940 | 0.850* (0.781 to 0.930) | 0.842* (0.769 to 0.915) | 0.684* (0.590 to 0.778) |
| | Care Mentor AI | 0.910 | 0.760* (0.672 to 0.844) | 0.705* (0.610 to 0.797) | 0.663* (0.568 to 0.758) |
| | Lunit INSIGHT CXR | N/A | 0.860 (0.790 to 0.930) | 0.787 (0.707 to 0.868) | 0.787 (0.704 to 0.870) |

The comparison relies on the ground truth markup. The values of the obtained metrics, taking into account 95% CI, which were less than those stated by the developer are marked with "*", and those which were more than those stated by the developer are marked with "**". The metrics named by the vendors are shown for detection of pathologies overall. AUC, area under the receiver operating characteristic curve; AI, artificial intelligence; CXR, chest X-ray; CI, confidence interval.

number of TP, TN, FP and FN interpretations (Table S3).

We found that at this stage the AUC and sensitivity of all 5 AI services were lower than the values declared by the developers. Also, the values of these parameters at stage 3 of the study were lower than at stage 2 for most of the AI services. This was due to the fact that some of the lung nodules found at the second stage were misclassified at the third stage (*Figure 6*).

Specificity at stage 3 remained 1.0 for all AI services, as none of the AI services detected non-existent lung nodules on 50 CXR without lung nodules.

### Analyzing FP and FN responses of AI services

Further, it was observed that the AI-based software yielded FP and FN results for some images more often than for the
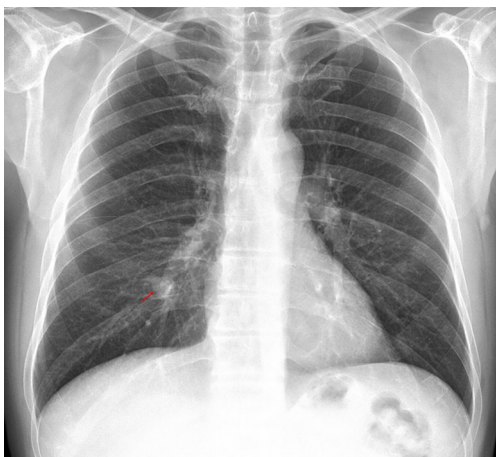
**Figure 4** An example of a clearly visible to the human eye nodule in the lung (indicated by the red arrow) that was not segmented by 3 out of 5 AI-based software. AI, artificial intelligence.
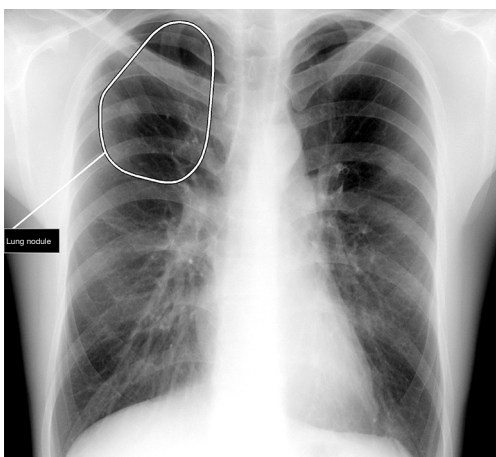


**Figure 5** An example of excessive segmentation of a lung nodule by one of the AI based software. In such cases, the nodule was considered to be incorrectly segmented. AI, artificial intelligence.

others. In the end, at least 4 out of 5 models misinterpreted 17 images from the dataset: 12 radiography studies were assigned a FN status and five images were labeled as FP. To identify possible causes of the errors, we submitted original radiography studies and those processed by the models to three radiologists for review. They concluded that the models failed to detect the lung nodules (i.e., produced FN conclusions) for the following reasons:

(I)   The shadow of the nodule was overlapped by the shadow either of the hilum, the rib, or both at the same time (11 out of 12 FN studies),
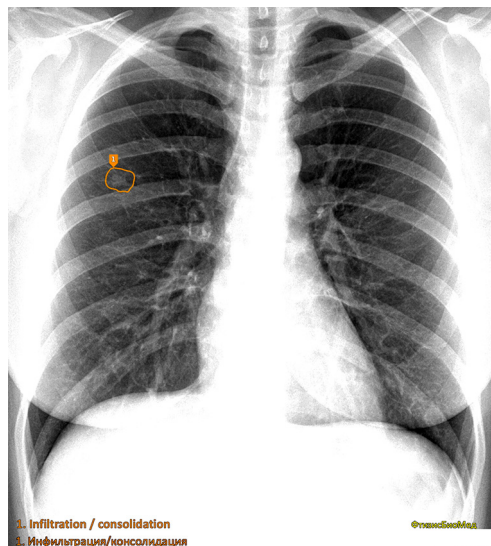


**Figure 6** Example of correct segmentation but misclassification (nodule defined as infiltration/consolidation) made by one of the AI services. AI, artificial intelligence.

(II)   The diameter, density, and location of the nodule shadow resembled the cross-section of an artery (1 of 12 FN studies).

The main reason for the FP results among the 5 images was the detection of non-target pathology, most often cardiomegaly (4 cases). We must point out that in each of these situations, the findings 'observed' by AI were not actually present in the images. When analyzing the total number of the estimates, 46 results were FNs and 23 were FPs.

### Paired comparisons of the diagnostic metrics obtained during the previous stages

At all stages of the study, no statistically significant differences were observed between the sensitivity values of the three models that demonstrated the highest values of this parameter (Figure S2). At the second and third stages, the specificity of all models reached the maximum possible value.

### Discussion

#### Comparison of diagnostic accuracy metrics of AI services obtained in our and other studies

When analyzing the 5 AI services operating within the Moscow Experiment, we obtained AUC values from 0.810

to 0.956 in the first stage of our study and from 0.667 to 0.812 in the third stage of our study.

It is interesting to compare the results with those of other authors. A study by Kufel *et al.* reported reported that their AI model achieved an AUC of 0.771 for lung nodules. Also according to the comparative analysis done by the authors, the AUC in other studies ranged from 0.669 to 0.811 (34). In a systematic review and meta-analysis by Aggarwal *et al.*, the AUC when lung nodules are detected on CXR is 0.884 [95% confidence interval (CI): 0.842 to 0.925] (32). In a recent study by van Leeuwen *et al.* evaluating the ability of four commercial AI services to detect pulmonary nodules, the AUC was in the range of 0.86 to 0.93 (35). We can see that the AUC obtained in other studies varies widely. Nevertheless, we can say that the AUC obtained for 5 AI services at stage 1, 2 and, most importantly, 3 of our study is generally comparable to the AUC of other AI software in studies by other authors.

However, we still believe that it is not entirely appropriate to compare the AUC that was obtained in different studies on different datasets. For a more objective comparison of AI services it is necessary to evaluate their performance on the same dataset. This is a logical continuation of our work, which we plan to carry out in the future.

### Comparison of AI service diagnostic accuracy metrics obtained in our study and radiologists' diagnostic accuracy metrics obtained in other studies

Another issue worth discussing is the comparison of diagnostic accuracy metrics between radiologists and AI services. According to a number of authors, the AUC of radiologists in the lung nodule detection task ranges from 0.810 to 0.839 (35-37). Compared to this range of values, the AUC values we obtained in stage 1 of the study were higher in 4 of the 5 AI services. Only one of the 5 AI services showed an AUC value comparable to the radiologists' AUC values reported in the literature in the third stage of the study.

In the future, we plan to evaluate the performance of radiologists on the dataset we used to compare it with the performance of the 5 AI services presented in this paper.

### Discussion of the causes of FNs

When analyzing images that confused most solutions, we highlighted the most common reasons for image

misattribution. The obtained data are consistent with the conclusions of several other authors, who believe the most common cause of FN results was the shadow of the lung nodule overlapping with that of the ribs and clavicles (38,39). This reason can be considered one of the most significant since the suppression of the rib and collarbone shadows on CXR with a neural network can boost the sensitivity of AI-based models from 79.8% to 91.5% (P<0.001) (40). The prevalence of FN results is consistent with the generally accepted view that the identification of lung nodules is one of the most challenging tasks in automated medical image analysis (41). The detection of complex (i.e., hard to differentiate) nodules in the lungs is difficult to both radiologists and AI (42).

On the one hand, AI-based software errors can help attract the specialist's attention to an image, and despite the mislabeling, once received by a radiologist, the image will be thoroughly analyzed. On the other hand, errors can undermine trust which means the radiologists would less likely seek help from AI.

### Relationship between the provenance of AI service training datasets and the resulting diagnostic accuracy metrics

It is also worth paying attention to the relationship between the data sets used to train the AI services under consideration and the diagnostic accuracy metrics obtained. It is known that unbalanced training data sets by race and ethnicity can lead to biases in the further operation of AI services in the practical healthcare of a particular country (43,44). In our study, we did not find a clear impact of datasets on the performance of AI services. For example, CareMentor AI, which demonstrated the lowest AUC among 5 AI services, was trained on datasets collected in the USA and Russia. On the contrary, Celsus, also trained on data from the USA, Russia and several other countries (UK, Vietnam, Belarus), showed the highest AUC among the 5 AI services. Two foreign AI services for Russia: qXR (trained on X-rays collected in 45 centers worldwide) and Lunit INSIGHT CXR (trained on X-rays collected in Korea) had AUC values close to Celsus.

### Scenarios for fine-tuning AI services to address specific clinical challenges

As mentioned above, the cut-off value was adjusted according to the maximum value of the Youden index only

at the first stage of the research. At the second and third stages, we had to use only cut-off values that corresponded to the internal settings of the AI-based software.

We assume that, unlike the 1st stage, the lower sensitivity and higher specificity at the 2nd and 3rd stages may partly be due to the choice of the cut-off values at each of the stages. This paves the way for fine-tuning AI-based software for clinical practice, depending on the task at hand. The first scenario suggests using AI as the only reader. In this case, the balance between sensitivity and specificity corresponding to the maximum value of the Youden index will be relevant. Another approach is using AI to assign patients with suspected pathology in CXR into a risk group, followed by an in-depth examination or dynamic observation. In this case, the settings that secure the highest sensitivity to make sure the AI would not miss a pathology will be preferred. Thirdly, the opposite task may be considered—ruling out patients with normal CXR to reduce the burden on radiologists. To achieve this, the optimal settings must provide maximum specificity to correctly determine the patients with normal CXR with little to no error.

Therefore, from a clinical standpoint, when it comes to nearly ideal specificity, FNs are the most unforgiving of errors, because late detection of lung nodules significantly increases the risk of adverse outcomes. It may also be assumed that the main focus of AI use in modern practice may change from detecting a pathology to triage.

Speaking of detecting target pathologies, none of the AI-based solutions lived up to their product claims. Therefore, in this study, none of the software was capable to secure 100% sensitivity, although all of them are certified as medical devices. According to the reviewed literature, such deviations are likely to be caused by either overfitting due to small and/or unbalanced training datasets, underfitting due to excessive regularization that reduces flexibility, or the number of features in the model is too small (45). To truly demonstrate the generalizability of AI models it is essential to use external validation with target populations not involved in training. Many publicly available datasets are heavily used in AI model training and are therefore unsuitable for independent external validation. Ideally, the performance of the AI models should be externally validated using real-world screening data to demonstrate generalizability and provide a rationale for clinical adoption (46,47). AI models, just like any diagnostic tool, must be evaluated using the objective assessment standards typical to

clinical setting. In our opinion, there are several solutions to these issues:

(I)   Additional training and testing of commercially available AI-based software using real-time data before introduction to routine clinical practice;
(II)  Training on larger and more diverse datasets that include difficult-to-define lung nodules;
(III) Providing enough features during training and using enough regularization to prevent both overfitting and underfitting.

Although AI services have not yet reached 100% sensitivity, we are seeing significant improvements in the performance of AI services in a variety of clinical applications. Much of this was fuelled by the COVID-19 pandemic, when it was proven in the most challenging environment for healthcare systems around the world that AI services can improve the diagnostic quality of radiological examinations (48,49).

### *Method limitations*

The results were obtained in early 2023 using the software versions available at the moment. Updates to the versions may lead to changes in the diagnostic metrics. Here, we did not aim to compare the technical specifications of the software, therefore a thorough analysis of the model architectures was not performed.

## Conclusions

The present study assesses the effectiveness of AI-based software in CXR analysis and its suitability for detecting lung nodules. The AI solutions showed high AUC (up to 0.956) in distinguishing between normal and pathological cases. However, radiologists found that the AI's correct interpretation of CXRs with lung nodules, based on probability and a cut-off value, didn't always mean the AI segmented and classified the nodules accurately. The highest AUC among the 5 algorithms decreased to 0.885 when the segmentation correctness was examined. Moreover, when evaluating both the segmentation and classification of the lung nodules, the highest AUC dropped even further to 0.812.

Hence, for a comprehensive assessment of AI-based software's ability to detect lung nodules, a binary evaluation alone is not enough. Expert validation is required to determine whether the AI is correctly segmenting and classifying lung nodules.

5300

Arzamasov et al. Comparing AI-based software aimed at CXR assessment

## Acknowledgments

## Footnote

*Reporting Checklist:* The authors have completed the STARD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-24-160/rc

*Conflicts of Interest:* All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-24-160/coif). The authors report the funding from the autonomous non-profit organization "Moscow Center for Innovative Technologies in Healthcare" (No. USIS 122112400040-1). The authors have no other conflicts of interest to declare.

*Ethical Statement:* The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013), and was approved by the Independent Ethics Committee of the Moscow Regional Office of the Russian Society of Radiologists and Radiographists (approval number 2, protocol code 2/2020 and date of approval 20.02.2020). Informed consent form was signed during the clinical routine.

*Open Access Statement:* This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: https://creativecommons.org/licenses/by-nc-nd/4.0/.

## References

1. Kufel J, Bargieł-Łączek K, Kocot S, Koźlik M, Bartnikowska W, Janik M, Czogalik Ł, Dudek P, Magiera M, Lis A, Paszkiewicz I, Nawrat Z, Cebula M, Gruszczyńska K. What Is Machine Learning, Artificial Neural Networks and Deep Learning?-Examples of Practical Applications in Medicine. Diagnostics (Basel) 2023;13:2582.

2. Choe J, Lee SM, Hwang HJ, Lee SM, Yun J, Kim N, Seo JB. Artificial Intelligence in Lung Imaging. Semin Respir Crit Care Med 2022;43:946-60.

3. Hasani N, Morris MA, Rhamim A, Summers RM, Jones E, Siegel E, Saboury B. Trustworthy Artificial Intelligence in Medical Imaging. PET Clin 2022;17:1-12.

4. Fasterholdt I, Kjølhede T, Naghavi-Behzad M, Schmidt T, Rautalammi QTS, Hildebrandt MG, Gerdes A, Barkler A, Kidholm K, Rac VE, Rasmussen BSB. Model for ASsessing the value of Artificial Intelligence in medical imaging (MAS-AI). Int J Technol Assess Health Care 2022;38:e74.

5. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, Allison T, Arnaout O, Abbosh C, Dunn IF, Mak RH, Tamimi RM, Tempany CM, Swanton C, Hoffmann U, Schwartz LH, Gillies RJ, Huang RY, Aerts HJWL. Artificial intelligence in cancer imaging: Clinical challenges and applications. CA Cancer J Clin 2019;69:127-57.

6. Joshi G, Jain A, Araveeti SR, Adhikari S, Garg H, Bhandari M. FDA-approved artificial intelligence and machine learning (AI/ML)-enabled medical devices: an updated landscape. Electronics 2024;13:498.

7. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol 2021;31:3797-804.

8. Kelly BS, Judge C, Bollard SM, Clifford SM, Healy GM, Aziz A, Mathur P, Islam S, Yeom KW, Lawlor A, Killeen RP. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur Radiol 2022;32:7998-8007.

9. Omoumi P, Ducarouge A, Tournier A, Harvey H, Kahn CE Jr, Louvet-de Verchère F, Pinto Dos Santos D, Kober T, Richiardi J. To buy or not to buy-evaluating commercial AI solutions in radiology (the ECLAIR guidelines). Eur Radiol 2021;31:3786-96.

10. Vasilev YA, Vladzymyrskyy AV, Arzamasov KM, Andreichenko AE, Gombolevsky VA, Kulberg NS, Omelyanskaya OV, Pavlov NA, Reshetnikov RV, Sergunova KA, Sharova DE, Shulkin IM. Computer vision in radiation diagnostics: the first stage of the Moscow experiment. Moscow: Limited Liability Company Publishing Solutions, 2022.

11. Gusev AV, Vladzymyrskyy AV, Sharova DE, Arzamasov KM, Khramov AE. Evolution of research and development in the field of artificial intelligence technologies for healthcare in the Russian Federation: results of 2021. Digital Diagnostics 2022;3:178-94.

12. Mahboub B, Tadepalli M, Raj T, Santhanakrishnan R, Hachim MY, Bastaki U, Hamoudi R, Haider E, Alabousi A. Identifying malignant nodules on chest X-rays: A validation study of radiologist versus artificial intelligence diagnostic accuracy. Advances in Biomedical and Health Sciences 2022;1:137-43.

13. Nitris L, Zhukov E, Blinov D, Gavrilov P, Blinova E, Lobishcheva A. Advanced neural network solution for detection of lung pathology and foreign body on chest plain radiographs. Imaging Med 2019;11:57-66.

14. Hwang EJ, Nam JG, Lim WH, Park SJ, Jeong YS, Kang JH, Hong EK, Kim TM, Goo JM, Park S, Kim KH, Park CM. Deep Learning for Chest Radiograph Diagnosis in the Emergency Department. Radiology 2019;293:573-80.

15. Gavrilov PV, Roitberg PG, Blinov DS, Goldin MG, Blinova EV, Leontiev VS, Kamishanskaya IG, Dorofeev AA, Cheremisin VM, Novokhatko YA, Sushkov EV, Shmatok DO, Sokolov AI. Artificial intelligence-based algorithms in detection and 3D reconstruction of lung nodules on chest computed tomography scans. Russian Journal of Operative Surgery and Clinical Anatomy 2021;5:15-22.

16. Rozhkova NI, Roitberg PG, Varfolomeeva AA, Mazo MM, Dobrenkii AN, Blinov DS, Sushkov EV, Deryabina ON, Sokolov AI. Neural network-based segmentation model for breast cancer X-ray screening. Sechenov Medical Journal 2020;11:4-14.

17. Roitberg P, Blinov D, Cheremisin V. Integrating AI technologies in radiology workflows. Health Care Standardization Problems 2020;9:29-33.

18. Karpov OE, Bronov OY, Kapninskiy AA, Pavlovich PI, Abovich YA, Subbotin SA, Sokolova SV, Rychagova NI, Milova AV, Nikitin ED. Comparative study of data analysis results of digital mammography AI-based system "Celsus" and radiologists. Bulletin of Pirogov National Medical & Surgical Center 2021;16:86-92.

19. Dydykin SS, Blinova EV, Vasilyev YL, Blinov DS. Congress of the International Federation of Associations of Anatomists (the 19 Congress of the IFAA), August 9–11, 2019, London. Russian Journal of Operative Surgery and Clinical Anatomy 2019;3:4-9.

20. Zhukov EA, Blinov DS, Leontiev VS, Gavrilov PV, Smolnikova UA, Blinova EV, Kamishanskaya IG. System of digital vision for X-ray lung pathology and foreign body detection. Vrach 2020;31:34-41.

21. Nitris L, Varfolomeeva A, Blinov D, Gavrilov P, Kamishanskaya I, Lobishcheva A. Artificial intelligence-based solution for x-ray longitudinal flatfoot determination and scaling. Imaging Med 2019;11:67-75.

22. Klassen VI, Safin AA, Maltsev AV, Andrianov NG, Morozov SP, Vladzymyrskyy AV, Ledikhova NV, Sokolina IA, Kulberg NS, Gombolevsky VA, Kuzmina ES. AI-based screening of pulmonary tuberculosis: diagnostic accuracy. Journal of eHealth Technology and Application 2018;16:28-32.

23. Singh R, Kalra MK, Nitiwarangkul C, Patti JA, Homayounieh F, Padole A, Rao P, Putha P, Muse VV, Sharma A, Digumarthy SR. Deep learning in chest radiography: Detection of findings and presence of change. PLoS One 2018;13:e0204155.

24. Nash M, Kadavigere R, Andrade J, Sukumar CA, Chawla K, Shenoy VP, Pande T, Huddart S, Pai M, Saravu K. Deep learning, computer-aided radiography reading for tuberculosis: a diagnostic accuracy study from a tertiary hospital in India. Sci Rep 2020;10:210.

25. Mushtaq J, Pennella R, Lavalle S, Colarieti A, Steidler S, Martinenghi CMA, Palumbo D, Esposito A, Rovere-Querini P, Tresoldi M, Landoni G, Ciceri F, Zangrillo A, De Cobelli F. Initial chest radiographs and artificial intelligence (AI) predict clinical outcomes in COVID-19 patients: analysis of 697 Italian patients. Eur Radiol 2021;31:1770-9.

26. Nam JG, Kim M, Park J, Hwang EJ, Lee JH, Hong JH, Goo JM, Park CM. Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs. Eur Respir J 2021;57:2003061.

27. Jang S, Song H, Shin YJ, Kim J, Kim J, Lee KW, Lee SS, Lee W, Lee S, Lee KH. Deep Learning-based Automatic Detection Algorithm for Reducing Overlooked Lung Cancers on Chest Radiographs. Radiology 2020;296:652-61.

28. Vasilev Y, Vladzymyrskyy A, Omelyanskaya O, Blokhin I, Kirpichev Y, Arzamasov K. AI-Based CXR First Reading: Current Limitations to Ensure Practical Value. Diagnostics

(Basel) 2023;13:1430.

29. Pavlou M, Qu C, Omar RZ, Seaman SR, Steyerberg EW, White IR, Ambler G. Estimation of required sample size for external validation of risk models for binary outcomes. Stat Methods Med Res 2021;30:2187-206.

30. Vasilev YA, Arzamasov KM, Vladzimirskiy AV, Omelyanskaya OV, Bobrovskaya TM, Sharova DE, Nikitin NY, Kodenko MR. Preparing a dataset for training and testing software based on artificial intelligence technology. Moscow: Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department, 2023.

31. Chetverikov SF, Arzamasov KM, Andreichenko AE, Novik VP, Bobrovskaya TM, Vladzimirsky AV. Approaches to Sampling for Quality Control of Artificial Intelligence in Biomedical Research. Sovrem Tekhnologii Med 2023;15:19-25.

32. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, Ashrafian H, Darzi A. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med 2021;4:65.

33. Morozov SP, Andreychenko AE, Chetverikov SF, Arzamasov KM, Kirpichev YS, Semenov SS, Kuritsyn SO, Logunova TA, inventors; Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department, assignee. Web-based tool for performing ROC analysis of diagnostic test results. Russian Federation Certificate of state registration of a computer program 2022617324. 2022. Available online: https://telemedai.ru/en/nauka/razrabotki/veb-instrument-dlya-roc-analiza

34. Kufel J, Bielówka M, Rojek M, Mitręga A, Lewandowski P, Cebula M, Krawczyk D, Bielówka M, Kondoł D, Bargieł-Łączek K, Paszkiewicz I, Czogalik Ł, Kaczyńska D, Wocław A, Gruszczyńska K, Nawrat Z. Multi-Label Classification of Chest X-ray Abnormalities Using Transfer Learning Techniques. J Pers Med 2023;13:1426.

35. van Leeuwen KG, Schalekamp S, Rutten MJCM, Huisman M, Schaefer-Prokop CM, de Rooij M, et al. Comparison of Commercial AI Software Performance for Radiograph Lung Nodule Detection and Bone Age Prediction. Radiology 2024;310:e230981.

36. Rudolph J, Schachtner B, Fink N, Koliogiannis V, Schwarze V, Goller S, Trappmann L, Hoppe BF, Mansour N, Fischer M, Ben Khaled N, Jörgens M, Dinkel J, Kunz WG, Ricke J, Ingrisch M, Sabel BO, Rueckel J. Clinically focused multi-cohort benchmarking as a tool for external

validation of artificial intelligence algorithm performance in basic chest radiography analysis. Sci Rep 2022;12:12764.

37. Cha MJ, Chung MJ, Lee JH, Lee KS. Performance of Deep Learning Model in Detecting Operable Lung Cancer With Chest Radiographs. J Thorac Imaging 2019;34:86-91.

38. Woźniak M, Połap D, Capizzi G, Sciuto GL, Kośmider L, Frankiewicz K. Small lung nodules detection based on local variance analysis and probabilistic neural network. Comput Methods Programs Biomed 2018;161:173-80.

39. Suzuki K, Abe H, MacMahon H, Doi K. Image-processing technique for suppressing ribs in chest radiographs by means of massive training artificial neural network (MTANN). IEEE Trans Med Imaging 2006;25:406-16.

40. Kim H, Lee KH, Han K, Lee JW, Kim JY, Im DJ, Hong YJ, Choi BW, Hur J. Development and Validation of a Deep Learning-Based Synthetic Bone-Suppressed Model for Pulmonary Nodule Detection in Chest Radiographs. JAMA Netw Open 2023;6:e2253820.

41. Homayoun H, Ebrahimpour-Komleh H. Automated Segmentation of Abnormal Tissues in Medical Images. J Biomed Phys Eng 2021;11:415-24.

42. 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;4:e2141096.

43. Glocker B, Jones C, Roschewitz M, Winzeck S. Risk of Bias in Chest Radiography Deep Learning Foundation Models. Radiol Artif Intell 2023;5:e230060.

44. Tripathi S, Gabriel K, Dheer S, Parajuli A, Augustin AI, Elahi A, Awan O, Dako F. Understanding Biases and Disparities in Radiology AI Datasets: A Review. J Am Coll Radiol 2023;20:836-41.

45. Tandon YK, Bartholmai BJ, Koo CW. Putting artificial intelligence (AI) on the spot: machine learning evaluation of pulmonary nodules. J Thorac Dis 2020;12:6954-65.

46. Anderson AW, Marinovich ML, Houssami N, Lowry KP, Elmore JG, Buist DSM, Hofvind S, Lee CI. Independent External Validation of Artificial Intelligence Algorithms for Automated Interpretation of Screening Mammography: A Systematic Review. J Am Coll Radiol 2022;19:259-73.

47. Hadjiiski L, Cha K, Chan HP, Drukker K, Morra L, Näppi JJ, et al. AAPM task group report 273: Recommendations on best practices for AI and machine learning for computer-aided diagnosis in medical imaging. Med Phys 2023;50:e1-e24.

48. Kufel J, Bargieł K, Koźlik M, Czogalik Ł, Dudek P,

Jaworski A, Cebula M, Gruszczyńska K. Application of artificial intelligence in diagnosing COVID-19 disease symptoms on chest X-rays: A systematic review. Int J Med Sci 2022;19:1743-52.

49. Parczewski M, Kufel J, Aksak-Wąs B, Piwnik J, Chober D, Puzio T, Lesiewska L, Białkowski S, Rafalska-Kosior M, Wydra J, Awgul K, Grobelna M, Majchrzak A, Dunikowski K, Jurczyk K, Podyma M, Serwin K, Musiałek J. Artificial neural network based prediction of the lung tissue involvement as an independent in-hospital mortality and mechanical ventilation risk factor in COVID-19. J Med Virol 2023;95:e28787.