Supplementary Methods

CXR dataset construction of common thoracic diseases

We constructed CXR datasets for the development and evaluation of an AI model for common thoracic diseases. We used the NLP pipeline to extract labels from reports for each CXR image. Two datasets were constructed. The SYSU dataset is composed of 120,702 CXR images from 92,327 patients between October 2018 and July 2020 in both inpatient and outpatient centers. The SYSU-PE dataset is composed of 24,500 CXR images from 23,585 patients for the health check.

We selected fourteen common thoracic diseases according to their clinical significance and prevalence, as defined in the NIH Chest X-ray dataset. These disease labels included atelectasis, cardiomegaly, consolidation, edema, effusion, emphysema, fibrosis, hernia, infiltration, nodule, mass, pleural thickening, pneumonia, and pneumothorax. We also defined another label of “No finding” that is positive if all other labels of the CXR image are negative. Thus, each CXR image in the dataset was annotated by the presence and absence of the fifteen labels.

We developed a rule-based NLP pipeline to extract disease labels from CXRs’ associated radiology reports. The pipeline included disease concept detection and negation classification, similar to CheXpert and NIH Chest X-ray dataset. We manually annotated 2,000 reports for model development and testing. A list of disease phrases was manually annotated by multiple board-certified radiologists. The radiologists read the 1,000 developmental reports to further improve the phrase list. Based on this list, we used regular expressions to detect disease concepts for each report. We evaluated the performance of this extraction module by the testing reports, achieving an average F-score of 0.99. Next, we developed the model using several hand-crafted rules for negation classification of the developmental reports, and evaluated them on the testing dataset. The model achieved an average F-score of 0.99. Finally, we used this NLP pipeline for label extraction.

Overview of the AI system

Our proposed AI system applied a modular pipeline approach, which consisted of three main components: a CXR standardization module, a common thoracic disease classification module, and a final pneumonia analysis module.

Automated CXR standardization module

The standardization module for CXR image processing consisted of an invert-grayscale CXR detection model, an anatomical landmarks detection model, and image registration techniques (Fig. 1 and Extended Data Fig. 7). The pipeline was designed to overcome the notorious problem and well-known challenges of data diversity/variations and non-standardization of CXR images.
**Invert-grayscale CXR detection model**

We trained an invert-grayscale CXR detection model to detect whether the input of CXR was inverted-grayscale and further automatically converted it into a conventional CXR image. We sampled 10,000 conventional CXR from the CheXpert dataset and 10,000 negative samples from CC-CXR and the inverted ones as positive samples. Binary Cross Entropy (BCE) loss was used for this task. Our invert-grayscale CXR detection model achieved an AUC>0.999 on the test set.

**Anatomical landmarks detection model**

We formulated the anatomical landmark detection as a heatmap regression, an approach widely adopted in recent work implemented by fully convolutional neural networks. For each landmark, we first generated a heatmap by a fully convolutional neural network, where the heatmap predicts the probability of the landmark occurring at each position. Namely, each output channel was predicted for one landmark. And then, rather than directly reporting the position with the maximum probability, we calculated the position by expectation over the heatmap. We adopted three fully convolutional architectures for this task, i.e., U-Net, Fully Convolutional Network, and DeepLabv3.

The main novelty of the DeepLabv3 model is to adopt an improved version of the atrous spatial pyramid pooling (ASPP) structure with a global average pooling layer (GAP), which helped to capture multi-scale information. The number of output channels was set to $K$ for the detection of $K$ landmarks ($K = 12$ for 12 anatomical landmarks in this study). We then apply softmax to the heatmap to give probability predictions for each anatomical landmark. To estimate a landmark through the heatmap, we applied the differentiable spatial to numerical transform (DSNT), which considered the heatmap as a probability distribution and took the expectation of location over the distribution as the estimated landmark. The loss function was based on the estimated landmark and the ground truth landmark, and the network was trained end-to-end. During training, the Adam optimizer was employed with a learning rate of 0.001. The training batch size was 8. A validation set was used for early-stopping with a patience of 10 to avoid overfitting. The model with the best validation loss was finally selected.

**CXR image registration**

After training the anatomical landmarks detection network, we used the results for CXR image registration via optimization. Specifically, the algorithm aims to solve an affine transformation $R \in \mathbb{R}^{2 \times 2}, t \in \mathbb{R}^2$ s.t. $R^TR = I$ that minimize the loss function $\sum_{i=1}^{n} \|Y_i - (RX_i + t)\|^2$, where $X_i, Y_i$ denote the i-th origin landmark coordinate and target landmark coordinate (which is fixed in this task), respectively. Then, we applied the solved transformation $(R, t)$ to all the pixels for CXR image registration.

**Common thoracic disease classification module**

By utilizing the CXR standardization module, this registered CXR image was used for the final input to the model for common thoracic disease classification and made an output prediction.
Pneumonia analysis module

CXR images labeled as pneumonia were further analyzed by this module. The pneumonia analysis module is a two-stage architecture based on the previous standardization module. In the first stage, a detector identifies suspicious lung-lesion regions in CXR images. Next, the generated lung maps, lesion maps, and registered original CXR are fed to a classifier to produce probability outputs. We used the pneumonia analysis module to improve pneumonia diagnostic performance and for COVID-19 severity assessment.

Lung-lesion segmentation module

Based on the landmark detection and registration, the lung and the lesion segmentation on CXR were performed using the aligned CXR images. Both the lung segmentation model and the lesion detection models were developed based on DeepLabv3, U-Net, and Fully Convolutional Network. We finally chose the DeepLabv3 because of its superior performance in this task. We modified the number of final output channels for the different segmentation tasks.

The networks were trained sequentially: the lung segmentation network was trained using the registered CXR images as inputs, and then extracted lung regions were used to train the lesion segmentation network. Since the raw CXR images may consist of irrelevant information for lesion segmentation (e.g. body parts not related to the lungs), a lung segmentation network was employed to discard such information and let the lesion segmentation network concentrate only on the lung area. The number of output channels was set to 5 for pixel-level classification, including background and four classes of lung region. The input was a CXR image resized into \(512 \times 512\) by bilinear interpolation, and the output was a heatmap with the same size. The size of the output heatmap was restored to the original size using up-sampling of bilinear interpolation. We then considered the binary mask generated by considering all background predictions below 0.5 to be the initial lung mask. Only the two largest connected regions of the initial lung mask were used for the final lung mask. A minimized bounding box overlapping the mask was used to clip the original image and to output an image with a segmented lung area. All processed CXR images were segmented for the purpose of guiding AI models to focus on the lung region.

The lesion segmentation network was trained to identify lesions using the output from the lung segmentation network, i.e., the image cropped to the lung bounding box, as input, and output the final probability heatmap. The number of output channels was set to 3 for pixel-level classification, including background and two classes of the lesion (including opacification and interstitial pattern). Similar to the approach of lung segmentation network, the input was resized to \(512 \times 512\) by bilinear interpolation, and the bilinear interpolation method is adopted to upsample the output to the original size. The lung and lesion segmentation networks were respectively trained using CXR images with pixel-level lung and lesion region annotations. For training, we used the pixel-level cross-entropy loss as the loss function, and Adam optimizer was employed with a learning rate at 0.001 and batch size of 8. A five-fold cross-validation test was applied, with each running process including four folds for development (70% of training and 10% of validation) and
one-fold for evaluation. A validation set was used for early-stopping with a patience of 10 to avoid overfitting. The model with the best validation loss was finally selected.

For the lung-lesion segmentation model, 1,016 CXR images including 228 viral pneumonia (including 121 COVID-19), 350 other pneumonia, and 438 normal CXR subjects were employed. We performed a manual segmentation task on 1,016 CXR images and divided a CXR image into four anatomical categories: lung field (left and right), and periphery of the lung field (left and right). In addition, two common lung lesions (opacification and interstitial pattern) were segmented. Two metrics, Dice coefficient and Intersection over Union (IOU), were employed to evaluate lung segmentation and lesion segmentation performance (Extended Data Table 4 and 5). For the two metrics, the DeepLabv3 outperformed FCN and U-Net on all four anatomical segmentation work as well as on the interstitial patterns. DeepLabv3 performed slightly worse than FCN on opacification lesions.

**Lung-lesion map for pneumonia detection enhancement**

The diagnostic model used three outputs from the previous models as input, namely the anatomical landmarks detection model, the lung segmentation model, and the lesion segmentation model. The generated three outputs were used to simulate three channels of RGB as input for image classification networks and provide more information to the classifier than simply using three repeated channels of grayscale CXR images. Given an original CXR image $x_{ori}$, We generated a registered CXR image $x_{reg}$ via the anatomical landmark detection model and the image registration algorithm. Since $x_{reg}$ was a grayscale image, we could see it as a brightness map, which could reflect information of lesion density. The segmented lung map $x_{lung}$ of the registered CXR image was generated by the lung segmentation network, where $x_{lung}$ was a foreground probability heatmap (without category information). And the segmented lesion map $x_{lesion}$ (without category information) was generated by the lesion segmentation network. These three $x_{reg}$, $x_{lung}$, and $x_{lesion}$ are stacked and used as channels for the final input to the DenseNet-121 for CXR image classification and made an output prediction. We conducted experiments to study the multi-channel model in comparison with the single-channel model. The results showed the superior performance of using the three maps as multi-channel input (Extended Data Fig. 11).
Supplementary Figure 1 | External validation of the AI system on the online RSNA dataset. The RSNA dataset consisted of three classes: Normal, Lung Opacity, and Not normal and not lung opacity. 

a, For Lung opacity detection: AUC = 0.964 (95% CI: 0.961-0.967). b, For Abnormal (the rest) detection: AUC = 0.944 (95% CI: 0.942-0.946).
Supplementary Figure 2 | Performance of the AI system in differentiating among three types of pneumonia: viral pneumonia, other types of pneumonia and normal.

Receiver operating characteristic curves (ROC) and normalized confusion matrices of the multiclass model. 

**a**, International validation cohorts. For three-way classification: accuracy = 86.92%, AUC = 0.934 (95% CI: 0.917-0.950). For viral pneumonia vs the rest: accuracy=85.69%, AUC = 0.920 (95% CI: 0.891-0.942).

**b**, Kaggle competition cohort. For three-way classification: accuracy = 83.82%, AUC = 0.948 (95% CI: 0.943-0.953). For viral pneumonia vs the rest: accuracy=85.61%, AUC = 0.916 (95% CI: 0.907-0.924). Normal, absence of pneumonia; Other, other types of pneumonia; Viral, viral pneumonia.
Supplementary Figure 3 | Performance of the AI system in differentiating COVID-19 pneumonia from other viral pneumonia using CXR images.

Receiver operating characteristic curves (ROC) and evaluation matrices for binary classification. External validation on the online BIMCV dataset of the AI system for detecting COVID-19 as viral pneumonia from normal CXRs.
Supplementary Figure 4 | Assessment of COVID-19 severity by using the CXR severity index provided by the AI reviewer and radiologists.

a, An illustration of the segmentation of the lung on the CXR image. Each CXR image is divided into 12 sections and then each section is assigned a severity index to quantify the extent of the opacity (see more details in Methods). b and c, Receiver operating characteristic (ROC) curve and confusion matrix using the deep learning. On the test data set: accuracy = 80.84%, sensitivity = 82.35%, specificity = 80.17%, AUC = 0.915 (95% CI: 0.869-0.954). d, The COVID-19 severity analysis using the CXR severity index by radiologists: accuracy = 76.39%, sensitivity = 79.49%, specificity = 74.84%.
Supplementary Figure 5 | Performance comparisons between the AI model and radiologists in differentiating COVID19 pneumonia patients from the other types of pneumonia and normal CXRs.

a. The performance of the AI system and eight radiologists. Receiver operating characteristic (ROC) curves for diagnosing viral pneumonia from the rest (other types of pneumonia and normal CXRs). The star denoted the operating point of the AI system. Filled dots denote the junior and senior radiologists’ performance (four junior radiologists and four senior radiologists), while the hollow dots denote each junior radiologist’s
performance with the AI’s assistance. Dashed lines linked the paired performance values of each junior radiologist. b-e, Confusion matrices of the three-way classification. b, Confusion matrix of the mean diagnostic performance of the four junior radiologists. c, Confusion matrix of the mean diagnostic performance of the four junior radiologists with the AI’s assistance. d, Confusion matrix of the mean diagnostic performance of the four senior radiologists. e, The AI system demonstrated performance comparable to that of the senior practicing radiologists: accuracy = 91.14%, sensitivity = 95.63%, specificity = 88.57%, AUC = 0.981 (95% CI: 0.970-0.990). The AI models demonstrated non-inferior performance comparable to senior radiologists and superior performance to junior radiologists. f, Penalty metric.
Supplementary Figure 6 | Illustration of the AI system for diagnosing pneumonia with application to COVID-19 patients during clinical deployment.
Supplementary Figure 7 | Schematic illustration of an explainable AI system for pneumonia detection with CXR images.

The system presented main analysis pipeline modules, including the CXR automated standardization module, lung-lesion segmentation, and classification model. **a**, The CXR standardization module consisted of anatomical landmark detection and registration techniques. **b**, The lung and lesion segmentation module. The classification model took the three-channel image produced from the previous process for further analysis.
Supplementary Figure 8 | Distribution of the detection rate over the error margin for 12 anatomical landmarks on CXR images.

The blue curve denoted the DeepLabv3, the orange curve denoted FCN, and the green curve denoted U-Net. The error margin is defined as the localization precision of the normalized distances between the prediction and the ground truth. The performance of the three models on 12 anatomical landmarks: a, midpoint of clavicle left (MCL); b, midpoint of clavicle right (MCR); c, sternal end of clavicle left (SECL); d, sternal end of clavicle right (SECR); e, hilar angle left (HAL); f, hilar angle right (HAR); g, costophrenic angle left (CAL); h, costophrenic angle right (CAR); i, diaphragmatic dome left (DDL); j, diaphragmatic dome right (DDR); k, cardiac diaphragmatic angle left (CDAL); l, cardiac diaphragmatic angle right (CDAR).
Supplementary Figure 9 I Performance of the landmark detection and lung-lesion segmentation on CXR images.

a, An example of anatomical landmark detection. The AI's performance compared with the ground-truth landmarks for detecting 12 landmarks on a CXR image of a normal subject. Yellow points denoted the AI detected landmarks, and green points represented the ground-truth landmarks by the radiologists. The normalized distance errors for the 12 landmarks are 0.032, 0.034, 0.036, 0.016, 0.018, 0.010, 0.015, 0.006, 0.002, 0.004, 0.058, 0.010, respectively. b, Comparison of the overall performance for detecting left and right anatomical landmarks by the AI model based on a. c and d, Validation of the AI system on the publicly available JSRT database.
Supplementary Figure 10 I Performance study of the AI system for a three-way classification.

a, The AI’s classification performance with different datasets for training. The orange curve denoted the AI model trained on CC-CXRI, with an AUC of 0.963. The blue curve denoted the AI trained with the combination of CC-CXRI and CheXpert-P, with an AUC of 0.977. The CXR images in CC-CXRI were diagnosed by RT-PCR, “the gold-standard labels”. The CXR images in CheXpert-P were manually annotated with the “silver-standard labels”.  

b, Performance of the proposed pipeline approach when skipped image registration, lung-lesion segmentation, or both modules. Performance dropped from AUC = 0.977 to AUC = 0.956 when validated on the test set.
Supplementary Figure 11 | Performance study of the AI system using the multi-channel input compared with the single-channel model.

The best results were achieved when using the three-channel model on the test dataset, with an AUC of 0.977 (95% CI: 0.971-0.982).
Supplementary Table 1 | Performance for the multi-label classification on the SYSU testing dataset and the SYSU-PE external dataset.

| Common thoracic disease | Testing dataset | External validation (Physical examination) |
|-------------------------|-----------------|---------------------------------------------|
| Atelectasis             | 0.925 (0.891-0.942) | 0.947 (0.911-0.975) |
| Cardiomegaly            | 0.962 (0.952-0.964) | 0.980 (0.975-0.989) |
| Consolidation           | 0.914 (0.903-0.927) | -- |
| Edema                   | 0.985 (0.977-0.994) | -- |
| Effusion                | 0.978 (0.976-0.980) | 0.939 (0.905-0.968) |
| Emphysema               | 0.945 (0.928-0.950) | 0.936 (0.930-0.967) |
| Fibrosis                | 0.891 (0.881-0.899) | 0.923 (0.915-0.935) |
| Hernia                  | 0.996 (0.996-0.997) | 0.647 (0.643-0.648) |
| Infiltration            | 0.930 (0.926-0.934) | 0.933 (0.896-0.947) |
| Mass                    | 0.926 (0.905-0.940) | 0.914 (0.867-0.965) |
| No finding              | 0.913 (0.909-0.919) | 0.872 (0.858-0.876) |
| Nodule                  | 0.864 (0.845-0.872) | 0.835 (0.813-0.848) |
| Pleural thickening      | 0.898 (0.887-0.912) | 0.913 (0.904-0.921) |
| Pneumonia               | 0.914 (0.910-0.918) | 0.921 (0.907-0.937) |
| Pneumothorax            | 0.967 (0.953-0.979) | -- |
| Overall                 | 0.930 (0.925-0.936) | 0.919 (0.897-0.929) |

AUC (95% CI) was used to evaluate multi-label performance.
Supplementary Table 2 | The AI system’s performance for predicting the pneumonia label on the SYSU testing dataset and the SYSU-PE external dataset.

Each row represents metrics based on the corresponding operation point set at various levels of sensitivity (85%, 90%, 95%, respectively) based on the validation dataset. PPV, positive predictive value; NPV, negative predictive value.

| Sensitivity level on validation set | Cohorts       | Sensitivity | Specificity   | PPV   | NPV   |
|------------------------------------|---------------|-------------|--------------|-------|-------|
| 85%                                | SYSU Testing set | 86.0%       | 84.0%        | 44.3% | 97.6% |
|                                   |               | (84.5-87.1) | (83.5-84.3)  | (42.8-45.2) | (97.3-97.8) |
|                                   | SYSU-PE       | 58.1%       | 98.2%        | 17.2% | 99.7% |
|                                   |               | (52.8-64.2) | (98.1-98.4)  | (15.2-20.1) | (99.7-99.8) |
| 90%                                | SYSU Testing set | 90.3%       | 78.0%        | 37.8% | 98.2% |
|                                   |               | (89.2-91.3) | (77.5-78.5)  | (36.4-38.6) | (98.0-98.4) |
|                                   | SYSU-PE       | 67.6%       | 96.9%        | 12.2% | 99.8% |
|                                   |               | (62.5-72.9) | (96.7-97.1)  | (10.8-14.2) | (99.7-99.8) |
| 95%                                | SYSU Testing set | 95.7%       | 64.4%        | 28.5% | 99.0% |
|                                   |               | (94.9-96.3) | (63.8-65.0)  | (27.5-29.1) | (98.8-99.2) |
|                                   | SYSU-PE       | 78.7%       | 93.0%        | 6.7%  | 99.9% |
|                                   |               | (73.5-82.5) | (92.7-93.3)  | (5.8-7.6)  | (99.8-99.9) |
Supplementary Table 3 | The performance of anatomical landmark detection on CXR images with three backbone models.

We used normalized detection error (with standard deviation) and successful detection rate (SDR) to evaluate the performance of detection the 12 landmarks. L, left; R, right.

| Number | Anatomical landmarks | Normalized distance error | SDR (successful detection rate) at 0.03 error margin |
|--------|---------------------|--------------------------|-----------------------------------------------------|
|        |                     | DeepLabv3  | FCN  | U-Net  | DeepLabv3 | FCN  | U-Net  |
| 1      | Midpoint of clavicle L | 0.026 (±0.032) | 0.037 (±0.051) | 0.039 (±0.056) | 81.68% | 67.87% | 66.97% |
| 2      | Midpoint of clavicle R | 0.027 (±0.035) | 0.038 (±0.054) | 0.041 (±0.053) | 81.98% | 63.66% | 66.67% |
| 3      | Sternal end of clavicle L | 0.022 (±0.030) | 0.040 (±0.041) | 0.042 (±0.040) | 90.09% | 57.36% | 57.96% |
| 4      | Sternal end of clavicle R | 0.023 (±0.028) | 0.042 (±0.045) | 0.043 (±0.047) | 88.89% | 54.05% | 55.26% |
| 5      | Hilar angle L        | 0.030 (±0.035) | 0.040 (±0.044) | 0.042 (±0.043) | 77.18% | 54.95% | 54.95% |
| 6      | Hilar angle R        | 0.029 (±0.035) | 0.043 (±0.047) | 0.043 (±0.043) | 70.27% | 52.85% | 54.05% |
| 7      | Costophrenic angle L | 0.029 (±0.046) | 0.039 (±0.061) | 0.038 (±0.048) | 82.88% | 72.37% | 72.07% |
| 8      | Costophrenic angle R | 0.026 (±0.050) | 0.034 (±0.061) | 0.037 (±0.062) | 88.89% | 79.58% | 78.68% |
| 9      | Diaphragmatic dome L | 0.034 (±0.035) | 0.049 (±0.053) | 0.052 (±0.054) | 64.86% | 47.15% | 50.45% |
| 10     | Diaphragmatic dome R | 0.027 (±0.032) | 0.039 (±0.046) | 0.041 (±0.045) | 77.18% | 56.16% | 56.46% |
| 11     | Cardiac diaphragmatic angle L | 0.025 (±0.036) | 0.036 (±0.057) | 0.040 (±0.056) | 84.38% | 75.38% | 74.17% |
| 12     | Cardiac diaphragmatic angle R | 0.021 (±0.034) | 0.029 (±0.049) | 0.030 (±0.049) | 91.89% | 82.58% | 81.98% |
Supplementary Table 4 | The AI system’s lung segmentation results with three segmentation models on CXR images.

We used Dice Coefficient (with standard deviation) and Intersection over Union (IOU, with standard deviation) metrics to evaluate lung field segmentation performance. L, left; R, right.

| Lung segmentation        | Dice Coefficient | Intersection Over Union |
|--------------------------|------------------|-------------------------|
|                          | DeepLabv3        | FCN                     | U-Net       | DeepLabv3 | FCN | U-Net |
| Lung field L             |                  |                         |             |            |     |       |
|                         | 0.873 (±0.013)   | 0.871 (±0.010)          | 0.860 (±0.013) | 0.775 (±0.021) | 0.771 (±0.015) | 0.755 (±0.020) |
| Lung field R             |                  |                         |             |            |     |       |
|                         | 0.910 (±0.008)   | 0.907 (±0.009)          | 0.903 (±0.012) | 0.835 (±0.014) | 0.829 (±0.015) | 0.824 (±0.020) |
| Periphery of the lung field L |                  |                         |             |            |     |       |
|                         | 0.864 (±0.010)   | 0.861 (±0.007)          | 0.856 (±0.010) | 0.761 (±0.015) | 0.756 (±0.011) | 0.748 (±0.015) |
| Periphery of the lung field R |                  |                         |             |            |     |       |
|                         | 0.893 (±0.007)   | 0.889 (±0.006)          | 0.889 (±0.007) | 0.807 (±0.012) | 0.801 (±0.009) | 0.801 (±0.012) |
Supplementary Table 5 I Performance comparison for lesion segmentation on CXR images: three segmentation models and the human baseline.

We used Dice Coefficient (with standard deviation) and Intersection over Union (IOU, with standard deviation) metrics to evaluate lesion segmentation performance. L, left; R, right.

| Lesion segmentation | Dice Coefficient | Intersection Over Union |
|---------------------|-------------------|--------------------------|
|                     | AI models         |                          |
|                     | DeepLabv3 FCN U-Net | Human baseline          |
| Opacification       | 0.665 (±0.050)    | 0.677 (±0.033)           |
|                     | 0.621 (±0.023)    | 0.705 (±0.23)            |
|                     | 0.500 (±0.058)    | 0.513 (±0.037)           |
|                     | 0.451 (±0.024)    | 0.545 (±0.012)           |
| Interstitial pattern| 0.592 (±0.051)    | 0.588 (±0.039)           |
|                     | 0.480 (±0.055)    | 0.688 (±0.013)           |
|                     | 0.422 (±0.051)    | 0.417 (±0.039)           |
|                     | 0.317 (±0.048)    | 0.524 (±0.414)           |
| Opacity (All lesions)| 0.765 (±0.006)  | 0.756 (±0.014)           |
|                     | 0.741 (±0.020)    | 0.778 (±0.013)           |
|                     | 0.620 (±0.008)    | 0.607 (±0.018)           |
|                     | 0.589 (±0.025)    | 0.636 (±0.014)           |
Supplementary Table 6 | Performance comparison for viral pneumonia discrimination

| Viral pneumonia discrimination | Performance  | Sensitivity   | Specificity   |
|-------------------------------|--------------|--------------|--------------|
| Average-juniors               | 66.36 (62.16-70.25) | 85.64 (83.35-87.96) |
| Average-seniors               | 91.07 (88.70-93.33) | 96.24 (95.09-97.29) |
| AI model                      | 95.62 (91.82-98.60) | 88.57 (84.62-92.17) |
| Juniors + AI model            | 87.22 (84.42-89.93) | 94.86 (93.38-96.19) |
Supplementary Table 7 | The AI system’s performance for landmark detection evaluated on the JRST-SCR dataset.

| Anatomical landmarks         | Normalized distance error | Physical distance error (mm) | Pixel distance error |
|-----------------------------|---------------------------|-------------------------------|---------------------|
| Midpoint of clavicle left   | 0.016(±0.013)             | 5.587(±4.661)                 | 31.926(±26.631)     |
| Midpoint of clavicle right  | 0.013(±0.011)             | 4.627(±3.768)                 | 26.440(±21.533)     |
| Sternal end of clavicle left| 0.010(±0.006)             | 3.579(±2.232)                 | 20.453(±12.752)     |
| Sternal end of clavicle right| 0.010(±0.006)             | 3.603(±2.054)                 | 20.587(±11.735)     |
| Costophrenic angle left     | 0.020(±0.020)             | 7.315(±7.218)                 | 41.802(±14.1248)    |
| Costophrenic angle right    | 0.018(±0.013)             | 6.363(±4.732)                 | 36.362(±27.039)     |
| Cardiac diaphragmatic angle left| 0.026(±0.034)           | 9.380(±12.133)                | 53.600(±69.333)     |
| Cardiac diaphragmatic angle right| 0.011(±0.009)           | 4.086(±3.176)                 | 23.351(±18.148)     |
| Overall                     | 0.016(±0.017)             | 5.568(±6.175)                 | 31.815(±35.287)     |
Supplementary Table 8 | Evaluation metrics of the proposed pipeline approach when skipped image registration, lung-lesion segmentation, or both modules.

Each row represents metrics based on the corresponding operation point set at the sensitivity of 90%.

| Models                   | Sensitivity | Specificity | PPV       | NPV       |
|--------------------------|-------------|-------------|-----------|-----------|
|                          |             |             |           |           |
| Full pipeline            | 90.1%       | 91.1%       | 75.4%     | 96.8%     |
|                          | (87.0-93.0) | (89.7-92.5) | (71.9-79.2)| (95.9-97.8)|
| w/o LungLesionMaps       | 90.1%       | 83.2%       | 61.9%     | 96.5%     |
|                          | (85.9-93.4) | (81.3-85.1) | (57.2-65.4)| (95.0-97.7)|
| w/o ImgReg               | 90.4%       | 83.1%       | 61.9%     | 96.6%     |
|                          | (86.9-92.9) | (81.0-85.0) | (57.4-65.3)| (95.6-97.7)|
| Baseline (w/o both)      | 90.3%       | 76.2%       | 55.4%     | 96.0%     |
|                          | (86.9-93.2) | (74.2-78.5) | (51.6-59.6)| (94.2-97.5)|
**Supplementary Table 9 | Characteristics of the cohort datasets.**

Shown are the numbers of images with ground truth labels by RT-PCR.

| Cohorts          | CC-CXRI | External validation | Population screening |
|------------------|---------|----------------------|----------------------|
| Subjects         | 16,196  | 1,899                | 1,034                |
| Female (%)       | 8,794 (54.3%) | 1,037 (54.6%) | 543 (52.6%)          |
| Age (sd.)        | 31.8 (22.8)     | 41.3 (17.9)          | 33.7 (17.7)          |
| Pneumonia (%)    | 10,718 (66.2%)  | 850 (44.8%)          | 266 (25.7%)          |

**Laboratory tests**

|                       |         |         |         |
|-----------------------|---------|---------|---------|
| Bacterial pneumonia   | 4,471   | 256     | 10      |
| Mycoplasma pneumonia  | 1,499   | 101     | 38      |
| Viral pneumonia       | 4,436   | 240     | 46      |
| COVID-19              | 1,571   | 98      | 0       |
| Other viral pneumonia | 2,865   | 142     | 46      |
| Other Pneumonia       | 312     | 253     | 172     |
| Total                 | 10,718  | 850     | 266     |
References

1. Musalar E, Ekinci S & Ünek O. Conventional vs invert-grayscale X-ray for diagnosis of pneumothorax in the emergency setting. The American journal of emergency medicine 35(9): 1217-1221(2017).

2. Tompson, J., Jain, A., Lecun, Y. & Bregler, C. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. in neural information processing systems 1799-1807 (2014).

3. Newell, A., Yang, K. & Deng, J. Stacked hourglass networks for human pose estimation. in european conference on computer vision 483-499 (2016).

4. Yang, W., Li, S., Ouyang, W., Li, H. & Wang, X. Learning Feature Pyramids for Human Pose Estimation. in international conference on computer vision 1290-1299 (2017).

5. Nibali, A., He, Z., Morgan, S. & Prendergast, L. Numerical Coordinate Regression with Convolutional Neural Networks. (2018).

6. Myronenko, A. & Song, X. Point set registration: coherent point drift. IEEE transactions on pattern analysis and machine intelligence 32, 2262-2275 (2010).