scientific data

Check for updates

OPEN

DATA DESCRIPTOR

REFLACX, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays

Ricardo Bigolin Lanfredi[1] ✉, Mingyuan Zhang[2], William F. Auffermann[3], Jessica Chan[3], Phuong-Anh T. Duong[3], Vivek Srikumar[4], Trafton Drew[5], Joyce D. Schroeder[3] & Tolga Tasdizen[1]

Deep learning has shown recent success in classifying anomalies in chest x-rays, but datasets are still small compared to natural image datasets. Supervision of abnormality localization has been shown to improve trained models, partially compensating for dataset sizes. However, explicitly labeling these anomalies requires an expert and is very time-consuming. We propose a potentially scalable method for collecting implicit localization data using an eye tracker to capture gaze locations and a microphone to capture a dictation of a report, imitating the setup of a reading room. The resulting REFLACX (Reports and Eye-Tracking Data for Localization of Abnormalities in Chest X-rays) dataset was labeled across five radiologists and contains 3,032 synchronized sets of eye-tracking data and timestamped report transcriptions for 2,616 chest x-rays from the MIMIC-CXR dataset. We also provide auxiliary annotations, including bounding boxes around lungs and heart and validation labels consisting of ellipses localizing abnormalities and image-level labels. Furthermore, a small subset of the data contains readings from all radiologists, allowing for the calculation of inter-rater scores.

Background & Summary

Deep learning has been successfully applied to medical image analysis, including abnormality detection in chest x-rays (CXRs)[1,2]. However, even though, in the medical image context, very large published CXRs datasets are available[3–6], with sizes in the hundreds of thousands of images, they are much smaller than some natural images datasets[7,8]. These large CXRs datasets were labeled through the mining of clinical reports. However, given the relatively small size of the datasets, additional labels may contribute to the training of data-demanding deep convolutional neural networks. We present a dataset named REFLACX containing eye-tracking data collected from radiologists while they dictate reports. This dataset was built as a proof-of-concept for a data collection method that expands the labels of a medical dataset, providing additional supervision during deep learning training.

Li *et al.* have shown that bounding boxes localizing abnormalities on CXRs can be used to supervise a convolutional neural network (CNN) to improve accuracy and localization scores[9]. They used the manually annotated bounding boxes provided for 880 images of the 112,200 CXRs from the ChestX-ray14 dataset[5]. Even though other datasets provide similar labels[10,11], such localization labels are rare, and when present, are usually provided in limited quantities, showing the difficulty of scaling up the manual labeling. To try to solve this issue, and in accordance with the prioritization of the development of automated image labeling and annotation methods of the NIH's roadmap for AI in medical imaging[12], we collected eye-tracking data from radiologists for implicit localization of anomalies. We believe that the proposed collection method has the potential to be scaled up and used as a nonintrusive annotation approach deployed in a radiology reading room.

Other works have used eye-tracking data to support models in training or inference. Templier *et al.* and Stember *et al.* used eye tracking for interactive segmentation of biological/medical images, intending to achieve faster labeling[13,14]. Khosravan *et al.* proposed a method of integrating an eye tracker into the reading room to

[1]Scientific Computing and Imaging Institute, University of Utah, 72 S Central Campus Drive, Room 3750, Salt Lake City, UT, 84112, USA. [2]Department of Population Health Sciences, University of Utah, 295 Chipeta Way, Williams Building, Room 1N410, Salt Lake City, UT, 84108, USA. [3]Department of Radiology and Imaging Sciences, University of Utah, 30 North 1900 East #1A071, Salt Lake City, UT, 84132, USA. [4]School of Computing, University of Utah, Room 3190, 50 Central Campus Dr., Salt Lake City, UT, 84112, USA. [5]Department of Psychology, University of Utah, 380 S 1530 E Beh S 502, Salt Lake City, UT, 84112, USA. ✉e-mail: ricbl@sci.utah.edu

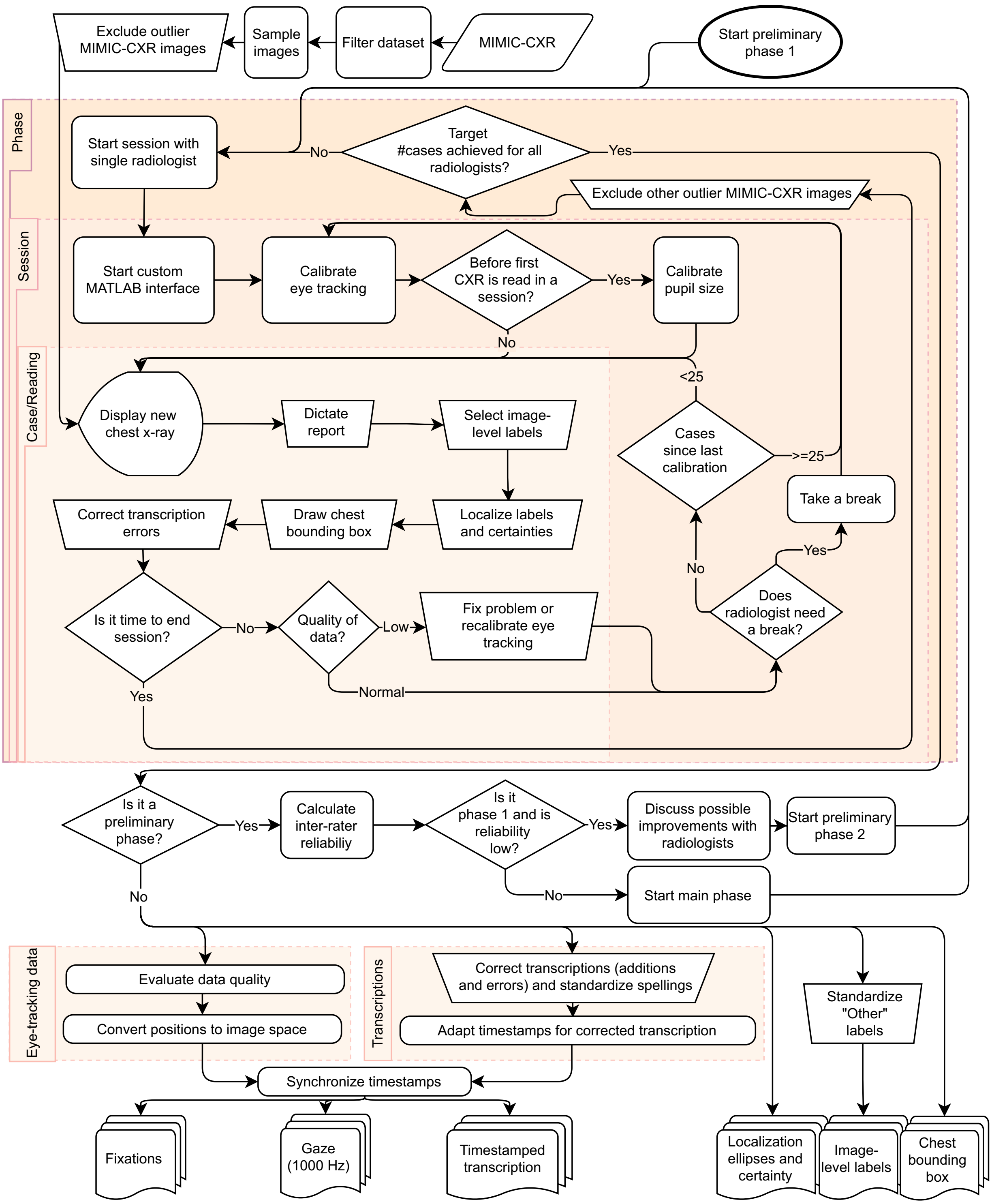

Exclude outlier MIMIC-CXR images
Sample images
Filter dataset
MIMIC-CXR
Start preliminary phase 1
Phase
Start session with single radiologist
No
Target #cases achieved for all radiologists?
Yes
Exclude other outlier MIMIC-CXR images
Session
Start custom MATLAB interface
Calibrate eye tracking
Before first CXR is read in a session?
Yes
Calibrate pupil size
No
<25
Case/Reading
Display new chest x-ray
Dictate report
Select image-level labels
Cases since last calibration
>=25
Take a break
Correct transcription errors
Draw chest bounding box
Localize labels and certainties
Yes
No
Is it time to end session?
No
Quality of data?
Low
Fix problem or recalibrate eye tracking
Does radiologist need a break?
Normal
Yes
Is it a preliminary phase?
Yes
Calculate inter-rater reliabiliy
Is it phase 1 and is reliability low?
Yes
Discuss possible improvements with radiologists
Start preliminary phase 2
No
No
Start main phase
Eye-tracking data
Evaluate data quality
Convert positions to image space
Transcriptions
Correct transcriptions (additions and errors) and standardize spellings
Adapt timestamps for corrected transcription
Standardize "Other" labels
Synchronize timestamps
Fixations
Gaze (1000 Hz)
Timestamped transcription
Localization ellipses and certainty
Image-level labels
Chest bounding box

Fig. 1 Overview of the steps in the building of the dataset.

combine a radiologist's reading with computer-aided diagnosis (CAD) systems[15]. Gecer *et al.* used the navigation behavior of pathologists during readings to improve the detection of cancer in histopathology images[16]. Stember *et al.* showed that the gaze of a radiologist when dictating standardized sentences for presence/absence of brain tumors in MRIs could be used to localize lesions[17]. In parallel to our project, Karargyris *et al.* built a dataset of eye-tracking data and respective dictation of CXRs reports, containing 1,083 readings by only one radiologist[18]. Saab *et al.* built a dataset of eye-tracking data for the task of identifying pneumothoraces for 1,170 CXRs[19]. Still, their dataset does not contain dictations.

To collect the proposed dataset, as described in Fig. 1, five board-certified subspecialty-trained thoracic radiologists, who closely worked on aspects of the study design, used a custom-built user interface mimicking some

Label	FK P1	FK P2	IoU P1	IoU P2
Airway wall thickening (P1)	0.03 ± 0.20		0.28 (n = 4)	
Atelectasis (P1, P2)	0.34 ± 0.05	0.25 ± 0.08	0.29 ± 0.03 (n = 33)	0.37 ± 0.04 (n = 19)
Consolidation (P1, P2)	0.36 ± 0.07	0.50 ± 0.07	0.36 ± 0.04 (n = 25)	0.37 ± 0.04 (n = 17)
Emphysema (P1) & High lung volume/emphysema (P2)	0.26 ± 0.32	0.10 ± 0.34	0.43 (n = 3)	0.53 (n = 2)
Enlarged cardiac silhouette (P1, P2)	0.56 ± 0.07	0.55 ± 0.08	0.75 ± 0.02 (n = 20)	0.75 ± 0.02 (n = 20)
Enlarged hilum (P2)		−0.03 ± 0.36		Undefined
Fibrosis (P1) & Interstitial lung disease (P2)	0.29 ± 0.44	0.16 ± 0.57	0.53 (n = 1)	0.23 (n = 1)
Fracture (P1) & Acute fracture (P2)	0.25 ± 0.25	0.71 ± 0.36	0.42 (n = 4)	0.21 (n = 1)
Groundglass opacity (P1, P2)	0.10 ± 0.17	0.21 ± 0.11	0.42 (n = 6)	0.38 ± 0.05 (n = 16)
Hiatal hernia (P2)		Undefined		Undefined
Mass (P1)	−0.01 ± 0.70		Undefined	
Nodule (P1)	0.29 ± 0.25		0.38 (n = 2)	
Lung nodule or mass (P2)		0.36 ± 0.49		0.83 (n = 1)
Pleural abnormality (P2)		0.67 ± 0.07		0.27 ± 0.02 (n = 18)
Pleural effusion (P1)	0.53 ± 0.06		0.37 ± 0.03 (n = 22)	
Pleural thickening (P1)	0.06 ± 0.40		0.29 (n = 1)	
Pneumothorax (P1, P2)	0.55 ± 0.25	0.62 ± 0.28	0.41 (n = 3)	0.27 (n = 3)
Pulmonary edema (P1, P2)	0.22 ± 0.13	0.10 ± 0.14	0.25 ± 0.03 (n = 11)	0.36 ± 0.06 (n = 11)
Quality issue (P1)	0.02 ± 0.30			
Support devices (P1, P2)	0.77 ± 0.05	0.47 ± 0.06		
Wide mediastinum (P1) & Abnormal mediastinal contour (P2)	0.23 ± 0.34	0.05 ± 0.25	0.57 (n = 2)	0.58 (n = 3)

Table 1. Inter-rater scores for validation of the quality of the data. For phases 1 (P1) and 2 (P2), we present reliability on image-level labels, calculated using Fleiss' Kappa (FK), and average IoU of the abnormality ellipses. All scores are paired with standard errors. The number of samples given for the IoU values represents the number of independent CXRs used in the calculation. The phases in which each label was present is listed in parenthesis. Table cells are left blank for labels that were not present in a specific phase.

of the functionalities from clinical practice. They dictated reports for images sampled from the MIMIC-CXR dataset[6] while eye-tracking data were collected, including gaze location and pupil area. The audio of the dictation was automatically transcribed and manually corrected. The transcriptions included word timestamps for synchronization with the rest of the data. During dictation, we also recorded the zooming, panning, and windowing state of the CXR at all times. After dictating each case, radiologists provided a set of manual labels: the abnormalities they found, selected from a list, and the location of those abnormalities. These manual labels were collected to validate the data collection method and any automatically extracted labels generated from the dataset. They are not easily scalable and would not be collected in more extensive implementations of the collection method. Radiologists also drew a bounding box around the heart and lungs for normalization of chest position.

The collection of data was separated into three phases. The two first preliminary phases were used to adjust minor data collection details and to estimate inter-rater scores for the labels of the dataset, with radiologists reading a shared set of images. For the third and main phase, the sets of CXRs for each radiologist were independent. Considering all phases, REFLACX contains 3,052 cases, for a total of 2,616 unique CXRs. Of the 3,052 cases, 3,032 contain eye-tracking data.

Methods

Data collection was divided into three phases. Phases 1 and 2 were preliminary phases where five radiologists read the same set of 59 and 50 CXRs, respectively. The set of possible image-level labels was chosen after a discussion among radiologists. After estimating the inter-rater reliability for phase 1, as shown in Table 1, a meeting was organized where radiologists discussed the labeling differences for the five cases that had the most negative impact on the reliability scores. The set of labels for phase 2 was slightly modified to clarify labeling and reduce its complexity. An electronic document, included as Supplementary File 1, was distributed to all radiologists with labeling instructions. A glossary by Hansell *et al.* provided some of the labeling definitions for this document[20]. Phase 3 had the same five radiologists reading independent sets of around 500 CXRs each and used the same set of anomaly labels as phase 2. This phase constitutes the main content of the dataset and has a slightly higher quality of eye-tracking data. The set of labels used for each phase is listed in Table 2. Phase 1 went from November 11, 2020 to January 4, 2021, phase 2 from March 1, 2021 to March 11, 2021, and phase 3 from March 24, 2021 to June 7, 2021. Data collection sessions took on average 2.21 hours, with a maximum of 3.92 hours and a minimum of 0.2 hours.

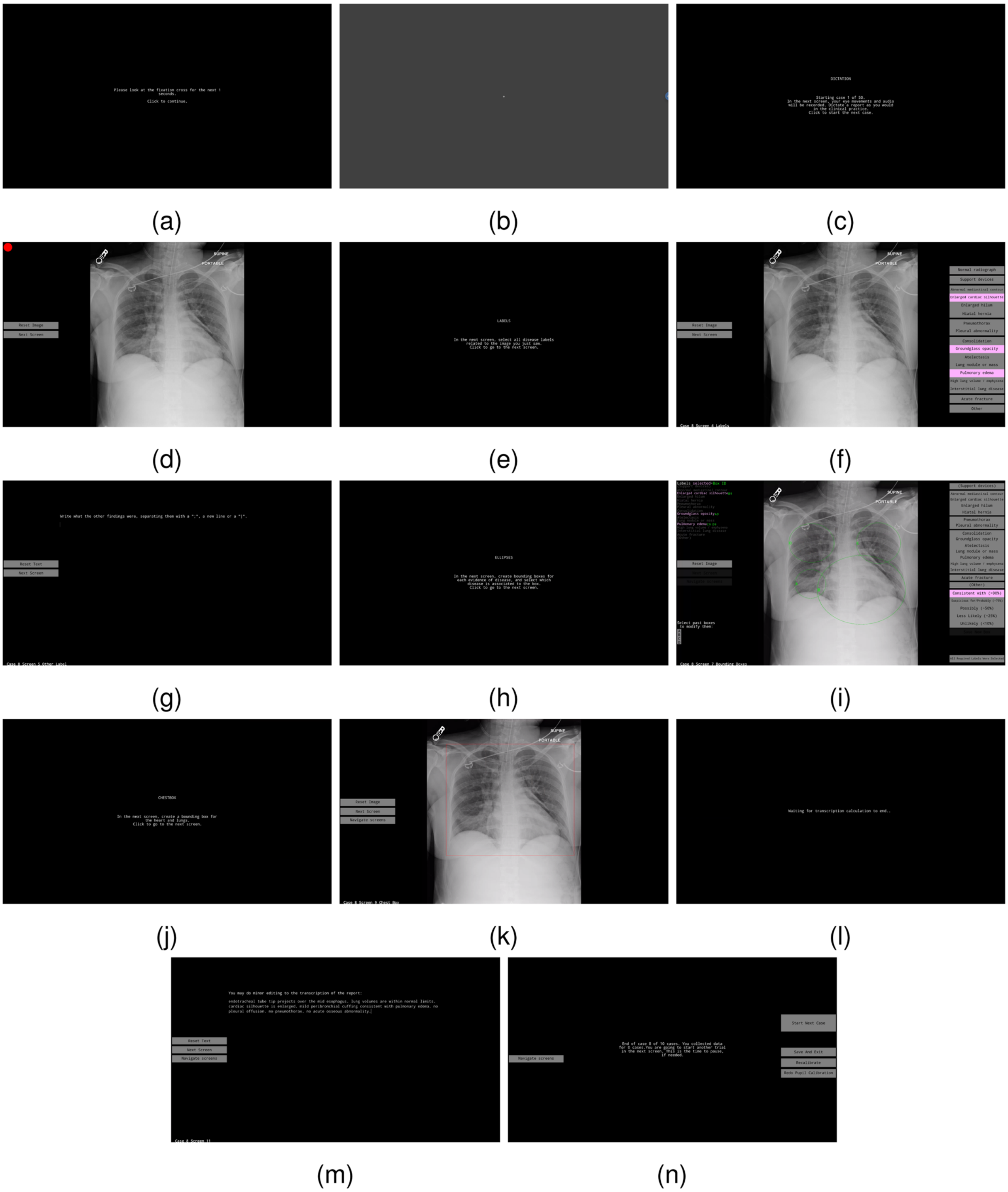
Dataset	Phase 1(P1)	Phase 2(P2)	Phase 3(P3)	MIMIC-CXR filtered (M)
# cases	295	250	2,507	194,495
# cases studies with eye-tracking data	285	240	2,507	0
# MIMIC-CXR images	59	50	2,507	
# subjects	58	50	2,110	60,018
% female	63.8	54.0	50.7	53.9
% male	36.2	46.0	49.1	45.7
% test set	15.3	14.0	20.2	1.4
% Normal Radiograph (P1, P2, P3) & No Finding (M)	18.0	24.4	22.8	32.9
% Abnormal mediastinal contour (P2,P3) & Wide mediastinum (P1)	2.7	5.6	2.7	
% Acute fracture (P2,P3) & Fracture (P1, M)	5.1	2.8	1.0	1.9
% Airway wall thickening (P1)	7.1			
% Atelectasis (P1,P2,P3,M)	41.4	27.6	25.8	20.5
% Cardiomegaly (M)				19.8
% Consolidation (P1,P2,P3,M)	28.5	28.8	25.9	4.7
% Enlarged cardiac silhouette (P1,P2,P3)	28.1	28.4	21.8	
% Enlarged Cardiomediastinum (M)				3.2
% Enlarged hilum (P2,P3)		2.8	1.9	
% Groundglass opacity (P1,P2,P3)	9.2	18.8	12.6	
% Hiatal hernia (P2,P3)		0.0	0.9	
% High lung volume/emphysema (P2,P3) & Emphysema (P1)	3.1	3.2	2.9	
% Interstitial lung disease (P2,P3) & Fibrosis (P1)	1.7	1.2	1.0	
% Lung nodule or mass (P2,P3) & Lung Lesion (M)		1.6	5.1	2.7
% Lung Opacity (M)				22.8
% Mass (P1)	0.7			
% Nodule (P1)	4.7			
% Other (P1,P2,P3)	13.9	8.8	6.0	
% Pleural abnormality (P2,P3)		30.0	29.5	
% Pleural Effusion (P1,M)	31.2			24.2
% Pleural thickening (P1)	2.0			
% Pleural Other (M)				0.9
% Pneumonia (M)				7.2
% Pneumothorax (P1,P2,P3,M)	4.7	4.4	2.9	4.6
% Pulmonary edema (P1,P2,P3) & Edema (M)	13.9	13.6	13.7	12.1
% Quality issue (P1)	3.4			
% Support devices (P1,P2,P3,M)	36.9	34.8	44.8	29.3

Table 2. Statistics of each phase of data collection and the subset of the MIMIC-CXR dataset from which images were sampled. The dataset where each label was present is shown inside parentheses. "Normal radiograph" represents CXRs for which no other label was selected. Table cells are left blank for labels that were not present in that dataset. For how the labels of the different datasets are related, check Fig. 5.

This study complied with all relevant ethical regulations. Eye-tracking data collection was exempted from approval by the University of Utah Institutional Review Board (IRB), and no informed consent was necessary. The use of the MIMIC-CXR CXRs was exempted from approval for this study since the images came from a publicly available and de-identified dataset. The MIMIC-CXR dataset was originally approved by the IRB of BIDMC, and patient consent was waived.

Image data. CXRs were sampled from the MIMIC-CXR dataset[6,21,22], a publicly available deidentified dataset. Before a random sampling of the images shown to radiologists, the dataset was filtered for including only CXRs that contained "ViewPosition" metadata with value "PA" or "AP", were the only frontal CXRs in their study, and were present in the label table from the MIMIC-CXR-JPG dataset[22–24]. Images were sampled to include 20% of images from the test set of the MIMIC-CXR dataset and 80% from the other splits, and uniformly at random from each of these two splits. After the sampling for phase 3, images with anonymizing black boxes that intersected the lungs were manually excluded before presentation to radiologists. After the sessions, we manually excluded outlier images that, according to the radiologists, had major parts of the lung missing from the field of view, were digitally horizontally flipped, or had a rotation of 90°.

Data-collection sessions. Data collection for each CXR involved two main parts: a dictation and transcription of a free-form radiology report while collecting eye-tracking data, and the selection of labels and ellipses to



Please look at the fixation cross for the next 1 seconds.
Click to continue.

DICTATION

Starting case 1 of 50.
In the next screen, your eye movements and audio will be recorded. Dictate a report as you would in the clinical practice.
Click to start the next case.

LABELS

In the next screen, select all disease labels related to the image you just saw.
Click to go to the next screen.

Write what the other findings were, separating them with a ";", a new line or a "|".

ELLIPSES

In the next screen, create bounding boxes for each evidence of disease, and select which disease is associated to the box.
Click to go to the next screen.

Select past boxes to modify them:

CHESTBOX

In the next screen, create a bounding box for the heart and lungs.
Click to go to the next screen.

Waiting for transcription calculation to end..

You may do minor editing to the transcription of the report:

endotracheal tube tip projects over the mid esophagus. lung volumes are within normal limits. cardiac silhouette is enlarged. mild peribronchial cuffing consistent with pulmonary edema. no pleural effusion. no pneumothorax. no acute osseous abnormality.

End of case 8 of 10 cases. You collected data for 0 cases.You are going to start another trial in the next screen. This is the time to pause, if needed.

Fig. 2 Screens of the data-collection interface in the sequence they are presented to a radiologist, including instruction screens (**a,c,e,h,j,n**), calibration of pupil size (**b**), dictation of reports (**d**), choice of global labels (**f,g**), selection of ellipses and certainties (**i**), drawing of lung/heart box (**k**), and editing of transcription (**l,m**). Digital visualization is recommended for reading the content.

use as evaluation ground-truth for anomalies present in a CXR and their location. The data collection interface was developed in MATLAB R2019a/Psychtoolbox 3.0.17[25–27]. The code for the interface is available at https://github.com/ricbl/eyetracking[28] The interface is shown in Fig. 2 and as a video in Supplementary File 2, where the moving semitransparent red ellipse, with an axis length of 1° of visual angle, represents stabilized gaze, i.e., fixations, and the moving blue ellipse represents the instantaneous gaze location sample. The cursor was not drawn in the video and the audio is a digitally generated voice representing the timestamped transcription. We did not include the original dictation in the video for anonymization purposes. This interface allowed for:

- Dictation of reports. A CXR was displayed to radiologists, and they dictated a report using a handheld microphone while eye-tracking data were collected. Radiologists did not have previous access to the CXRs or their reports. Eye-tracking data were collected as soon as the CXR was shown to radiologists, including data containing a reasoning period when they first saw the image and a dictation period. Radiologists were instructed to dictate reports as they would dictate in clinical practice. We also asked them to dictate free-form reports, so radiologists that used templates in clinical practice had to change their style to free-form.
- Editing of mistakes of the automated transcription. Radiologists were instructed to not add or change report content, and only correct transcription mistakes.
- Selection of image-level labels and ellipses localizing each label, including assessing the certainty of the presence of anomaly for each ellipse. Following Panicek *et al.*[29], the allowed certainties were: Unlikely (<10%), Less Likely (25%), Possibly (50%), Suspicious for/Probably (75%), Consistent with (>90%). Radiologists were instructed to add image-level labels even in case they forgot to mention them in the original report.
- Image windowing, zooming, and panning. These features were available while dictating and labeling, mimicking features available to radiologists in their clinical practice.
- Drawing of chest bounding boxes that encompass lungs and heart, for normalization of the variations in CXRs.
- Calibration of pupil size and access to the eye-tracker calibration screens.

In more detail, at the beginning of the session, the calibration interaction adopted the following order:

1. Eye-tracking calibration screen, where 13 circles, distributed throughout the screen, are displayed in random order for the radiologists to fixate.
2. Calibration of pupil size.

For the rest of the session, for each reading of CXR, the interaction of radiologists followed:

1. Dictation of report. Radiologists were not allowed to return to this screen for changes in dictation.
2. Image-level label selection.
3. Text input for “Other” label, in case it was selected as an image-level label.
4. Localization of image-level labels by the drawing of ellipses and choice of certainty for each ellipse.
5. Drawing of a chest bounding box including lung and heart.
6. Text editing to correct transcription errors.
7. Screen allowing proceeding to the subsequent CXR, displaying warnings for low-quality eye-tracking data, and optional return to calibration screens.

Equipment. To collect eye-tracking data, we used an Eyelink 1000 Plus system, which allows for high spatial (less than 1° of visual angle) and temporal (1,000 Hz) resolution. It also allows the radiologists to move their heads within a small area while maintaining this high degree of spatial and temporal acuity. Given our interest in ensuring that we had high-quality eye-tracking data, the experimenter calibrated and validated each radiologist multiple times throughout their viewing sessions. This process can be time-consuming, particularly if the clinician wears bifocals, which often lead to poor calibration. One alternative to this setup would have been to use a mobile eye-tracking system. These systems typically have the eye-tracking apparatus embedded within glasses worn by the subject. A good calibration can be achieved much more easily and possibly without an experimenter guiding this process. However, at present, the resultant data are much more difficult and time-consuming to analyze because most mobile eye-tracking systems do not co-register eye movements with precise screen coordinates. In practice, this typically means that the experimenters must hand-code the co-registration from a video output. This process is time-consuming and can suffer from bias, requiring multiple coders to examine the data to ensure reliable coding.

The Eyelink 1000 Plus was equipped with a 25 mm lens and managed by an Eyelink 1000 Host PC recording gaze at 1,000 Hz. The eye tracker was configured in the remote setup, for which the radiologists put a sticker on their forehead. Radiologists had the freedom to move their heads as long as the sticker and tracked eye stayed within the camera’s field of view, and the sticker stayed between 55 cm and 65 cm from the camera. The camera was positioned below a 27 inches BenQ PD2700U, 3,840 × 2,160 pixels, 60 Hz screen, connected to a Display PC running Ubuntu 18.04. For phase 3, the camera was positioned 11 cm in front of the screen. Eye distance to the bottom of the screen was around 71 cm, while it was around 65 cm to the top of the screen.

Calibration. Eye-tracker calibrations are necessary for finding the correspondence between pupil and cornea positions in the image captured by the eye tracker’s camera and locations on the screen where the radiologists are looking. During the calibration, radiologists had to look at 13 targets in several locations on the screen for the eye-tracker to register the pupil and cornea positions for each location and interpolate for other intermediate locations on the screen. We performed an eye-tracker calibration at the beginning of each session, every 25 cases, every time radiologists took a break, or when noticeable quality problems with the data could be solved with a new calibration. These quality problems mainly involved moments when the cornea or the pupil was not recognized for specific eye positions or when the eye-tracker gave incorrect locations for the pupil or the cornea, e.g., glasses recognized as the cornea. The experimenter identified them with access to real-time data-collection information on the Eyelink 1000 Host PC. During calibration, radiologists were positioned such that the forehead sticker was between 59 cm and 61 cm away from the camera. For phases 1 and 2, calibrations were not performed at a regular 25-case interval. For phase 3, the calibration was considered successful if the average error was less than 0.6°, and the maximum error was less than 1.5°. The left eye was tracked by default, and the right eye was only tracked when the left eye calibration was repeatedly faulty. At each session beginning, radiologists were asked to look at the center of the screen for 15 s for measuring a constant used to normalize the pupil size.

Report transcription. We collected the dictation audio at 48,000 Hz using a handheld PowerMic II microphone. The audio was transcribed using the IBM Watson Speech to Text service, which provides timestamps for each transcribed word, with the "en-US_BroadbandModel" model. Before phase 1, the service's custom language model was trained with sentences from the "Findings" and "Impressions" sections of the MIMIC-CXR reports, which were filtered to remove sentences that referenced other studies of a patient through the search of keywords. For phase 2 and phase 3, in addition to language training with the filtered MIMIC-CXR reports, models had language and acoustic training with the collected audio and corrected transcriptions from phase 1. Audio files had silence trimmed from the start and end of the file to speed up transcriptions. Silence was detected using Otsu's thresholding over the average audio level (dBFS) of 500 ms chunks. Word timestamps were adjusted for the trimming of the beginning of the audio. After the report transcription was received from the cloud service, radiologists could make minor changes to it. Several of the common mistakes of the cloud service were programmatically corrected before providing transcriptions to the radiologists.

Postprocessing. The eye-tracking gaze samples were parsed for fixations, i.e., locations where gaze was spatially stabilized for a certain period; blinks, i.e., moments when the eye tracker did not capture pupil or cornea; and saccades, i.e., fast eye movements between fixations. Parsing was done in real time by the EyeLink 1000 Host PC, using a saccade velocity threshold of $35°/s$, a saccade motion threshold of 0.2°, and a saccade acceleration threshold of $9{,}500°/s^2$. Fixation locations were converted from screen space to image space by recording, at the start of the fixation, what image part was shown at what screen section. Fixations were synchronized with the transcriptions and other recorded data by synchronization messages sent by the Display PC to the EyeLink 1000 Host PC.

Pupil area data, captured by the eye tracker, was normalized by the average value of pupil area from the calibration screen from the beginning of the session. Radiologists were asked to look directly at the center of the screen, marked by a cross. The average value was weighted by fixation durations and calculated only for fixations at most 2° from the screen center.

After all data-collection sessions, transcriptions were checked and corrected by another person, who looked for additions of content during the radiologist correction screen, which were not allowed, and for out-of-context words and other clear mistakes, which were corrected by relistening to the recorded audios of those cases. During this process, spellings of a few words were standardized. The labels listed as "Other" were also standardized. Since the transcriptions were corrected for mistakes after receiving the output from the cloud service, timestamps had to be adapted to the new set of words. We used the counting of syllables to perform a linear interpolation between the times of both texts. Interpolations were calculated for each difference found between strings, as given by the difflib Python library. When there was an addition of words with no removal, the new words used the time range defined by the end time of the previous word and the start time of the next one, making it possible for words to have the same timestamp for start and end.

Data quality. Readings were evaluated for the quality of eye-tracking data by measuring the times that data were classified as a blink during collection. Eye-tracking data were discarded in cases with a blink longer than 3 s or when blinks corresponded to more than 15% of the data. Warnings were shown between cases for blinks longer than 1.5 s or when blinks corresponded to more than 10% of the data. Cases that had eye-tracking data discarded were not included in the dataset for phase 3 but were included for phases 1 and 2 to make possible the evaluation of inter-rater scores for other labels. The threshold values were chosen by qualitative observation of blink data histograms from phases 1 and 2 before collecting phase 3 data. Eye-tracking data were also discarded when the radiologist unintentionally clicked the "Next Screen" button before completing the dictation, when the eye-tracking data were not correctly saved because of various software problems, and when the eye tracker identified the glasses of the radiologist as their eyes. We discarded eye-tracking data for 7 cases for incomplete dictation, 6 for software problems when saving the data, 2 for large parts of the lungs missing, 1 for a horizontally flipped image, 1 for extreme rotation of the MIMIC-CXR image, 41 for low data quality, and 2 for glasses being confused for eyes. The total of discarded eye-tracking data was 10 cases for phase 1, 10 cases for phase 2, and 40 cases for phase 3. Additionally, one CXR in phase 1 had large parts of the lung missing and is not included in the 59 images present in the dataset. The 2,616 unique CXRs for which data are provided in the REFLACX dataset were included after manual image quality exclusions and data quality exclusions.

Data Records

For each reading of a MIMIC-CXR image, the labels of this dataset consist of eye-tracking data, formatted as fixations and as gaze position samples, a timestamped report transcription, ellipses localizing anomalies associated with a certainty, and a chest bounding box. For each case/reading, there is a subfolder containing these labels, separated into individual tables. Subfolders from all three phases are grouped in the same folder (main_data/ and gaze_data/), and the phase to which each subfolder belongs is listed in metadata tables, one for each data collection phase. The statistics of the resulting dataset are presented in Table 2. Genotypical sex information was extracted from the MIMIC-IV dataset[22,30], and was missing for around 0.35% of the 2,199 unique subjects. The dataset is available on Physionet, at https://doi.org/10.13026/e0dj-8498[22,31].

Description of tables and their columns.

- **main_data.zip/main_data/metadata_phase_<phase>.csv**: list of all the subfolders/cases corresponding to a specific phase and their metadata.

    - **id**: subfolder name and unique identifier for a reading of a specific CXR by a specific radiologist.
    - **split**: the split given by the MIMIC-CXR dataset for that specific image. The possible values are "train", "validate", "test". Images were sampled so that 20% of the images were from the test set of the MIMIC-CXR dataset.
    - **eye_tracking_data_discarded**: for phases 1 and 2, even when the eye-tracking data were discarded for low quality, the anomaly labels, localizing ellipses, and chest bounding box were collected and included in the dataset. This column is "True" when the eye-tracking data has been discarded, and "False" otherwise. Transcriptions are also not included for these cases. Such cases should not be used for analysis if eye-tracking data or transcription are required. For phase 3, no case with discarded eye-tracking data is included.
    - **image**: path to the DICOM file from the MIMIC-CXR dataset used in this reading, with the same folder structure as provided by that dataset.
    - **dicom_id**: unique identifier of the image that can be used to join tables with the metadata from MIMIC-CXR.
    - **subject_id**: unique identifier for the patient of that study.
    - **image_size_x**: horizontal size of the CXR in pixels
    - **image_size_y**: vertical size of the CXR in pixels
    - **Other columns**: the rest of the columns represent the possible presence of anomaly evidence in the image, as selected by the radiologist. Most of these columns contain values between 0 and 5, representing a certainty of the presence of such anomaly, according to the scale:
        - 0: not selected by radiologist,
        - 1: Unlikely,
        - 2: Less Likely,
        - 3: Possibly,
        - 4: Suspicious for/Probably,
        - 5: Consistent with.

Certainties were associated with each localizing ellipse in the image, so each label's maximum certainty is reported. Radiologists were asked not to draw ellipses for the anomaly labels "Support devices," "Quality issue" and "Other," so there is no certainty associated with these labels.

- **Support devices** and **Quality issue**: the presence of these labels is reported, using "True" or "False."
- **Other**: A list of the other anomalies reported by radiologists not included in the rest of the labels, separated by "|." If empty, no other anomaly was reported.
- **main_data.zip/main_data/<id>/fixations.csv**: eye-tracking data summarized as fixations and collected during the dictation of the report.
- **timestamp_start_fixation / timestamp_end_fixation**: the time in seconds when the fixation started/ended, counting from the start of the case.
- **average_x_position/average_y_position**: average position for the fixation, given in pixels and in the image coordinate space, where (0,0) is the top left corner.
- **pupil_area_normalized**: pupil area, normalized by the calibration performed at the beginning of each session.
- **window_level/window_width**: average values of the windowing used for the image during a fixation.
- **angular_resolution_x_pixels_per_degree/angular_resolution_y_pixels_per_degree**: number representing how many image pixels fit in 1° of visual angle for each axis (x or y). It is dependent on the position of the fixation and the zoom applied to the image.
- **xmin_shown_from_image/ymin_shown_from_image/xmax_shown_from_image/ymax_shown_from_image:** bounding box given in image-space representing what part of the image was shown to the radiologist at the start of the fixation. The reading/case always started with the whole image being shown, but zooming and panning were allowed.
- **xmin_in_screen_coordinates/ymin_in_screen_coordinates/xmax_in_screen_coordinates/ymax_in_screen_coordinates:** bounding box given in screen space representing where the part of the image was shown.

- **gaze_data.zip/gaze_data/<id>/gaze.csv**: complete eye-tracking data during the dictation of the report, collected at 1,000 Hz. Even though this data are not necessary for accomplishing the main research goals of this dataset, these data are included for any other analyses that need the gaze location in more detail than provided by the fixations.csv table. Compared to the fixations.csv table, the **timestamp_start_fixation** and **timestamp_end_fixation** columns were replaced by the **timestamp_sample** column. The remaining columns are the same in both tables, but they represent values when the eye tracker's camera captured the gaze sample in gaze.csv.

    - **timestamp_sample**: timestamps do not start at 0 because audio recording started before gaze recording.

  - **x_position/y_position/pupil_area_normalized/angular_resolution_x_pixels_per_degree/angular_resolution_y_pixels_per_degree:** these columns are empty for timestamps when the eye tracker could not find the radiologist's pupil or cornea, making it impossible to calculate gaze at that moment. These rows are usually associated with moments when radiologists blinked.

- **main_data.zip/main_data/<id>/timestamps_transcription.csv**: timestamped corrected transcriptions of the reports. Radiologists were allowed to delete parts of the report and to modify transcription errors but not to add content.

  - **word**: the word that was spoken. Periods (.), commas (,) and slashes (/) occupy one row.
  - **timestamp_start_word/timestamp_end_word**: the time in seconds when the dictation of each word started/ended, counting from the start of the case.

- **main_data.zip/main_data/<id>/transcription.txt**: the transcription in text form, without timestamps.
- **main_data.zip/main_data/<id>/anomaly_location_ellipses.csv**: bounding ellipses drawn by radiologists for each label present in the image. Each label may appear in more than one ellipse, and each ellipse may contain more than one label. Radiologists were instructed to select more than one label for an ellipse when a single image finding may be evidence of one or another label. Each row of the table represents one ellipse.

  - **xmin/ymin/xmax/ymax**: coordinates representing the extreme points, in image pixels, of the full horizontal and vertical axes of the ellipse. Coordinate (0,0) represents the top left corner of the image.
  - **certainty**: value from 0 to 5 representing a certainty of the finding presence, according to the same scale as in the metadata table.
  - **Other columns**: the rest of the columns have a Boolean value representing the presence of evidence for each anomaly label in the ellipse. Since radiologists were asked not to draw ellipses for "Support devices", "Quality issue", and "Other," most of the rows have the value "False" for these labels. For all other labels present in the image, at least one ellipse is drawn.

- **main_data.zip/main_data/<id>/chest_bounding_box.csv**: single-row table containing information for the bounding box drawn around the lungs and the heart.

  - **xmin/ymin/xmax/ymax**: coordinates representing the extreme points, in image pixels, of the bounding box. Coordinate (0,0) represents the top left corner of the image.

Technical Validation

For all reported technical validation values, we also report standard error and sample size when applicable. Standard errors were only reported for $n > 10$. For measurements that had more than one score for the same CXR, e.g., intersection over union (IoU) calculated for all pairs of radiologists, we averaged the scores for each CXR before calculating the final average. The sample size given is the number of independent CXRs involved in the calculation.

Eye-tracking data.

Considering the errors provided by each of the calibrations used in data collection, we calculated the average and maximum calibration errors for each phase. Phase 1 had an average calibration error of $0.43 \pm 0.02°$ ($n = 25$) and a maximum error of 2.79°, whereas, for phase 2, it was $0.43 \pm 0.03°$ ($n = 13$) and 1.09°, respectively. Phase 3 had an average error of $0.44 \pm 0.01°$ ($n = 128$) and a maximum error of 1.5°.

To validate the presence of abnormality location information in the eye-tracking data, we calculated the presence of fixations inside the abnormality ellipses, as exemplified in Fig. 3. For each reading that had at least one drawn ellipse, we calculated the normalized cross-correlation (NCC) between a fixation heatmap and a mask generated from the union of ellipses, using

$$\mathrm{NCC}(x_f, x_e) = \frac{1}{P-1}\sum_{p}\frac{(x_f(p) - \mu_{x_f})}{\sigma_{x_f}} \times \frac{(x_e(p) - \mu_{x_e})}{\sigma_{x_e}}, \tag{1}$$

where $\mathrm{NCC}(x_f, x_e)$ is the NCC between the fixation heatmap x_f and the ellipse heatmap x_e, σ_x is the standard deviation of a heatmap x, μ_x is the average value of a heatmap x, $x(p)$ is the value of a heatmap x at pixel p, and P is the number of pixels in the heatmaps. The fixation heatmaps were generated by drawing Gaussians centered on every fixation, with the standard deviation equal to 1° in each axis and with intensity proportional to the fixation duration.

To check if the fixations heatmap of a specific CXR has more localization information than the heatmaps from unrelated CXRs, we compared the NCC for two types of fixations heatmap: heatmaps generated from the eye-tracking fixations of each CXR reading, as shown in Fig. 3c, and a baseline heatmap representing the average gaze over all CXRs, as shown in Fig. 3d.

To calculate the baseline heatmap shown in Fig. 3d, we normalized all heatmaps to the same location using the labeled chest bounding boxes to compensate for the variation in the location of the lungs for each CXR. We calculated the average chest bounding box, transformed all heatmaps to this same space, and averaged the heatmaps. We finally transformed the average heatmap back to the space of each CXR to calculate the NCC against the labeled ellipses.

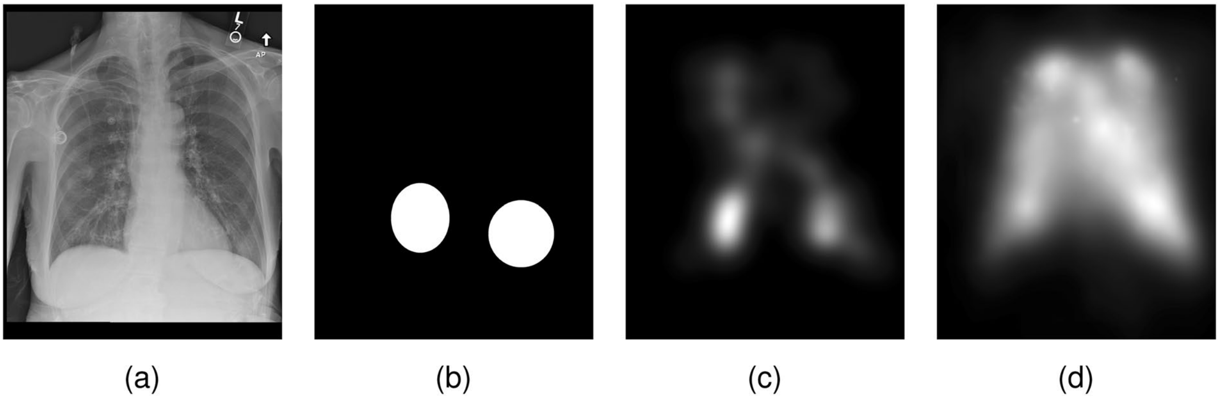


Fig. 3 Example of the localization information provided by the eye-tracking data and how it was validated. (**a**) CXR read by the radiologist. (**b**) Union of the abnormality ellipses selected by radiologists used to compare against heatmaps. (**c**) Heatmap generated by the fixations made by the radiologist while dictating the report. (**d**) Average heatmap for all radiologists and CXRs read in phases 1 and 2, normalized to the location of lung and heart of the CXR.

For the fixation heatmaps specific to each CXR, the average NCC achieved over all applicable CXRs was 0.380 ± 0.014 (n = 96), against a baseline score of 0.326 ± 0.013 (n = 96). This result shows that there is more abnormality location information in the fixations for each CXR than on a heatmap built from the usual areas looked at by a radiologist.

To analyze if the localization information given by the eye-tracking data correlates with the time that the presence of an abnormality was mentioned during the dictation, we produced the graph shown in Fig. 4. To develop this analysis, we annotated the location where labels were mentioned in the report for 200 non-test CXRs from Phase 3. For annotating, we used a mix of a modified version of the chexpert labeler[4] and hand-labeled corrections. CXRs that had image-level labels not mentioned in the dictation were not included in the 200 randomly selected CXRs.

We separated the time before the end of mentions of a label into bins of same size. For each bin, we calculated the percentage of fixations localized inside an ellipse for the same abnormality and CXR. The percentage was calculated considering the duration of the fixation inside the bin. Besides analyzing the delay in time units, we also used sentences units. To calculate the sentence units of each timestamp, we separated the dictation into mid-sentence moments and in-between-sentences pause moments. The timestamps of each transition between these were represented by integer numbers. Timestamps in the middle of a sentence or pause had their representation calculated through linear interpolation of the start and end of their sentence or pause. For example, a 12 s timestamp within the second sentence, dictated from 10 s to 15 s, would be associated with 3.4 sentence units. The sentence units shown in Fig. 4 represent the difference between the sentence units of the mentions and the fixations before the mentions.

We divided the full range of data into 75 bins, of width 1.03 s and 0.21 sentences, and only kept bins up to before the first bin with less than or equal to ten fixations inside ellipses. We calculated 95% confidence intervals through bootstrapping, randomly sampling with replacement 200 CXRs from the 200 annotated CXRs. We performed this sampling 800 times and display in Fig. 4 the 2.5% and 97.5% percentiles for each bin.

As shown in Fig. 4, there are peaks of correlation between the location of ellipses and the radiologists' fixations at around 2.5 s before the mention of the respective abnormality. The correlation peak was also calculated to be around 0.6 to 1.25 sentence units before the mention. This correlation shows that the transcription timestamps could be used to get label-specific localization information. With the shown delay between label fixation and mention, our data might need alignment algorithms for the correct association between fixations and label. We leave the exploration of the application of such algorithms to future works.

Validation labels.

Image-level labels. The inter-rater reliability was measured through Fleiss' Kappa[32] for the image-level labels, calculated using the statsmodels library in Python[33]. Image-level labels were considered positive when the maximum certainty, among all ellipses of a given label, was "Possibly" or higher. The inter-rater reliability scores for phases 1 and 2 are shown in Table 1. The achieved inter-rater reliability scores are relatively low but in line with scores obtained in other similar studies for readings of CXR[34,35]. Some of the low values might be caused by the low prevalence of some of the labels[36,37].

Localization ellipses. For each CXR, for each label with more than one radiologist selecting certainty "Possibly" or higher, we calculated the average paired IoU between all pairs of respective radiologists. We then calculated the average IoU over all CXRs of each phase, with results presented in Table 1. Between phases 1 and 2, we discussed a few examples of CXR that had low IoU.

Chest bounding boxes. Similar to the localization ellipses, the quality of the bounding boxes containing the heart and the lungs was measured through IoU. IoU was calculated between all the pairs of radiologists for every

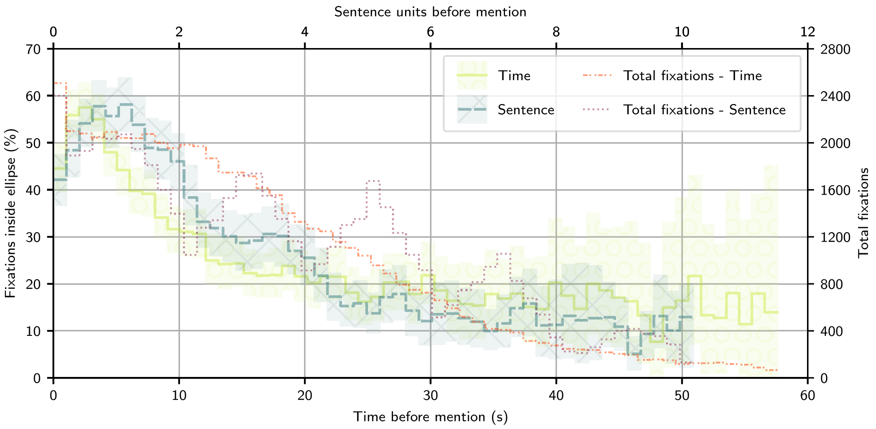


Fig. 4 Time analysis of the correlation between each mention of a label and what percentage of fixations were located inside the ellipses that localized each respective label. We present two lines, one as a function of time and another as a function of the counting of sentences and pauses before the mention. The step lines represent the percentage for separate data bins. We also draw the 95% confidence interval for each bin in each line, calculated with bootstrapping. The number of fixations used to calculate each bin is shown in separate lines.

CXR of the preliminary phases. For phase 1, the average IoU was 0.917 ± 0.004 (n = 59), and for phase 2, it was 0.920 ± 0.004 (n = 50). We organized no discussion for improvement of this label between phases 1 and 2.

Usage Notes

To have access to the CXRs that the radiologists read, access to the MIMIC-CXR dataset[6,21–23] is necessary. Both datasets are accessible only on Physionet, requiring the signature of a data use agreement by a logged-in user. Access to the MIMIC-CXR dataset requires free online courses in HIPAA regulations and human research subject protection.

The main uses intended for our data include:

- combining fixations into heatmaps, for use as an attention label and research on saliency maps and related subjects;
- using the fixations as a nonuniformly sampled sequence of attention locations;
- combining the timestamped transcriptions with the fixations for more specific localization to each abnormality;
- associating the pupil data with the fixations for more information on the cognitive load of each fixation;
- validating abnormalities parsed from the transcriptions using the image-level labels;
- validating the locations found from the eye-tracking data through the abnormality ellipses;
- using the chest bounding boxes for normalizing the location of the lungs while performing other analyses; and
- using the chest bounding boxes for training a model to output bounding boxes for unseen data.

Other possible uses of the data may include using the certainties provided by the radiologists in uncertainty quantification research and the reports and their transcriptions in image captioning for chest x-rays. In https://github.com/ricbl/eyetracking, we provide examples on how to generate heatmaps, how to normalize the location of heatmaps using the chest bounding box, how to filter the fixations, how to calculate brightness of the shown CXR in any given time during dictation, and how to load and use the tables from the dataset.

Eye-tracking data. There are uncertainties in the eye-tracking measurement pipeline. To represent them, we suggest following a method used in the generation of heatmaps in the visual attention modeling literature and modeling the location of each fixation as a Gaussian with a standard deviation of 1° of visual angle[38]. We provide pixel resolution per visual angle for each axis of the image, so the Gaussian will be slightly anisotropic in the image space. For some applications, e.g., when generating one embedding vector per fixation for sequence analysis, it might be beneficial to filter out fixations that happened outside of the image. Among other reasons, fixations may have happened outside of the image because there were two buttons in the dictation screen: one to indicate that the dictation was over, and one to reset the windowing, zooming, and panning of the image.

Abnormality labels. Figure 5 shows the hierarchy between the labels of our study and the labels from the MIMIC-CXR dataset. Not all labels present in one dataset have an equivalent in other datasets. This hierarchy was produced to the best of their understanding of the MIMIC-CXR labels. Supplementary File 1 provides the definition of the labels agreed upon among the radiologists who participated in the study.

Pupil data. Brunyé *et al.*[39] showed that the pupil diameter of pathologists reliably increased for more difficult cases, providing an indicator of cognitive engagement. In our dataset, we provide the normalized pupil area, whose square root is equivalent to the normalized pupil diameter. The normalization was performed by a

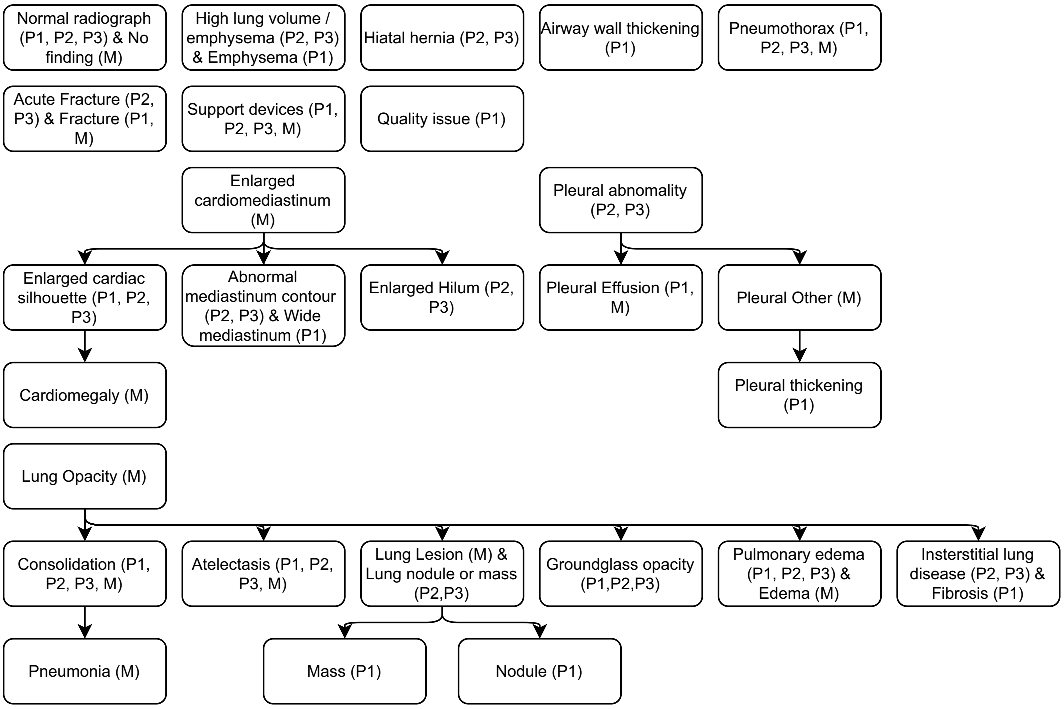

Normal radiograph (P1, P2, P3) & No finding (M)
High lung volume / emphysema (P2, P3) & Emphysema (P1)
Hiatal hernia (P2, P3)
Airway wall thickening (P1)
Pneumothorax (P1, P2, P3, M)
Acute Fracture (P2, P3) & Fracture (P1, M)
Support devices (P1, P2, P3, M)
Quality issue (P1)
Enlarged cardiomediastinum (M)
Pleural abnomality (P2, P3)
Enlarged cardiac silhouette (P1, P2, P3)
Abnormal mediastinum contour (P2, P3) & Wide mediastinum (P1)
Enlarged Hilum (P2, P3)
Pleural Effusion (P1, M)
Pleural Other (M)
Cardiomegaly (M)
Pleural thickening (P1)
Lung Opacity (M)
Consolidation (P1, P2, P3, M)
Atelectasis (P1, P2, P3, M)
Lung Lesion (M) & Lung nodule or mass (P2,P3)
Groundglass opacity (P1,P2,P3)
Pulmonary edema (P1, P2, P3) & Edema (M)
Insterstitial lung disease (P2, P3) & Fibrosis (P1)
Pneumonia (M)
Mass (P1)
Nodule (P1)

Fig. 5 Hierarchy of the labels of all the phases of our dataset and the labels of the MIMIC-CXR dataset. Arrows point to a subset of the originating label. The datasets to which each label belongs are listed inside parentheses, according to P1 (Phase 1), P2 (Phase 2), P3 (Phase 3), and M (MIMIC-CXR). Labels that do not have a hierarchical relationship with other labels are not connected to any arrows.

division by the area of the pupil in a standardized screen. However, the variation of the screen brightness, caused by windowing, zooming, and panning, may cause more variation in pupil area than psychological reasons. We suggest including another normalization, using the division by a value representing the screen brightness at each moment, similarly to what was done by Brunyé *et al.*[39]. For the calculation of this value, we suggest summing the intensity of pixels of the shown CXR. The part of the image shown in each moment and the part of the screen where it was shown are provided in the dataset. When considering the windowing of the CXR, images are shown following

$$shown\_image = min\left(max\left(\frac{original\_image - window\_level}{window\_width} + 0.5,\ 0\right),\ 1\right), \tag{2}$$

where *window_width* can have values from 1.5e-05 to 2 and is usually initialized to 1, *window_level* can have values from 0 to 1 and is usually initialized to 0.5, *shown_image* is the image sent to the screen, and *original_image* is the loaded DICOM image normalized so that its possible range is from 0 to 1, which usually means dividing the image intensities by 4,096. The initial *window_width*, *window_level*, and the maximum intensity value were loaded from the DICOM tags of each image file.

Limitations of the study.

- Limited information presented to radiologists: our setup used a single screen, but multiple screens are used in a clinical setting. With multiple screens, the eye-tracker setup is more complex and would have to be validated. We limited the study to show only frontal CXR. Lateral views, past CXRs, and clinical information were not presented.
- Report: in clinical practice, reports can be modified after a first transcription. We limited the editing to corrections of the transcription of the original dictation and deletions of dictation mistakes. This limitation was needed to assign timestamps to each word and ensure the radiologist saw the finding while the eye tracker was on. Several radiologists use templates for their reports in clinical practice, only dictating small parts of the report. We did not test such a dictation method.
- Head position: even though we used the remote mode for the eye tracker, which allows for some freedom of movement, the head movement, posture, and the distance from head to the screen were still more limited than in clinical practice, when radiologists can get closer to the screen to see a detail, for example. Because of this limitation, the use of zooming was probably more frequent than in clinical practice. Furthermore, radiologists mentioned that, with the limitations in position, they became fatigued faster than usual.

- CXR dataset: we collected readings for images of only one dataset, so the current dataset may have ensuing biases. The radiologists also characterized images from the MIMIC-CXR dataset as having lower quality than usual for their practice, in aspects like the field-of-view excluding small parts of the lung and the blurring present in some images.
- Display: the GPU of the computer used to display the CXRs supported only 8-bit display, so not all intensities of the original DICOM were shown, reducing the image quality and possibly changing the way radiologists interacted with the CXR. This limitation was partially remediated by allowing the windowing of the image to be changed.
- Calibration cost: calibrations happened every 45 to 60 minutes, and sometimes more than five retries were needed to reach quality thresholds. The clinical implementation of the data collection method described in this paper for the collection of larges quantities of data might cause an undesired cost to the radiologist reading process.
- Unautomated processes: another person was in the room coordinating calibrations and checking for low data quality. This same person raised the need for a recalibration or a change in position of the radiologist. This person might have to be replaced by automated processes if this data collection method is implemented in clinical practice.

Code availability

The code used for all automatic processes described in this paper, involving sampling, collection, processing, and validation of data, is available at https://github.com/ricbl/eyetracking[28] The software and versions we used were: MATLAB R2019a, Psychtoolbox 3.0.17[25–27], Python 3.7.7, edfapi 3.1, EYELINK II CL v5.15, Eyelink GL Version 1.2 Sensor = AC7, EDF2ASC 3.1, librosa 0.8.0[40], numpy 1.19.1[41], pandas 1.1.1[42,43], matplotlib 3.5.1[44,45], statsmodels 0.12.2[33], shapely 1.7.1[46], scikit-image 0.17.2[47], pyrubberband 0.3.0, pydicom 2.1.2[48], pydub 0.24.1, soundfile 0.10.3.post1, pyttsx3 2.90, pillow 8.0.1[49], scikit-learn 0.23.2[50], nltk 3.5[51], syllables 1.0.0, moviepy 1.0.3[52], opencv 3.4.2[53], Ubuntu 18.04.5 LTS, espeak 1.48.04, joblib 1.1.0, ffmpeg 3.4.8, and rubberband-cli 1.8.1.

Received: 8 November 2021; Accepted: 1 June 2022;
Published online: 18 June 2022

References

1. Rajpurkar, P. *et al*. CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. Preprint at http://arxiv.org/abs/1711.05225 (2017).
2. Lakhani, P. & Sundaram, B. Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* **284**, 574–582, https://doi.org/10.1148/radiol.2017162326 (2017).
3. Bustos, A., Pertusa, A., Salinas, J. M. & de la Iglesia-Vayá, M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Anal.* **66**, 101797, https://doi.org/10.1016/j.media.2020.101797 (2020).
4. Irvin, J. *et al*. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 590–597, https://doi.org/10.1609/aaai.v33i01.3301590 (AAAI Press, 2019).
5. Wang, X. *et al*. ChestX-Ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017*, 3462–3471, https://doi.org/10.1109/CVPR.2017.369 (IEEE Computer Society, 2017).
6. Johnson, A. E. W. *et al*. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data* **6**, 317, https://doi.org/10.1038/s41597-019-0322-0 (2019).
7. Deng, J. *et al*. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255, https://doi.org/10.1109/CVPR.2009.5206848 (IEEE, 2009).
8. Wu, B. *et al*. Tencent ML-Images: A large-scale multi-label image database for visual representation learning. *IEEE Access* **7**, https://doi.org/10.1109/ACCESS.2019.2956775 (2019).
9. Li, Z. *et al*. Thoracic disease identification and localization with limited supervision. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018*, 8290–8299, https://doi.org/10.1109/CVPR.2018.00865 (IEEE Computer Society, 2018).
10. Shih, G. *et al*. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence* **1**, e180041, https://doi.org/10.1148/ryai.2019180041 (2019).
11. Nguyen, H. Q. *et al*. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. Preprint at https://arxiv.org/abs/2012.15029 (2021).
12. Langlotz, C. P. *et al*. A roadmap for foundational research on artificial intelligence in medical imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop. *Radiology* **291**, 781–791, https://doi.org/10.1148/radiol.2019190613 (2019).
13. Templier, T., Bektas, K. & Hahnloser, R. H. R. Eye-trace: Segmentation of volumetric microscopy images with eyegaze. In Kaye, J., Druin, A., Lampe, C., Morris, D. & Hourcade, J. P. (eds.) *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7–12, 2016*, 5812–5823, https://doi.org/10.1145/2858036.2858578 (ACM, 2016).
14. Stember, J. N. *et al*. Eye tracking for deep learning segmentation using convolutional neural networks. *J. Digit. Imaging* **32**, 597–604, https://doi.org/10.1007/s10278-019-00220-4 (2019).
15. Khosravan, N. *et al*. A collaborative computer aided diagnosis (C-CAD) system with eye-tracking, sparse attentional model, and deep learning. *Medical Image Anal.* **51**, 101–115, https://doi.org/10.1016/j.media.2018.10.010 (2019).
16. Gecer, B. *et al*. Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks. *Pattern Recognit.* **84**, 345–356, https://doi.org/10.1016/j.patcog.2018.07.022 (2018).
17. Stember, J. N. *et al*. Integrating eye tracking and speech recognition accurately annotates mr brain images for deep learning: Proof of principle. *Radiology: Artificial Intelligence* **3**, e200047, https://doi.org/10.1148/ryai.2020200047 (2021).
18. Karargyris, A. *et al*. Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for AI development. *Scientific Data* **8**, https://doi.org/10.1038/s41597-021-00863-5 (2021).
19. Saab, K. *et al*. Observational supervision for medical image classification using gaze data. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021*, 603–614, https://doi.org/10.1007/978-3-030-87196-3_56 (Springer International Publishing, 2021).

20. Hansell, D. M. *et al*. Fleischner society: Glossary of terms for thoracic imaging. *Radiology* **246**, 697–722, https://doi.org/10.1148/radiol.2462070712 (2008).
21. Johnson, A. E. W., Pollard, T., Mark, R., Berkowitz, S. & Horng, S. The MIMIC-CXR database (version 2.0.0). *PhysioNet* https://doi.org/10.13026/C2JT1Q (2019).
22. Goldberger, A. L. *et al*. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **101**, e215–e220, https://doi.org/10.1161/01.CIR.101.23.e215 (2000).
23. Johnson, A. *et al*. MIMIC-CXR-JPG - chest radiographs with structured labels (version 2.0.0). *PhysioNet* https://doi.org/10.13026/8360-t248 (2019).
24. Johnson, A. E. W. *et al*. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. Preprint at https://arxiv.org/abs/1901.07042 (2019).
25. Brainard, D. H. The Psychophysics Toolbox. *Spatial Vision* **10**, 433–436, https://doi.org/10.1163/156856897X00357 (1997).
26. Pelli, D. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial vision* **10**, 437–442, https://doi.org/10.1163/156856897x00366 (1997).
27. Kleiner, M. *et al*. What's new in psychtoolbox-3. *Perception* **36**, 1–16, https://doi.org/10.1177/03010066070360S101 (2007).
28. Bigolin Lanfredi, R. ricbl/eyetracking: Code for REFLACX dataset v1.2, https://doi.org/10.5281/zenodo.6419833 (2022).
29. Panicek, D. M. & Hricak, H. How sure are you, doctor? a standardized lexicon to describe the radiologist's level of certainty. *AJR. American journal of roentgenology* **207**, 2–3, https://doi.org/10.2214/ajr.15.15895 (2016).
30. Johnson, A. *et al*. MIMIC-IV (version 1.0). *PhysioNet* https://doi.org/10.13026/S6N6-XD98 (2021).
31. Bigolin Lanfredi, R. *et al*. REFLACX: Reports and eye-tracking data for localization of abnormalities in chest x-rays. *PhysioNet* https://doi.org/10.13026/E0DJ-8498 (2021).
32. Fleiss, J. Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76**, 378–382, https://doi.org/10.1037/h0031619 (1971).
33. Seabold, S. & Perktold, J. statsmodels: Econometric and statistical modeling with python. In Walt, S. v. d. & Millman, J. (eds.) *Proceedings of the 9th Python in Science Conference*, 92–96, https://doi.org/10.25080/Majora-92bf1922-011 (2010).
34. Balabanova, Y. *et al*. Variability in interpretation of chest radiographs among russian clinicians and implications for screening programmes: observational study. *BMJ* **331**, 379–382, https://doi.org/10.1136/bmj.331.7513.379 (2005).
35. Quekel, L. G., Kessels, A. G., Goei, R. & van Engelshoven, J. M. Detection of lung cancer on the chest radiograph: a study on observer performance. *European Journal of Radiology* **39**, 111–116, https://doi.org/10.1016/S0720-048X(01)00301-1 (2001).
36. Wongpakaran, N., Wongpakaran, T., Wedding, D. & Gwet, K. L. A comparison of cohen's kappa and gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology* **13**, https://doi.org/10.1186/1471-2288-13-61 (2013).
37. Sim, J. & Wright, C. C. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy* **85**, 257–268, https://doi.org/10.1093/ptj/85.3.257 (2005).
38. Le Meur, O. & Baccino, T. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods* 1–16, https://doi.org/10.3758/s13428-012-0226-9 (2012).
39. Brunyé, T. T. *et al*. Pupil diameter changes reflect difficulty and diagnostic accuracy during medical image interpretation. *BMC Medical Informatics and Decision Making* **16**, https://doi.org/10.1186/s12911-016-0322-3 (2016).
40. McFee, B. *et al*. librosa/librosa: 0.8.0, https://doi.org/10.5281/zenodo.3955228 (2020).
41. Harris, C. R. *et al*. Array programming with NumPy. *Nature* **585**, 357–362, https://doi.org/10.1038/s41586-020-2649-2 (2020).
42. McKinney, W. Data Structures for Statistical Computing in Python. In S., van der Walt & J. Millman (eds.) *Proceedings of the 9th Python in Science Conference*, 56–61, https://doi.org/10.25080/Majora-92bf1922-00a (2010).
43. Reback, J. *et al*. pandas-dev/pandas: Pandas 1.1.1, https://doi.org/10.5281/zenodo.3993412 (2020).
44. Hunter, J. D. Matplotlib: A 2d graphics environment. *Computing in Science Engineering* **9**, 90–95, https://doi.org/10.1109/MCSE.2007.55 (2007).
45. Caswell, T. A. *et al*. matplotlib/matplotlib: Rel: v3.5.1, https://doi.org/10.5281/zenodo.5773480 (2021).
46. Gillies, S. *et al*. Shapely: manipulation and analysis of geometric objects. *GitHub* https://github.com/Toblerity/Shapely (2007).
47. van der Walt, S. *et al*. scikit-image: image processing in Python. *PeerJ* **2**, e453, https://doi.org/10.7717/peerj.453 (2014).
48. Mason, D. *et al*. pydicom/pydicom: pydicom 2.1.2, https://doi.org/10.5281/zenodo.4313150 (2020).
49. van Kemenade, H. *et al*. python-pillow/pillow 8.0.1, https://doi.org/10.5281/zenodo.4118627 (2020).
50. Pedregosa, F. *et al*. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
51. Bird, S., Klein, E. & Loper, E. *Natural language processing with Python: analyzing text with the natural language toolkit* ("O'Reilly Media, Inc.", 2009).
52. Zulko *et al*. johncooper199/moviepy. *Zenodo* https://doi.org/10.5281/zenodo.4781125 (2021).
53. Bradski, G. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

Acknowledgements

This research was funded by the National Institute Of Biomedical Imaging And Bioengineering of the National Institutes of Health under Award Number R21EB028367. Christine Prickett provided copyediting support, Lauren Williams provided help in using Psychtoolbox 3, David Alonso provided advice in running eye-tracking sessions, Howard Mann was one of the five radiologists to read images, and Yichu Zhou was part of the team that modified the chexpert labeler to fit the labels of our dataset.

Author contributions

R.B.L. wrote the manuscript, did all the coding specific for this project, conducted the data-collection sessions, and ran the analyses. M.Z. participated in the design of the technical validation analysis. W.A. participated in the study design, was one of the readers and provided feedback for data-collection processes. J.C. was one of the readers and provided feedback for data-collection processes. P.A.D. was one of the readers and provided feedback for data-collection processes. V.S. participated in the study design. T.D. participated in the study design, conceived the fundamental parts of the eye-tracking side of the data-collection sessions, and wrote the paragraph comparing types of eye-tracking devices. J.S. participated in the study design, guided the clinical decisions for the data collection, produced the clinical instructions, was one of the readers, and provided feedback for data-collection processes. T.T. is the PI of the project, coordinating the study design, leading discussions about the project, and editing the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-022-01441-z.

Correspondence and requests for materials should be addressed to R.B.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.