Journal of Digital Imaging
https://doi.org/10.1007/s10278-022-00760-2

ORIGINAL PAPER

Changes in Radiologists' Gaze Patterns Against Lung X-rays with Different Abnormalities: a Randomized Experiment

Ilya Pershin[1] · Tamerlan Mustafaev[1,2] · Dilyara Ibragimova[3] · Bulat Ibragimov[4]

Received: 5 September 2022 / Revised: 23 November 2022 / Accepted: 15 December 2022

Abstract

The workload of some radiologists increased dramatically in the last several, which resulted in a potentially reduced quality of diagnosis. It was demonstrated that diagnostic accuracy of radiologists significantly reduces at the end of work shifts. The study aims to investigate how radiologists cover chest X-rays with their gaze in the presence of different chest abnormalities and high workload. We designed a randomized experiment to quantitatively assess how radiologists' image reading patterns change with the radiological workload. Four radiologists read chest X-rays on a radiological workstation equipped with an eye-tracker. The lung fields on the X-rays were automatically segmented with U-Net neural network allowing to measure the lung coverage with radiologists' gaze. The images were randomly split so that each image was shown at a different time to a different radiologist. Regression models were fit to the gaze data to calculate the treads in lung coverage for individual radiologists and chest abnormalities. For the study, a database of 400 chest X-rays with reference diagnoses was assembled. The average lung coverage with gaze ranged from 55 to 65% per radiologist. For every 100 X-rays read, the lung coverage reduced from 1.3 to 7.6% for the different radiologists. The coverage reduction trends were consistent for all abnormalities ranging from 3.4% per 100 X-rays for cardiomegaly to 4.1% per 100 X-rays for atelectasis. The more image radiologists read, the smaller part of the lung fields they cover with the gaze. This pattern is very stable for all abnormality types and is not affected by the exact order the abnormalities are viewed by radiologists. The proposed randomized experiment captured and quantified consistent changes in X-ray reading for different lung abnormalities that occur due to high workload.

Keywords Lung fields · U-Net · Eye-tracking · Human-AI interaction · Radiologist performance

Introduction

Over the past decade, the workload of radiologists has been growing rapidly, which is due to the increased use of magnetic resonance and computed tomography modalities and the more extensive use of imaging in diagnosis, treatment, and post-treatment monitoring [1–3]. Artificial intelligence (AI) has demonstrated the potential in assisting with various tasks in radiology [4], and the COVID-19 pandemic has accelerated the adoption of such systems for medical imaging [5]. Nevertheless, the problem of effective integration of artificial intelligence systems into the working activity of a radiologist is still under investigation with various subareas almost unaddressed [6, 7]. One of the promising areas of AI in medicine is radiologist-AI interaction, where AI does not replace radiologists by providing automatic diagnosis but provides some other assistance through continuous communication with radiologists.

The AI-assisted analysis of radiological eye movements is one of the emerging research fields. Visual reading of medical images can be divided into two stages: visual search for regions of interest and interpretation of them [8]. Kim and Mansfield [9] compiled a list of 12 types of errors in radiological interpretation. In 42%, the reason for misinterpretation was under-reading, i.e., when the reader simply overlooks an abnormality. Hanna et al. [10] suggested that

✉ Bulat Ibragimov
bulat@di.ku.dk

1 Innopolis University, Republic of Tatarstan, Innopolis, Russia

2 Swanson School of Engineering, University of Pittsburgh, Pittsburgh, PA, USA

3 Kazan State University Clinic, Republic of Tatarstan, Kazan, Russia

4 University of Copenhagen, Copenhagen, Denmark

Published online: 09 January 2023

such perceptual errors could be associated with the changes in visual search processes occurring due to fatigue.

Fatigue in the reading room is an important problem that leads to an increase in diagnostic errors [10]. At the same time, fatigue is a subjective feeling and therefore difficult to quantitatively assess. Krupinski et al. [11] used the Swedish Occupational Fatigue Inventory (SOFI) questionnaire and oculomotor strain subscale from the Simulator Sickness Questionnaire (SSQ) for radiologists' fatigue estimation. Their study involved 40 radiologists who viewed 60 X-rays and concluded that after a working day, there was an increase in symptoms of fatigue, oculomotor strain, and deterioration in diagnostic quality. Hanna et al. [10] investigated the effect of night shifts on the quality of a radiologist's diagnosis. The study involved 12 radiologists, each of whom viewed 20 images during a typical working day and in the morning after a night shift. A deterioration in the SOFI and diagnostic accuracy has been observed after night shifts. This study used eye-tracking and discovered that fatigued radiologists need more time to find the first area of interest in X-rays. A similar study was done on the interpretation of CT images by radiologists with different experiences, where all radiologists exhibit worse SOFI and SSQ scores, while less experienced radiologists also demonstrate a significant reduction in diagnostic accuracy [12].

Radiologists' gaze contains information about the quality and comprehension of image reading. Eye-tracking hardware can capture the longitudinal and spatial patterns in gaze data. Hosp et al. [13] presented an approach based on machine learning and several oculomotor features to classify the surgeon's experience during arthroscopic shoulder surgery. Tien et al. [14] previously concluded that less experienced surgeons have greater pupillary entropy during cognitively demanding tasks, while experts have greater fixation rates and spend more time in the areas of interest. Brunye et al. [15] observed significant changes in pupil size when physicians examine different areas of breast histopathology scans. Similarly, Castner et al. [16] captured the increased pupil diameter of expert physicians during the analysis of complex areas of dental X-rays in contrast to less experienced physicians, whose pupils were significantly smaller in the same areas. However, the moderate changes in pupil size cannot be accurately captured by regular eye-tracking devices, and usually require the installation of specific equipment that can potentially interfere with the radiological workstation setting. The fatigue-related changes can be also seen in gaze paths, which become sketchier and less focused [17]. Often these changes are not reflected in numeric metrics such as gaze path length, and the number of fixation points, and therefore require a deeper analysis of gaze paths.

This study investigates how lung field coverage with radiologists' gaze changes with time against different abnormality types. A workstation that mimics a radiology workstation has been developed and equipped with a framework for X-ray reading. Four practicing radiologists with different levels of experience have been recruited to analyze lung X-rays while their eye movement and voice and workstation controller commands were recorded. A deep learning-based algorithm was applied to segment lung fields which enabled the automated calculation of the gaze coverage for the target anatomy. A specific randomization protocol was designed to minimize the effect of reading patterns for individual radiologists and capture the overall reading trends in lung field coverage for different abnormality types and different time points of the experiment.

Methodology

Database

A public database of chest X-rays was utilized in this experiment [18]. The X-rays in the database were annotated by three radiologists, which were recruited by the database authors. From the database, 400 chest X-rays with unambiguous labels were randomly selected. A healthy chest X-ray is considered unambiguous if no radiologist found any abnormality on it. An abnormality is considered to be unambiguously present if all three reference radiologists found it. At the same time, a chest X-ray could contain other abnormalities detected by some but not all radiologists. Such a database composition ensures that the reference labels had minimum uncertainty. In total, the database had 168 X-rays with no lung abnormalities, i.e., normal subjects, 60 X-rays with unambiguous nodule/masses labels, 72 X-rays with chest infiltrations, 48 X-rays with pneumothorax, 12 X-rays with atelectasis, and 40 X-rays with cardiomegaly. Other abnormalities diagnosed by some radiologists included aortic enlargement, pleural thickening, calcifications, and other lesions. The authors of the chest X-rays database removed most of the attributes from the DICOM headers of the X-rays. However, gender and age were available in some DICOMs, which we extracted to get an overview of the patient demographics. From the extracted metadata, the average age of the patients was 49 years (144 X-rays), with the gender compositing being 61% males and 39% females (260 X-rays). The X-rays were acquired by various scanners with resolutions ranging from 1624×1775 to 3320×3408 pixels.

In-House-Designed Radiological Workstation

A framework that mimics radiological workstations for the analysis of X-rays has been designed and augmented with eye-tracking modules [19]. Several calibration sessions that tested the overall correctness and robustness of the framework have demonstrated that the user's gaze moves too often

to the keyboard for typing the diagnosis and switching to a new X-ray. This observation stimulated us to redesign the framework by adding voice recording to it. This minimized the use of the keyboard and mouse. In the final framework design, the user dictated the diagnoses, while moving to the next X-ray, pausing, and finishing the reading were controlled by the same Enter button. The mouse could be used to change the brightness and contrast of the current X-ray image, which was also recorded by the framework. The framework recorded and synchronized eye-movement tracking, voice recording, and controller commands. The diagnostics labels were manually extracted after the experiment from the voice recordings.

From the hardware point of view, the framework was equipped with a 10-bit monitor with a resolution of 3840×2160 px and a pixel density of $\rho = 7.31$ px/mm to display X-ray images. Tobii Eye Tracker 4C was used for eye-movement tracking, while Logitech 960 headsets were used for voice recording.

Deep Learning-Based Lung Field Segmentation

One of the core ideas of the gaze analysis experiment is estimating how lung coverage with gaze changes over time and if this change is consistent for all radiologists. To automate gaze analysis, the lung fields were segmented from the chest X-ray using the U-Net [20] convolutional neural network architecture. We employed our previous modification of the U-Net [21], which captures not only the area inside the lungs but also their contours. The intuition of explicitly requesting lung contours from the U-Net is to introduce additional penalties during training for the errors located on the border of the target lungs. The segmentation errors on lung borders are more critical in contrast to segmentation errors inside the lungs, which can be fixed with simple post-processing steps like morphological operations and connected component analysis.

The original encoder of U-Net was replaced with ResNeXt50 [22] pre-trained on the ImageNet [23] public database. Contours were extracted from the lungs using morphological erosion. The U-Net for lung segmentation was trained on a public database from The Japanese Society of Thoracic Radiology (JSRT) database, which was manually segmented by van Ginneken et al. [24].

Chest X-ray Coverage

The segmented lungs combined with gaze data allow us to quantitatively estimate the coverage of the lung fields during X-ray reading. To estimate the coverage, we first need to define the view angle parameter θ that allows us to calculate the size of the image part captured by the gaze at each time point. There is no universal value of θ. However, some studies estimated the information perception on medical images against the distance between an abnormality and a gaze point [25]. Recently, Wolfe et al. [26] showed that radiologists have a 33% probability of moving their gaze to breast abnormality if it is inside a 2-degree view angle. This is supported by Stransburger et al. [27] who demonstrated that humans are able to perceive objects at a view angle of 1.5–2 degrees from the focus point. Accordingly, we set up $\theta = 2^{\circ}$.

The area A_t of the perceived visual information at the gaze point $g(t)$ at time t can be calculated as

$$A_t = \{(x, y) : \sqrt{(x - x_t) - (y - y_t)} \leq z_t \, \rho \tan\left(\frac{\theta}{2}\right)\}, \quad (1)$$

where x_t, $y_t \, \epsilon \, g(t)$ are the gaze coordinates on the monitor, $z_t \, \epsilon \, g(t)$ is the distance in mm between the participant and gaze coordinates, θ is the visual angle, ρ is the pixel density of the monitor, and x and y are the absolute image coordinates in pixels.

The lung coverage area C_τ^{lung} calculated for a time period of τ seconds from the beginning of the X-ray reading is

$$C_\tau^{\text{lung}} = \frac{|\bigcup_{t=0}^{\tau}(A_t \bigcap L)|}{|L|}, \quad (2)$$

where L is the set of pixels that belong to the lung fields and $|L|$ is the cardinality of L. The coverage information allows us to quantitatively assess the portion of the lungs observed by the reader at each time moment and the end of the reading process (Fig. 1).

Randomized Experiment Measuring Changes in Lung Coverage

This study hypothesizes that lung coverage significantly reduces with the number of X-rays a radiologist has read, and this reduction is invariant against different abnormality types. A straightforward approach to test this hypothesis would be to measure the lung coverage for all X-rays for a specific radiologist and fit a linear regression model to the resulting 400 data samples. The slopes of such models fitted to each radiologist's gaze data will estimate how the lung field coverage changes against the number of viewed images. The problem with such an approach is that reading time depends not only on the reader's fatigue but also, potentially to a greater extent, on the complexity of the depicted case. It is therefore possible that some radiologists got more attention-demanding X-rays at the start/end of the experiment, which could skew the observed regression trends.

To address the summarized above issue, the order in which radiologists view X-rays was individually randomized. The randomization was designed to ensure that each X-ray is given to some radiologists closer to the start of the

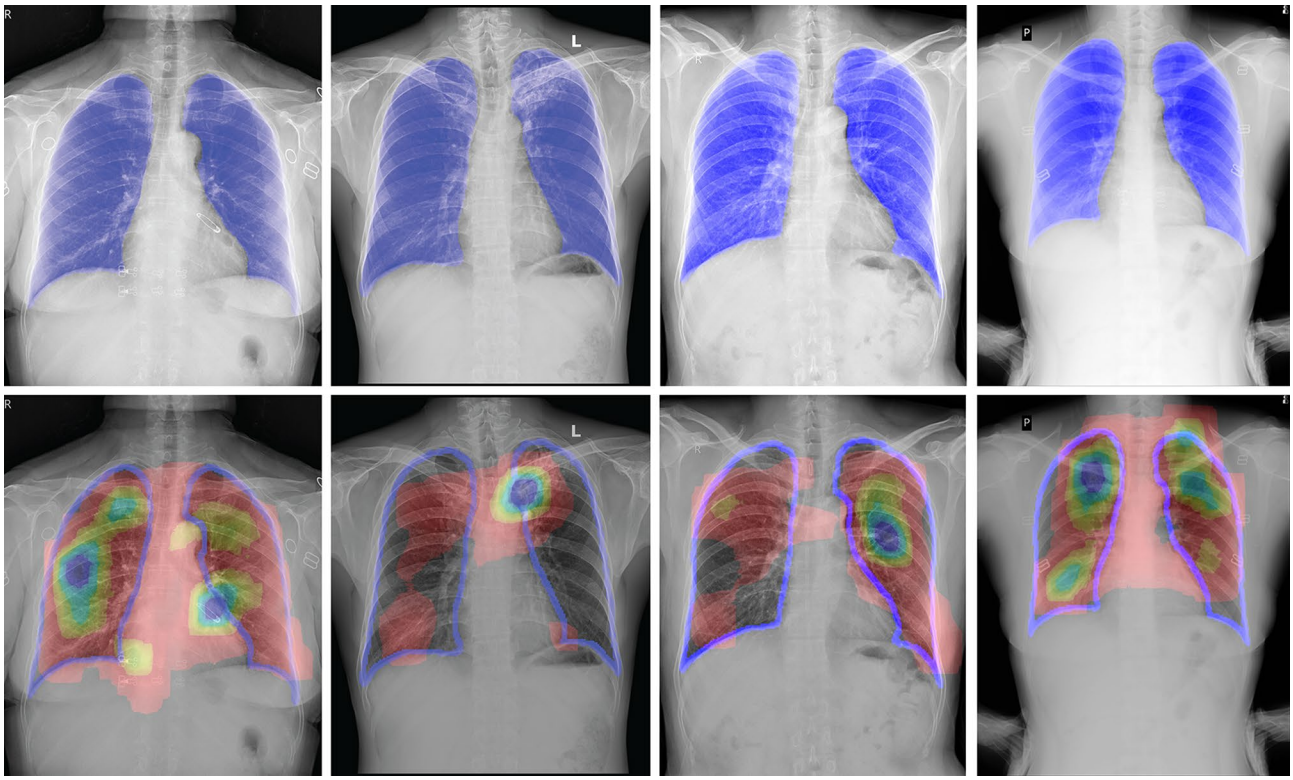

Fig. 1 Example of lung segmentation results (first row) and radiologists' gaze maps (second row) superimposed over four randomly selected X-rays. The segmentations were automatically generated using a modified U-Net algorithm. The heatmaps were recorded during X-ray reading

experiment, and to other radiologists closer to the end of the experiment (Fig. 2).

The image randomization was incorporated into a one-day X-ray reading experiment conducted by each radiologist. It was expected that radiological fatigue will grow during the experiment and the reading quality and potentially

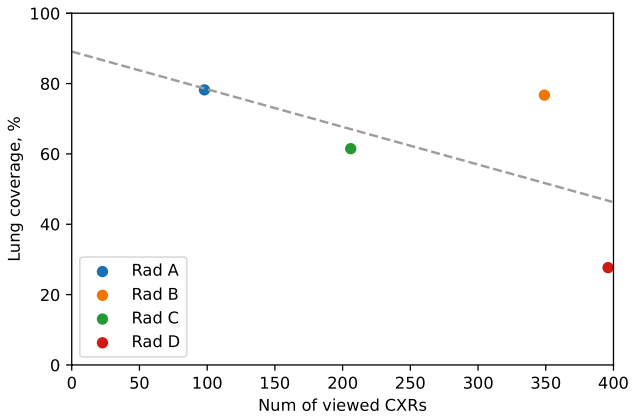

Fig. 2 The radiologists' performance on a randomly selected chest X-ray. The X-ray was given to radiologists at different time points, so each point (blue, red, orange, and green) corresponds to the lung coverage with gaze for a particular radiologist. The time points defined as the number of chest X-rays analyzed by the radiologist before he got the target X-ray define the *x*-axis

X-ray coverage with gaze will deteriorate with the number of images read. Considering that radiologists cannot analyze 400 X-rays without rest while continuously dictating the diagnoses, the same work-rest protocol has been introduced to all radiologists minimizing the potential influence of variable breaks on the experiment outcomes. The breaks were introduced after each batch of 100 X-rays was analyzed. At the start of the experiment and after each break, a short calibration session was carried out. During calibration, the position of the radiologist was selected, which would be comfortable for them throughout the analysis of 100 images. The framework was equipped with a module that controls that the reader's eyes are inside the eye tracker capture range. If the reader's eyes come to the borders of this capture range, the reader hears a warning sound. Each radiologist analyzed 100 X-rays, then had a short break, then analyzed the next 100 X-rays, then had a 40-min lunch break, then analyzed the third batch of 100 X-rays, then had a short break, and finally analyzed the last batch of 100 X-rays. During breaks, the radiologists also passed various fatigue/concentration tests. In particular, they filled out the oculomotor part of the Simulator Sickness Questionnaire (SSQ) to self-evaluate their level of general discomfort, headache, fatigue, blurred vision, eye strain, etc. This test was passed at the beginning and end of the experiment, and before and after the lunch break. During other breaks, the radiologists also passed

digital concentration tests, namely, digit symbol substitution, circle coverage, and reaction time tests [28, 29]. The results of the tests and their correlation with the number of analyzed X-rays are presented in our previous study [30].

Following the experiment protocol, the X-rays were separated into four batches with 100 X-rays in each batch. The proportion of healthy and pathological subjects in each batch reflected the proportion of healthy and pathological samples in the complete database. The batches were shown to radiologists in random order. Moreover, the X-rays inside batches were randomly shuffled for each radiologist.

As a consequence of this randomized experiment, each X-ray was analyzed by four radiologists at different time points. This resulted in four measurements of lung coverage with radiologists' gaze. We fitted a linear regression model to these measurements to estimate how the coverage changes for this particular X-ray depending on the time this image was viewed (Fig. 2).

Statistical Analysis

All linear regression models used to evaluate coverage results were augmented with 95% confidence intervals estimated with the bootstrap method [31]. The X-ray shuffling during the experiment was performed using the Fisher-Yates algorithm.

Experiment and Results

Experiment Details

Four practicing radiologists with experience ranging from 3 to 30 years were recruited to participate in the experiment. In his everyday practice, radiologist A analyzes both X-ray and CT images. Radiologist B analyzes only CT images and does not work with X-rays. In contrast, radiologists C and D analyze only X-ray images. We did not inform radiologists about the experiment aims to remove the risks that they will involuntarily try to change their image reading behavior. At the same time, they were informed that we will ask them to read 400 X-rays and dictate the diagnoses, which will be then compared against reference diagnoses. They were also informed about certain tests they need to pass during the experiment.

The diagnoses of the participating radiologists were manually extracted from voice recordings. The radiologists were asked to mention all abnormalities they observe and comment on the confidence they have in their decision. The decision was considered correct if the radiologist mentions, potentially among other abnormalities, the abnormality that was unanimously diagnosed by the reference radiological team [18]. Examples of lung segmentation with superimposed gaze heatmaps for several random cases are given in Fig. 1.

Results

The average ($\pm$ standard deviation) lung coverage was $64 \pm 16\%$, $65 \pm 14\%$, $58 \pm 17\%$, and $55 \pm 17\%$ for radiologists A, B, C, and D, respectively. The linear regression models fitted to the lung coverage data had slope coefficients -0.026, -0.013, -0.076, and -0.04 for radiologists A, B, C, and D, respectively. In other words, radiologist C covers 1% less of the lung fields after reading every 13 X-rays ($1/0.076 \approx 1$). Figure 3 depicts the regression models with the confidence intervals shaded. Linear regression models were also computed for each abnormality type (Fig. 3b). The slope coefficients were -0.041, -0.041, -0.035, -0.034, -0.034, and -0.041 for atelectasis, infiltration, pneumothorax, nodule/mass, cardiomegaly, and healthy chest X-rays, respectively. Figure 4 presents the box-whisker plots that measure the lung coverage for each abnormality and each radiologist.

Using the average slope of the linear models fitted to individual X-ray readings, we evaluated whether the lung coverage with gaze for each radiologist is on average above or below the overall trendline. In other words, we used the order number of an X-ray to calculate the expected gaze coverage and compared it with the observed gaze coverage for each radiologist. We observed that radiologist A is above the trendline in 64.5% of cases; that is, he covers a larger part of lung fields than is expected from an average radiologist for 64.5% of the analyzed X-rays. Radiologist B was above the trendline in 66.3% of cases, radiologist C — in 42% of cases, and radiologist D — in 38.5% of cases. Considering that radiologists can be fast/slow readers, we run an ablation analysis of artificially increasing/decreasing lung coverage for different radiologists. The analysis aim was to confirm that the trends observed for individual diseases (Fig. 3b) are invariant to radiologists' reading styles. In particular, we reduced the lung coverage by 10% for radiologists A and B or increased the lung coverage by 10% for radiologists C and D. Then, we computed the lung coverage slopes using the presented above methodology. The trends remain very consistent with the average reduction of lung coverage remaining in a narrow interval of 3.9–4.1% per 100 X-rays (Fig. 5).

The self-reported SSQ test results were recorded for radiologists B–D, while not recorded for radiologist A due to a software error. The utilized oculomotor part of the SSQ test result ranges from 7, indicating the lowest level of fatigue from all test subparts, to 28, indicating the maximal self-reported fatigue. All radiologists B–D have reported no fatigue at the start of the experiment; that is, they graded all seven oculomotor SSQ measurements with the lowest available grade "1." After analyzing 200 X-rays, the average self-reported fatigue grew from 7 to 11 ± 1.4. The lunch break returned the fatigue to a close-to-original level of 7.7 ± 0.5. At the end of the experiment, the average self-reported fatigue was 12.7 ± 2.1.

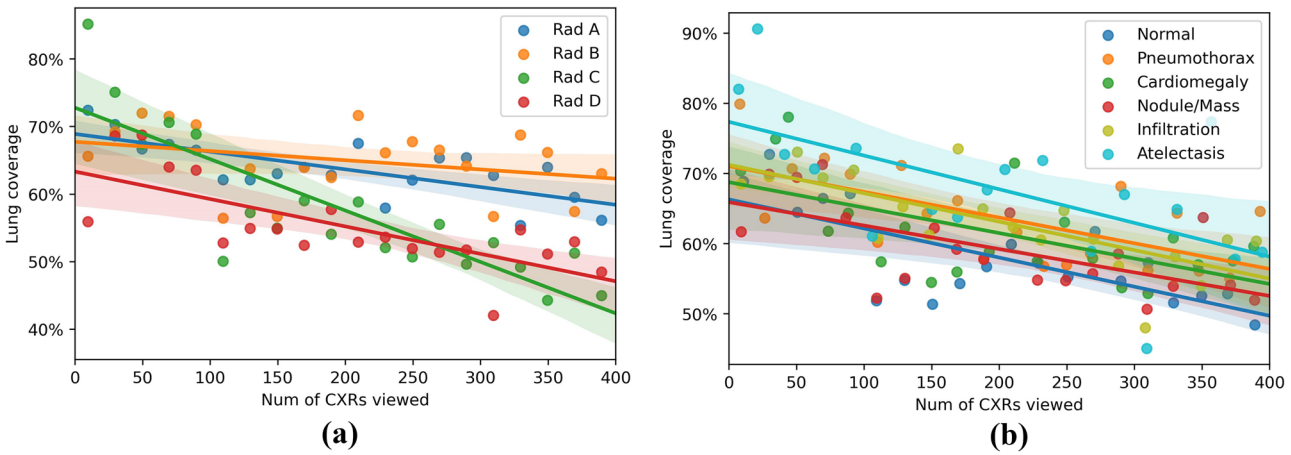

Fig. 3 These illustrations demonstrate how lung coverage with radiologists' gaze changes with the number of images analyzed by the radiologists. **a** Linear regression models fitted to the lung coverage metric computed for each of the four radiologists participating in the experiment. **b** Linear regression models fitted to gaze coverage data for all radiologist computed for each lung X-ray. These linear regression models are then aggregated for individual abnormalities. The linear regression models are overlapped with the corresponding data points; the points are averaged over 20 data samples for better visibility

Discussion

In this paper, we described a randomized experiment to capture the potential changes in chest X-ray reading patterns of radiologists by utilizing eye-tracking and AI-based lung segmentation. We hypothesized that there will be a trend of reduced quality of X-ray readings at the end of the experiment when radiologists are more tired. The reduction of the reading time at the end of the work shifts has been observed previously and can vary from insignificant for bone fracture diagnosis [11] to more than a 30% reduction for CT colonography [32]. At the same time, it is also observed that the reading time depends significantly on the abnormality type [11]. It was therefore essential to separate the reading changes related to fatigue from the reading changes related to X-ray complexity. We designed a data randomization protocol for the chest X-ray reading, which allowed us to measure the statistical changes in chest coverage for individual radiologists and individual lung abnormalities. By computing the changes in lung coverage for each radiologist

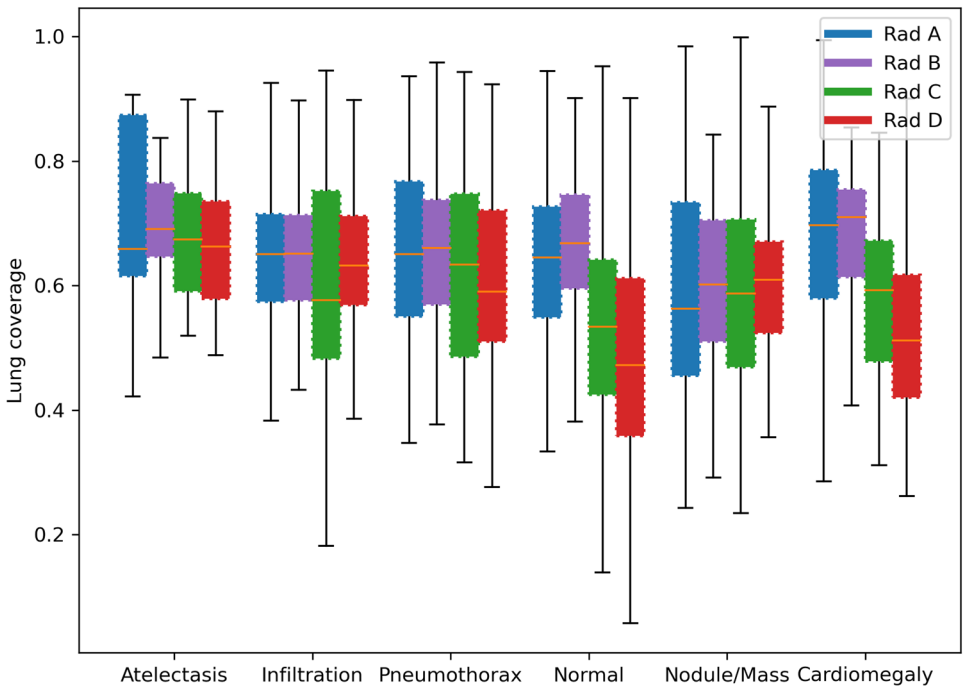

Fig. 4 Box-whisker plots show the lung coverage of an X-ray with a specific pathology by each radiologist. The orange line shows the median, the whiskers show the interquartile range, the boxes extend from the first to the third quartile, and the whiskers extend to the 1.5×of the interquartile range. Outliers outside whiskers are not visualized for figure clarity

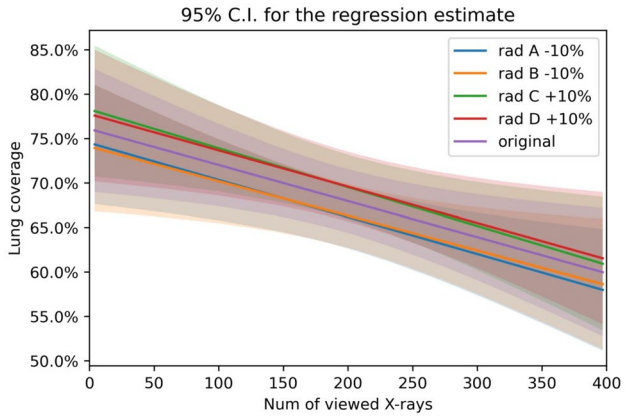

Fig. 5 An ablation study demonstrating the model that captures the lung coverage with gaze against the number of images viewed by radiologists. Each regression model computes the changes in lung coverage over all X-rays using the data from all radiologists. Alternative regression models were generated by artificially increasing/reducing the lung coverage for some radiologists

(Fig. 3a), the intuitive assumption that the lung coverage for all radiologists reduces with the growth of the X-rays analyzed has been confirmed. At the same time, this trend was much stronger for radiologists C and D and less pronounced for radiologists A and B. This fact could be potentially explained by induvial reading style of radiologists A and B, or by the coincidence, where radiologists A and B got more challenging-to-diagnose X-rays at the end of the experiment while radiologists C and D — more challenging-to-diagnose X-rays at the beginning of the experiment. Using binomial distribution, we estimated the probability of getting 10 more challenging-to-diagnose X-rays images at the beginning/end of the experiment to be above 50%.

The regression models fitted to individual images (Fig. 3b) demonstrate that the coverage trends are almost the same for each abnormality, ranging from $-0.54e^{-4}$ for lung infiltrations to $-0.42e^{-4}$ for cardiomegaly. The lung coverage with gaze approximately reduces by 20% after reading 400 X-rays. This confirms the assumption that the coverage drop is similar for all radiologists and the discrepancies we saw in Fig. 3a are due to different orders of X-rays assigned to each radiologist. The radiologists seem to cover a larger part of the lungs with their gaze when the patient has atelectasis. One of the possible explanations is that nonobstructive atelectasis is caused by lung diseases such as pleural effusion, pneumonia, pneumothorax, and fibrosis [33]. The radiologists, therefore, need to search more thoroughly for comorbidities if they discover atelectasis. It is also important to note that the lung coverage for healthy X-rays was on average lower than for the X-rays with pathologies.

Radiologists C and D on average cover smaller lung parts with their gaze than radiologists A and B. This observation may correlate with the level of expertise of the participants, as both radiologists C and D specialize exclusively in X-ray image analysis, while, in contrast, CT-specializing radiologist B demonstrated the highest level of lung coverage. Moreover, radiologist D with 30 years of practice is the most experienced reader in the study. The literature review confirms our observation of a negative correlation between reading time and the level of expertise. Manning et al. [34] documented that radiologists read a chest X-ray in around 30 s, while novices spend around 41 s per X-ray. Burling et al. [32] found that experienced radiologists spend on average 11 s to identify and interpret colorectal cancer and detect colon polyps or their absence in CT images. This performance compares favorably to the 16 s needed for novice radiologists, and the 17 s needed for radiographic technicians. Wood et al. [35] compared the performance of musculoskeletal radiologists on pelvic, spine, palm, wrist, elbow, and ankle X-ray analysis and found that experts read the images almost two times faster than novices. Similar trends are observed if reading time is replaced with anatomy coverage [17]. Manning et al. [34] separated chest X-rays into 14 anatomical zones and observed that radiologists on average cover 12 zones per reading, while novices — 12.5 zones. Drew et al. [36] separated the radiologists into drillers, which use bottom-up reading, and scanners, which use top-down reading and compared their performance [37]. The authors found out that the drillers on average have 5 years less experience but cover 5% more lung volume with the gaze. It is, however, important to note that only 20% of doctors were considered scanners, which may have skewed the results. Rubin et al. [38] observed that expert radiologists cover a 50% smaller portion of the lung volumes during decision-making in comparison to 1st-year radiology residents. In our study, the tendency that more experienced radiologists to cover a smaller portion of lung fields with gaze persists during the whole experiment. The most experienced radiologist D demonstrates the lowest lung coverage normalized to the X-ray viewing order in comparison to other participating radiologists.

One of the main advantages of the proposed randomized experiment is its ability to discover consistent image reading patterns invariant to the viewing order of the X-rays for different radiologists. There is a visual tendency for lung coverage to reduce with the number of viewed images for all radiologists (Fig. 3a). Despite the overall trend, the lung coverage reduction significantly varied from 1.3% per 100 X-rays for radiologist B to 7.6% per 100 X-rays for radiologist C. When computing trend image-wise, variability reduces dramatically to 3.4% per 100 X-rays for cardiomegaly and 4.1% per 100 X-rays for atelectasis. We wanted to additionally investigate whether this consistency is invariant to the performance of individual radiologists by artificially modifying their lung coverage trends (Fig. 3a). The artificial reduction/increase of coverage almost did not affect the

trend lines (Fig. 5). Such invariance is expected and can be explained using an example. Suppose the lung coverage for radiologist A was reduced by 10% of their average lung coverage. The slopes of the linear models (Fig. 3), where radiologist A reads the X-rays in the first half of the experiment, will reduce. The slopes of the linear models, where radiologist A reads the X-rays in the second half of the experiment, will increase. These changes largely compensate for each other resulting in a stable slope of the averaged regression model. The ablation analysis indicates that the changes we see in Fig. 3b are not only applicable to radiologists A–D from our experiment but also likely capture the overall reading pattern.

It is important to appreciate that the reported absolute values of the lung coverage depend on the view angle $\theta = 2°$. The assumption on the view angle is in agreement with other studies on eye tracking in radiology [25–27]. This, however, does not mean that radiologists do not respond to any visual stimuli outside the view angle as their response may depend on the type of the depicted abnormality, its relative appearance, and size. We can assume that visual information capturing depends on more parameters than a single view angle. On the other hand, the proposed study aims to find the trends, which are relatively invariant to the view angle parameter. The discovery of such trends opens new directions for eye-tracking integration into clinical practice. By recording the changes in organ coverage with gaze, we can potentially recognize the moments when a radiologist is tired and a second opinion may be needed.

The presented study has limitations. The prevalence of X-rays with abnormalities in the analyzed database does not necessarily match a typical X-ray composition in a radiology department. Moreover, radiologists have individual work schedules and can take short breaks outside predefined intervals. Another limitation is related to the gaze coverage approach. A fixed view angle value was used; however, this may differ for each radiologist. In addition, the peripheral vision was not considered when calculating eye coverage. Future discoveries and clarifications in the field of visual information perception will allow us to refine the absolute values of the lung field coverage and potentially discover additional patterns in chest X-ray readings.

Acknowledgements We thank Khanov A. N. MD and Zinnurov A. R. MD for participating in the experiment.

Author Contribution The study was designed by Bulat Ibragimov. Software for eye tracking and medical image analysis was developed by Ilya Pershin. Clinical experiments were developed and supervised by Tamerlan Mustafaev and Dilyara Ibragimova. The first draft of the manuscript was written by Bulat Ibragimov and Ilya Pershin, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This research has been financially supported by The Analytical Center for the Government of the Russian Federation (Agreement No. 70-2021-00143 dd. 01.11.2021, IGK 000000D730321P5Q0002).

Declarations

Conflict of Interest The authors declare no competing interests.

References

1. R. J. M. Bruls and R. M. Kwee, Workload for radiologists during on-call hours: dramatic increase in the past 15 years, *Insights Imaging*, vol. 11, no. 1, p. 121, Nov. 2020. https://doi.org/10.1186/s13244-020-00925-z.
2. A. M. Koc, L. Altin, T. Acar, A. Ari, and Z. H. Adibelli, How did radiologists' diagnostic performance has changed in COVID-19 pneumonia: a single-centre retrospective study, *Int. J. Clin. Pract.*, vol. 75, no. 10, Art. no. 10, Oct. 2021. https://doi.org/10.1111/ijcp.14693.
3. R. Alexander et al., Mandating limits on workload, duty, and speed in radiology, *Radiology*, vol. 304, no. 2, pp. 274–282, Aug. 2022. https://doi.org/10.1148/radiol.212631.
4. E. Ranschaert, L. Topff, and O. Pianykh, Optimization of radiology workflow with artificial intelligence, *Radiol. Clin.*, vol. 59, no. 6, Art. no. 6, Nov. 2021. https://doi.org/10.1016/j.rcl.2021.06.006.
5. J. Born et al., On the role of artificial intelligence in medical imaging of COVID-19, *Patterns N. Y. N*, vol. 2, no. 6, p. 100269, Jun. 2021. https://doi.org/10.1016/j.patter.2021.100269.
6. R. W. Filice and R. M. Ratwani, The case for user-centered artificial intelligence in radiology, *Radiol. Artif. Intell.*, vol. 2, no. 3, Art. no. 3, May 2020. https://doi.org/10.1148/ryai.2020190095.
7. E. Sorantin et al., The augmented radiologist: artificial intelligence in the practice of radiology, *Pediatr. Radiol.*, Oct. 2021. https://doi.org/10.1007/s00247-021-05177-7.
8. L. Lévêque, H. Bosmans, L. Cockmartin, and H. Liu, State of the art: eye-tracking studies in medical imaging, *IEEE Access*, vol. 6, pp. 37023–37034, 2018. https://doi.org/10.1109/ACCESS.2018.2851451.
9. Y. W. Kim and L. T. Mansfield, Fool me twice: delayed diagnoses in radiology with emphasis on perpetuated errors, *Am. J. Roentgenol.*, vol. 202, no. 3, Art. no. 3, Mar. 2014. https://doi.org/10.2214/AJR.13.11493.
10. T. N. Hanna et al., The effects of fatigue from overnight shifts on radiology search patterns and diagnostic performance, *J. Am. Coll. Radiol. JACR*, vol. 15, no. 12, pp. 1709–1716, Dec. 2018. https://doi.org/10.1016/j.jacr.2017.12.019.
11. E. A. Krupinski, K. S. Berbaum, R. T. Caldwell, K. M. Schartz, and J. Kim, Long radiology workdays reduce detection and accommodation accuracy, *J. Am. Coll. Radiol.*, vol. 7, no. 9, pp. 698–704, Sep. 2010. https://doi.org/10.1016/j.jacr.2010.03.004.
12. H. Zhan, K. Schartz, M. E. Zygmont, J.-O. Johnson, and E. A. Krupinski, The impact of fatigue on complex CT case interpretation by radiology residents, *Acad. Radiol.*, vol. 28, no. 3, Art. no. 3, Mar. 2021. https://doi.org/10.1016/j.acra.2020.06.005.
13. B. Hosp, M. S. Yin, P. Haddawy, R. Watcharopas, P. Sa-ngasoongsong, and E. Kasneci, Differentiating surgeons' expertise solely by eye movement features, in *Companion Publication of the 2021 International Conference on Multimodal Interaction*, New York, NY, USA, Oct. 2021, pp. 371–375. https://doi.org/10.1145/3461615.3485437.
14. T. Tien, P. H. Pucher, M. H. Sodergren, K. Sriskandarajah, G.-Z. Yang, and A. Darzi, Differences in gaze behaviour of expert and junior surgeons performing open inguinal hernia repair, *Surg.*

Endosc., vol. 29, no. 2, Art. no. 2, Feb. 2015. https://doi.org/10.1007/s00464-014-3683-7.
15. T. T. Brunyé, M. D. Eddy, E. Mercan, K. H. Allison, D. L. Weaver, and J. G. Elmore, Pupil diameter changes reflect difficulty and diagnostic accuracy during medical image interpretation, *BMC Med. Inform. Decis. Mak.*, vol. 16, p. 77, Jul. 2016. https://doi.org/10.1186/s12911-016-0322-3.
16. N. Castner et al., Pupil diameter differentiates expertise in dental radiography visual search, *PLOS ONE*, vol. 15, no. 5, Art. no. 5, May 2020. https://doi.org/10.1371/journal.pone.0223941.
17. A. van der Gijp et al., How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology, *Adv. Health Sci. Educ. Theory Pract.*, vol. 22, no. 3, pp. 765–787, Aug. 2017. https://doi.org/10.1007/s10459-016-9698-1.
18. H. Q. Nguyen, VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations, *Scientific Data*, vol. 9, no. 429, (2022). https://doi.org/10.13026/3akn-b287.
19. M. Kholiavchenko et al., "Contour-aware multi-label chest X-ray organ segmentation," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 15, no. 3, pp. 425–436, Mar. 2020. https://doi.org/10.1007/s11548-019-02115-9.
20. O. Ronneberger, P. Fischer, and T. Brox, U-Net: convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Cham, 2015, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.
21. M. Kholiavchenko et al., Contour-aware multi-label chest X-ray organ segmentation, *Int. J. Comput. Assist. Radiol. Surg.*, vol. 15, no. 3, pp. 425–436, Mar. 2020. https://doi.org/10.1007/s11548-019-02115-9.
22. S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, Aggregated residual transformations for deep neural networks, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 5987–5995. https://doi.org/10.1109/CVPR.2017.634.
23. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.
24. B. van Ginneken, M. B. Stegmann, and M. Loog, Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database, *Med. Image Anal.*, vol. 10, no. 1, pp. 19–40, Feb. 2006. https://doi.org/10.1016/j.media.2005.02.002.
25. H. L. Kundel, C. F. Nodine, D. Thickman, and L. Toto, Searching for lung nodules. A comparison of human performance with random and systematic scanning models, *Invest. Radiol.*, vol. 22, no. 5, pp. 417–422, May 1987. https://doi.org/10.1097/00004424-198705000-00010.
26. J. M. Wolfe, C.-C. Wu, J. Li, and S. B. Suresh, What do experts look at and what do experts find when reading mammograms?, *J. Med. Imaging Bellingham Wash*, vol. 8, no. 4, p. 045501, Jul. 2021. https://doi.org/10.1117/1.JMI.8.4.045501.
27. H. Strasburger, I. Rentschler, and M. Jüttner, Peripheral vision and pattern recognition: a review, *J. Vis.*, vol. 11, no. 5, p. 13, Dec. 2011. https://doi.org/10.1167/11.5.13.
28. J. Jaeger, Digit symbol substitution test, *J. Clin. Psychopharmacol.*, vol. 38, no. 5, pp. 513–519, Oct. 2018. https://doi.org/10.1097/JCP.0000000000000941.
29. H. C. Becker, W. J. Nettleton, P. H. Meyers, J. W. Sweeney, and C. M. Nice, Digital computer determination of a medical diagnostic index directly from chest X-ray images, *IEEE Trans. Biomed. Eng.*, vol. BME-11, no. 3, pp. 67–72, Jul. 1964. https://doi.org/10.1109/TBME.1964.4502309.
30. I. Pershin, M. Kholiavchenko, B. Maksudov, T. Mustafaev, D. Ibragimova, and B. Ibragimov, "Artificial Intelligence for the Analysis of Workload-Related Changes in Radiologists' Gaze Patterns," *IEEE J. Biomed. Health. Inform.*, vol. 26, no. 9, pp. 4541–4550, 2022. https://doi.org/10.1109/JBHI.2022.3183299.
31. B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, 1st edition. New York: Chapman and Hall/CRC, 1993.
32. D. Burling et al., CT colonography interpretation times: effect of reader experience, fatigue, and scan findings in a multi-centre setting, *Eur. Radiol.*, vol. 16, no. 8, pp. 1745–1749, Aug. 2006. https://doi.org/10.1007/s00330-006-0190-9.
33. A. E. O'Donnell, 90 — bronchiectasis, atelectasis, cysts, and localized lung disorders, in *Goldman's Cecil Medicine (Twenty Fourth Edition)*, L. Goldman and A. I. Schafer, Eds. Philadelphia: W.B. Saunders, 2012, pp. 548–552. https://doi.org/10.1016/B978-1-4377-1604-7.00090-7.
34. D. Manning, S. C. Ethell, and T. Crawford, Eye-tracking AFROC study of the influence of experience and training on chest X-ray interpretation, in *Medical Imaging 2003: Image Perception, Observer Performance, and Technology Assessment*, May 2003, vol. 5034, pp. 257–266. https://doi.org/10.1117/12.479985.
35. G. Wood, K. M. Knapp, B. Rock, C. Cousens, C. Roobottom, and M. R. Wilson, Visual expertise in detecting and diagnosing skeletal fractures, *Skeletal Radiol.*, vol. 42, no. 2, pp. 165–172, Feb. 2013. https://doi.org/10.1007/s00256-012-1503-5.
36. T. Drew, M. L.-H. Vo, A. Olwal, F. Jacobson, S. E. Seltzer, and J. M. Wolfe, Scanners and drillers: characterizing expert visual search through volumetric images, *J. Vis.*, vol. 13, no. 10, p. 3, Aug. 2013. https://doi.org/10.1167/13.10.3.
37. C. E. Connor, H. E. Egeth, and S. Yantis, Visual attention: bottom-up versus top-down, *Curr. Biol.*, vol. 14, no. 19, pp. R850–R852, Oct. 2004. https://doi.org/10.1016/j.cub.2004.09.041.
38. G. D. Rubin et al., Characterizing search, recognition, and decision in the detection of lung nodules on CT scans: elucidation with eye tracking, *Radiology*, vol. 274, no. 1, pp. 276–286, Jan. 2015. https://doi.org/10.1148/radiol.14132918.