TRACKING AND PREDICTING COVID-19 RADIOLOGICAL TRAJECTORY USING DEEP LEARNING ON CHEST X-RAYS: INITIAL ACCURACY TESTING

1-S. Duchesne[1,2], 1-D. Gourdeau[2,3], P. Archambault[4,5,6], C. Chartrand-Lefebvre[7], L. Dieumegarde[2], R. Forghani[8,9], C. Gagné[10], A. Hains[10], D. Hornstein[11,12], H. Le[12,13], S. Lemieux[1], M.H. Lévesque[1,14], D. Martin[8,9], L. Rosenbloom[12,13], A. Tang[7], F. Vecchio[15], O. Potvin[2], N. Duchesne[1,16]

1- Co-First Authors

Affiliations:

[1] Department of Radiology and Nuclear Medicine, Université Laval, Québec, Québec, Canada
[2] CERVO Brain Research Center, Québec, Québec, Canada
[3] Physics Department, Université Laval, Québec, Québec, Canada
[4] Department of Family and Emergency Medicine, Université Laval, Québec, Québec, Canada
[5] Centre de recherche intégrée pour un système apprenant en santé et services sociaux, Lévis, Québec, Canada
[6] Centre de recherche sur les soins et les services de première ligne de l'Université Laval, Québec, Québec, Canada
[7] University of Montreal Hospital Center; Centre de recherche du CHUM, Montréal, Canada
[8] Department of Diagnostic Radiology, McGill University, Montreal, Canada
[9] Augmented Intelligence & Precision Health Laboratory, Department of Radiology and the Research Institute of McGill University Health Center, Montreal, Canada
[10] Electrical and Computer Engineering Department, Université Laval, Québec, Canada
[11] Department of Internal Medicine, McGill University, Montreal, Canada
[12] Jewish General Hospital, Montreal, Canada
[13] Department of Diagnostic Radiology, McGill University, Montreal, Canada
[14] Institut universitaire de cardiologie et de pneumologie de Québec, Québec, Canada
[15] Brain Connectivity Laboratory, Department of Neuroscience and Neurorehabilitation, IRCCS San Raffaele Pisana, Rome, Italy
[16] Public Health Directory, Centre intégré universitaire santé et services sociaux de la Capitale Nationale, Québec, Québec, Canada

Corresponding author :
Simon Duchesne, P.Eng., Ph.D.
CERVO Brain Research Center
2601 de la Canardière, Québec, Québec
Canada G1J 2G3
simon.duchesne@fmed.ulaval.ca
+1 (418) 663-5741 ext. 4777

Abbreviations

AP:	anterior-posterior
AUC:	area under the (receiver operating) curve
COVID-19:	coronavirus disease
CXR:	chest X-ray
ICU:	intensive care unit

ABSTRACT

Background

Decision scores and ethically mindful algorithms are being established to adjudicate mechanical ventilation in the context of potential resources shortage due to the current onslaught of COVID-19 cases. There is a need for a reproducible and objective method to provide quantitative information for those scores.

Purpose

Towards this goal, we present a retrospective study testing the ability of a deep learning algorithm at extracting features from chest x-rays (CXR) to track and predict radiological evolution.

Materials and Methods

We trained a repurposed deep learning algorithm on the CheXnet open dataset (224,316 chest X-ray images of 65,240 unique patients) to extract features that mapped to radiological labels. We collected CXRs of COVID-19-positive patients from two open-source datasets (last accessed on April 9, 2020)(Italian Society for Medical and Interventional Radiology and MILA). Data collected form 60 pairs of sequential CXRs from 40 COVID patients (mean age ± standard deviation: 56 ± 13 years; 23 men, 10 women, seven not reported) and were categorized in three categories: "Worse", "Stable", or "Improved" on the basis of radiological evolution ascertained from images and reports. Receiver operating characteristic analyses, Mann-Whitney tests were performed.

Results

On patients from the CheXnet dataset, the area under ROC curves ranged from 0.71 to 0.93 for seven imaging features and one diagnosis. Deep learning features between "Worse" and "Improved" outcome categories were significantly different for three radiological signs and one diagnostic (“Consolidation”, “Lung Lesion”, “Pleural effusion” and “Pneumonia”; all $P < 0.05$). Features from the first CXR of each pair

could correctly predict the outcome category between "Worse" and "Improved" cases with 82.7% accuracy.

Conclusion

CXR deep learning features show promise for classifying the disease trajectory. Once validated in studies incorporating clinical data and with larger sample sizes, this information may be considered to inform triage decisions.

INTRODUCTION

The current outbreak of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and the subsequent pandemic of coronavirus disease (COVID-19) is imposing a substantial stress on healthcare systems worldwide. In the majority of COVID-19 cases admitted to intensive care units (ICU) for respiratory distress and hypoxaemia, endotracheal intubation and ventilation are the main treatment options. The high number of infected patients has highlighted the need for more precise decision support systems for determining the need and prognosis after ventilation, especially in healthcare networks where there is a risk of overwhelming system capacity. The recent surviving sepsis campaign recommendations do not make any specific recommendation about this triage decision making (1). Clinical prediction rules are therefore required to help caregivers during this delicate but necessary decision making process, and these rules should be based in part on the prognosis of possible outcomes (2).

While imaging is not indicated for diagnostic purposes in COVID-19, the use of chest radiography to inform prognosis was recommended by Rubin et al. in the recent consensus statement of the Fleischner society, published in this journal: “in a resource-constrained environment, imaging is indicated for medical triage of patients with suspected COVID-19 who present with moderate-severe clinical features and a high pre-test probability of disease” (3).

This recommendation rests on radiological findings for COVID-19, already reported in adults (4) (5) (6) (7). For CT imaging, they comprise (a) bilateral, subpleural, and peripheral ground-glass opacities; (b) crazy paving appearance (ground glass opacities and inter-/intra-lobular septal thickening); (c) air space consolidation; (d) bronchovascular thickening; and (e) traction bronchiectasis. COVID-19 appearance on CXR was reported more recently, with a handful of reports focusing specifically on anterior-posterior (AP CXR) at the bedside, the most common form of imaging in ICUs. CXR may be normal in early or mild disease, but commonly shows abnormal findings in patients requiring hospitalization, in 69% of patients

at the time of admission, and in 80% of patients sometime during hospitalization (8). Most frequent CXR findings are consolidation (59 %) and ground glass opacities (41 %)(8) (9), with a peripheral and lower zone distribution, that are commonly bilateral or multilobar and that tend to be patchy and asymmetric. Pneumothoraces are rare. The main finding over time on CXR was consolidation (8). These findings are not specific however, being similar to other causes of coronavirus and other viral pneumonias (10).

Given the critical nature of the triage decision, it is imperative that as much relevant information as possible be extracted from all available data. This information can help in assessing the risk of mortality, determine priority for initiating ventilation, determine improvements in condition and predict probable clinical trajectory. All of these must be considered in the intervention decision (11). We postulate that AP CXR images may provide such additional information, beyond simply assessing disease spread, in the form of radiomics-like features; and hypothesize that deep learning can extract these features in a reproducible and quantitative manner.

Towards this goal, we present our initial accuracy tests at tracking and predicting radiological evolution in a series of COVID-19 cases for a deep learning system adapted to extract features from AP CXR.

MATERIALS AND METHODS

Study design

This is a retrospective study of a large dataset of CXRs and one convenience series of COVID-19 cases, both open access. This study is conducted and reported based on the STARD criteria (12).

Ethics

The study was approved by the ethics and research review board of our institution [Information withheld to preserve blinding].

Dates of Study

The study was performed between March 15, 2020 and April 9, 2020.

Training, test, and validation sets

Training set: We used as training set the open "CheXpert" chest X-ray dataset from Stanford Hospital, comprised of 224,316 X-ray images taken from 65,240 unique patients (aged 60.4 ± 17.8 years (mean ± standard deviation); 132,636 CXRs from men / 90,777 CXRs from women)(**Table 1**)(13). The CheXpert database was originally extracted from the Stanford Hospital PACS system with the assistance of text mining from the associated radiological reports using natural language processing. The dataset includes both posterior-anterior, anterior-posterior and lateral images. None were from COVID-19 positive patients.

Validation set: The validation set (*n* = 234) for deep learning feature extraction was selected at random within the 500 validation set studies that forms part of the CheXpert dataset (https://stanfordmlgroup.github.io/competitions/chexpert/)(13). The latter was composed of randomly sampled studies from the full dataset with no patient overlap. Three board-certified radiologists from the CheXpert team individually assigned radiological findings and diagnoses to each of the studies in this validation set.

Test set: The test set for COVID-19 was curated from a convenience sample of 40 cases (aged 56 ± 13) years; 23 men, 10 women, seven not reported)(**Table 1**) with sequential AP CXRs accessible in two open access repositories, the Italian Society for Medical and Interventional Radiology (https://www.sirm.org/category/senza-categoria/covid-19/) and the MILA COVID-19 image data collection (https://github.com/ieee8023/covid-chestxray-dataset/). Italian reports were translated to English by one author (F.V.). Care was taken to eliminate double entries between the datasets. A full list of cases, including links to original sources, is included in **Supplementary Material 1**.

Inclusion/Exclusion/Eligibility

Training and Validation set: adult participants having visited Stanford Hospital who underwent CXR for any clinical presentation between October 2002 and July 2017 in both inpatient and outpatient settings (13).

Test set: adult participants initially admitted to emergency departments or ICUs for COVID-19 with sequential AP CXRs and access to summary radiological data.

Index test and reference standard

The canonical index test for confirmation of COVID-19 was a positive polymerase chain reaction test. By patient, sequential AP CXRs were grouped into pairs. There was a total of 60 such pairs, given that some patients received more than two CXRs. The primary outcome for each pair of sequential AP CXRs was a categorical classification of radiological evolution ("Worse"; "Stable"; "Improved") and defined based on the radiological case history provided with the open dataset as well as the images themselves (**Supplementary Material 1**). The history was performed by certified radiologists at the centers providing cases. The categorical classification was done by two authors ([N.D.] (25 years practice); [S.L.] (fourth year residency) for indications regarding radiological outcome. If, when compared to the first CXR of a pair, any additional findings (e.g. new lung opacities, or increase in lung opacities already present) were noted in the second CXR, then the pair was categorized as "Worse". If no change was reported, it was labeled "Stable"; and if improvements were described, the category was "Improved". In the case of discrepancy between authors' reading, a tie-break was provided by the lead study author. There were 44 pairs of successive CXR studies in the "Worse" outcome category; eight in the "Stable"; and eight in the "Improved" outcome categories (**Table 1**). Mean age for the Worse outcome group was 55.5 ± 13.2 years (mean ± standard deviation)(30 men, nine women, five not reported); for the Stable group 56.9 ± 16.7 years (six men, one women, one not reported); and the Improved group 54.7 ± 7.8 years (four men, three women, one not reported).

Deep learning

We trained a deep learning model for feature extraction, taking as input all single-view chest radiographs of the training CheXpert dataset (regardless of patient position) and providing as output the probability of nine radiological findings and one radiological diagnostic category ("Pneumonia", used as a label by the CheXpert authors to represent images that suggested primary infection as the diagnosis). The findings were defined by certified radiologists in CheXpert. We removed the following radiological findings from the training set, given their irrelevance to the purpose of our study: "No Findings"; "Fracture"; "Support Devices". We further removed "Pneumothorax", given its low occurrence in COVID-19. We used a DenseNet121 architecture for all our experiments as it was determined by Irving et al. to achieve the best results on the CheXpert dataset (13). Images were fed into the network with pretrained weights on Imagenet with a size of 320 × 320 pixels. We used the Adam optimizer with default β-parameters of $\beta 1 = 0.9$, $\beta 2 = 0.999$ and learning rate of 1×10^{-4} which was fixed for the duration of the training. Batches were sampled using a fixed batch size of 16 images. We used a weighted binary cross-entropy loss function to account for class imbalance and followed the U-zeroes policy from (13), replacing the uncertain findings with negative findings. We trained for three epochs, saving checkpoints every epoch and using the checkpoint with the lowest validation loss.

Outcome prediction

We then proceeded in testing our hypothesis as follows (**Figure 1**). First was whether deep learning features could track radiological evolution. We used the deep learning network as trained above to extract the findings probabilities from each CXR. We then computed the difference in findings probabilities between sequential CXRs in each pair and tested whether this difference was significant between outcome groups.

Secondly, we attempted to assess the predictive power of the extracted deep learning features, i.e. whether or not the features of the first CXR could predict the outcome category of the CXR pair (Worse or Improved)(**Figure 1**). Instead of using the predicted findings or the findings probabilities from the

deep learning network, we used the output of the last convolutional layer. We reduced the dimensionality of this feature space by selecting only the significantly different features between classes using a Chi-square test, and created a logistic regression model for the prediction of outcome category. We performed a classification using a leave-one-patient-out scheme, removing/testing all pairs associated with this patient in the learning/testing phase.

Software

Deep learning feature extraction was done in Python using the PyTorch library (version 1.4). Our source code is available in the following GitHub repository: https://github.com/medicslab/COVID-19-public.

Statistical Analysis

Demographics were expressed as mean (standard deviation) in years, and differences between outcome groups were tested using the SciPy library. Statistical analysis of the deep learning algorithm consisted in calculating the area under the (receiver operating) curve (AUC) for the determination of label learning on the training set for radiomics; and Mann-Whitney tests to compare results between outcome groups. We used a $P < 0.05$ threshold for significance and calculated effect size (Cohen's d) for each output.

RESULTS

Demographics

There were no statistical differences in age and sex between COVID-19 outcome groups, however CheXpert patients were older and the proportion of males was significantly lower than the whole COVID-19 cohort ($P < 0.05$)(**Table 1**).

Study flowchart

There were 40 patients with at least two sequential CXRs from the open datasets (**Figure 2**). The test was performed on 60 pairs of CXRs.

Deep learning feature extraction

We successfully trained the deep learning algorithm with the aforementioned architecture to extract salient radiological findings, attaining results comparable to those from the original authors of the CheXpert series (**Figure 3**) with AUCs ranging from 0.71 ("Enlarged Cardiomediastinum") to 0.93 ("Consolidation"). We were unable to ascertain AUCs for two radiological findings ("Lung Lesion", "Pleural - Other") due to a lack of sufficient number of cases in the validation dataset.

We generated a class-activation map for the highest-activated radiological sign ("Pneumonia:) on a random COVID-19 patient for illustrative purposes (**Figure 4**).

Outcome prediction

Testing whether deep learning features could track disease trajectory, we applied the learned classifier to the test set, extracted radiological sign probabilities, and computed the differences between sequential CXRs. The four main findings related to COVID-19 are shown in Figure 5 for each outcome group. There were significant inter-group differences (Worse vs. Improved outcome categories; Mann-Whitney $P < 0.05$) for four radiological findings and diagnoses ("Consolidation", "Lung Opacity", "Pleural effusion", and "Pneumonia"), and a significant difference only for the "Pleural effusion" sign between Worse vs. Stable groups (Mann-Whitney $P < 0.05$; **Table 2**). The Cohen's d effect sizes for the Worse vs Improved comparison were "Consolidation": 0.791; "Lung Opacity": 0.783; "Pleural effusion": 0.479, and "Pneumonia": 0.568. For the Worse vs. Stable case, the effect size of "Pleural effusion" was 0.764.

Testing whether deep learning features could predict future outcome, the last convolutional layer was reduced from 1,024 to five features using Chi-square tests ($P < 0.05$). These features were fed to the logistic regression model. Using a leave-one-patient-out cross-validation, performance measures were: accuracy: 82.7% (confidence interval (CI): 69.7% to 91.8%); sensitivity 86.4% (CI: 72.6% to 94.8%); specificity 62.5% (CI: 24.5% to 91.5%); positive likelihood ratio: 2.3 (CI: 0.93 to 5.68); negative likelihood ratio 0.22 (CI: 0.09 to 0.55); positive predictive value: 92.7% (CI: 83.7% to 96.9%); and negative

predictive value: 45.4% (CI: 25.0% to 67.6%). Reversing the order of the pairs (i.e. trying to predict the first CXR using the second of the pair) reduced accuracy to 59.6%, as expected close to chance.

DISCUSSION

Summary

Triage decisions to decide if and when patients should be admitted in the ICU and mechanically ventilated during the current COVID-19 pandemic must be based on sound ethical guidelines and all available prognostic evidence.

We hypothesized that deep learning analysis of baseline CXR and longitudinal changes in feature probabilities could provide objective information to help in these triage decisions. To this end we needed first to prove the ability of deep learning of assessing imaging features and predicting imaging outcomes related to the disease. Consequently, we used a deep learning architecture, pre-trained on a large CXR dataset, and able to learn image features related to nine radiological signs and one pneumonia diagnosis. We applied this algorithm to a series of sequential images from patients with suspected or proven COVID-19. The algorithm was able to significantly detect changes in the images related to either a worsening or improving outcome for the patient and predict the category from the first CXR with reasonably high accuracy (>80%).

Findings and implications for practice

We found that the proposed deep learning architecture was able to derive meaningful feature classes from a large yet disparate number of images. In effect the CheXpert dataset was not curated specifically for pneumonia; images were acquired in a variety of positions (e.g. anterio-posterior, posterior-anterior, and lateral views); and there were a number of non-pathologically related artefacts (e.g. various devices creating image shadows). Yet, it proved robust at extracting those deep learning features that best

correlated to the radiological findings in the validation and test sets, the latter only composed of AP CXRs. The class activation maps of Figure 3 are indicative of the process and show that the deep learning architecture is correctly focusing on relevant areas.

The value of the deep learning features to inform triage decision making however lies not so much in the identification of radiological findings; this task is being done by radiologists themselves in the course of their duty. Rather, it centers on the ability to extract image features, distributed over the image, that may prove salient at the task of predicting outcome. These may be subtle, counter-intuitive, and therefore not part of the usual radiological diagnostic checklist or report; be subject to inter-reader variability; or couched in language that would vary between readers and centers. By quantitatively calculating these features, the model provides objective, repeatable estimates that may have better predictive ability than the binarized appraisal of disease status as exemplified in clinical scores such as the SMART-COP ("multilobar: yes/no")(14).

Study Limitations

This study has some limitations. First, the small size of the test dataset, which inevitably must be augmented to avoid potential bias, most notably case selection, and to confirm generalizability. However, this study represents a proof of concept whose predictive performance can be reassessed as the research community shares additional cases of COVID-19-positive CXRs. Second, the time duration between sequential CXRs was not uniform, which may have diminished the appraisal of the features' sensitivity to change and the predictive ability of our model. Further, the design of the study is retrospective. However, as the pandemic unfolds, new clinical and radiological data will be continuously incorporated in the test set from the open source repositories, and in future cases from the authors' institutions, which will truly test generalizability and solve most of these limitations. The authors would be grateful to any reader that would be willing to contribute to this effort.

To a degree, this report has demonstrated that deep learning features can track radiological progression in COVID-19 but also predict temporal evolution, adding evidence to the conceptualization that there is directional information in static x-rays allowing this prediction. It should be restated however that the reference standard was categorization of *imaging* rather than *clinical* outcomes, such as duration of ICU stay or mortality. Hence, it remains to be determined whether these features can track clinical, rather than radiological, progression. Further studies should therefore assess the added value of deep learning features in clinical decision making using multivariate models incorporating additional variables such as vital signs, oxygenation and ventilation parameters, and assessment of imaging data.

Conclusion

We found that the results were sufficiently convincing to warrant further consideration of deep learning features being incorporated in a clinical prediction rule to support clinicians in making triage decisions. This being said, triage decision making and decisions to institute mechanical ventilation will not only rely on such prognostic decision rules. Shared decision making integrating the best available prognostic models, clinician experience and patient values and preferences about life-sustaining therapies will also be paramount in making these very difficult decisions. Depending on the phase of the COVID-19 viral pandemic, decisions may unfortunately only be based on prognosis, the ethical principle of social justice and availability of mechanical ventilation.

ACKNOWLEDGMENTS

We would like to sincerely thank all patients, members of the Italian Society for Medical and Interventional Radiology and MILA groups for aggregating and/or releasing data. We would further like to thank K. Duchesne (CERVO Brain Center), as well as F. Alù, F. Miraglia, and A. Orticoni from the Brain Connectivity Laboratory of IRCCS San Raffaele Pisana, Rome, Italy for help in data collection and translation. This study has been financed by a COVID-19 Pilot Project grant from the Quebec Bio-Imaging

Network as well as concurrent funding from a Discovery Award to the primary investigator (S.D.) from the National Science and Engineering Research Council of Canada.

There are no relevant conflicts for the authors for this study.

REFERENCES

1. Poston JT, Patel BK, Davis AM. Management of Critically Ill Adults With COVID-19. JAMA. 2020. Epub 2020/03/28. doi: 10.1001/jama.2020.4914. PubMed PMID: 32215647.

2. Arya A, Buchman S, Gagnon B, Downar J. Pandemic palliative care: beyond ventilators and saving lives. CMAJ. 2020. Epub 2020/04/03. doi: 10.1503/cmaj.200465. PubMed PMID: 32234725.

3. Rubin GD, Ryerson CJ, Haramati LB, Sverzellati N, Kanne JP, Raoof S, et al. The Role of Chest Imaging in Patient Management during the COVID-19 Pandemic: A Multinational Consensus Statement from the Fleischner Society. Radiology. 2020:201365. Epub 2020/04/08. doi: 10.1148/radiol.2020201365. PubMed PMID: 32255413.

4. Zhao W, Zhong Z, Xie X, Yu Q, Liu J. Relation Between Chest CT Findings and Clinical Conditions of Coronavirus Disease (COVID-19) Pneumonia: A Multicenter Study. AJR Am J Roentgenol. 2020:1-6. Epub 2020/03/04. doi: 10.2214/AJR.20.22976. PubMed PMID: 32125873.

5. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. JAMA. 2020. Epub 2020/02/08. doi: 10.1001/jama.2020.1585. PubMed PMID: 32031570; PubMed Central PMCID: PMCPMC7042881.

6. Lee EYP, Ng M-Y, Khong P-L. COVID-19 pneumonia: what has CT taught us? The Lancet Infectious Diseases. 2020. doi: 10.1016/s1473-3099(20)30134-1.

7. Salehi S, Abedi A, Balakrishnan S, Gholamrezanezhad A. Coronavirus Disease 2019 (COVID-19): A Systematic Review of Imaging Findings in 919 Patients. AJR Am J Roentgenol. 2020:1-7. Epub 2020/03/17. doi: 10.2214/AJR.20.23034. PubMed PMID: 32174129.

8. Wong HYF, Lam HYS, Fong AHT. Frequency and distribution of chest radiographic findings in COVID-19 positive patients. Radiology. 2020;In press.

9. Kong W, Agarwal PP. Chest Imaging Appearance of COVID-19 Infection. Radiology: Cardiothoracic Imaging. 2020;2(1):In press.

10. Rodrigues JCL, Hare SS, Edey A, Devaraj A, Jacob J, Johnstone A, et al. An update on COVID-19 for the radiologist - A British society of Thoracic Imaging statement. Clinical Radiology. 2020. doi: 10.1016/j.crad.2020.03.003.

11. Truog RD, Mitchell C, Daley GQ. The Toughest Triage - Allocating Ventilators in a Pandemic. N Engl J Med. 2020. Epub 2020/03/24. doi: 10.1056/NEJMp2005689. PubMed PMID: 32202721.

12. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. Radiology. 2015;277(3):826-32. Epub 2015/10/29. doi: 10.1148/radiol.2015151516. PubMed PMID: 26509226.

13. J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, et al., editors. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19); 2019; Honolulu, HW: AAAI.

14. Charles PG, Wolfe R, Whitby M, Fine MJ, Fuller AJ, Stirling R, et al. SMART-COP: a tool for predicting the need for intensive respiratory or vasopressor support in community-acquired pneumonia. Clin Infect Dis. 2008;47(3):375-84. Epub 2008/06/19. doi: 10.1086/589754. PubMed PMID: 18558884.

FIGURES

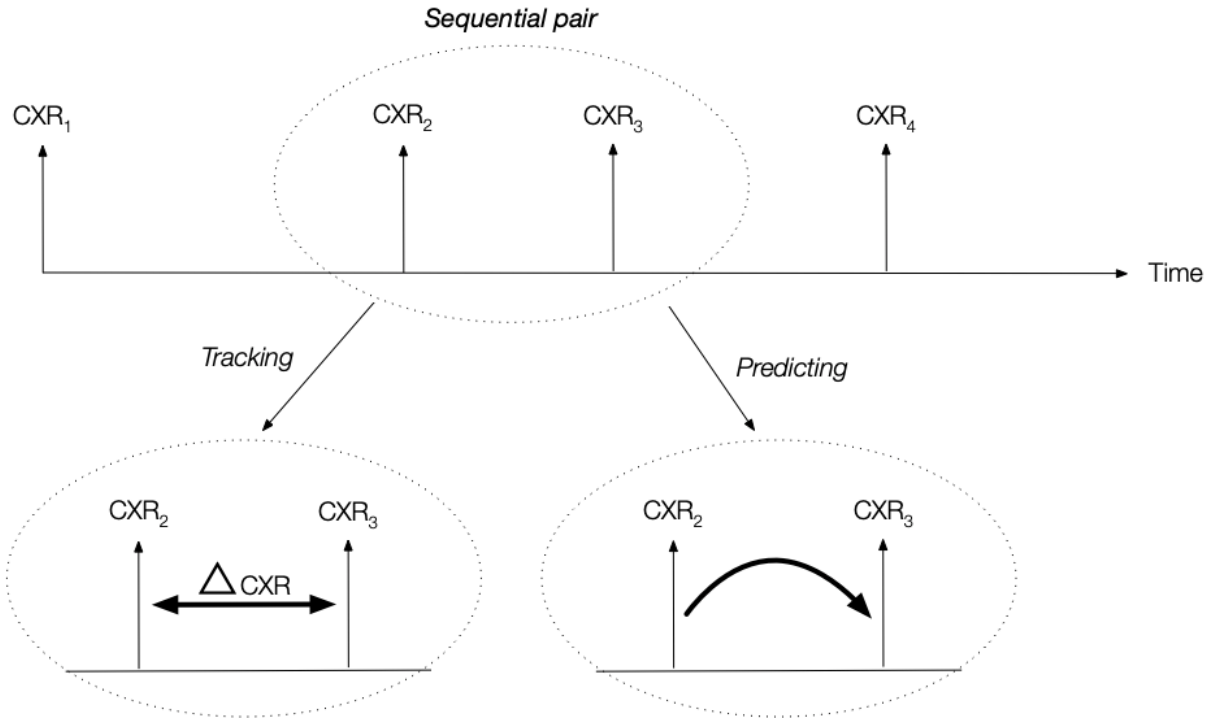

Figure 1 - Experimental design. For each patient, the acquired CXRs formed a series of sequential pairs. For each pair (example shown for CXR_2 and CXR_3), an outcome was defined by judging if the radiological evolution of second CXR of the pair was worse, stable or improved compared to the first. We then tested whether the difference (ΔCXR) in radiological findings probabilities would be statistically different between outcome categories; and secondly if deep learning features from the first CXR of the pair would predict radiological evolution.

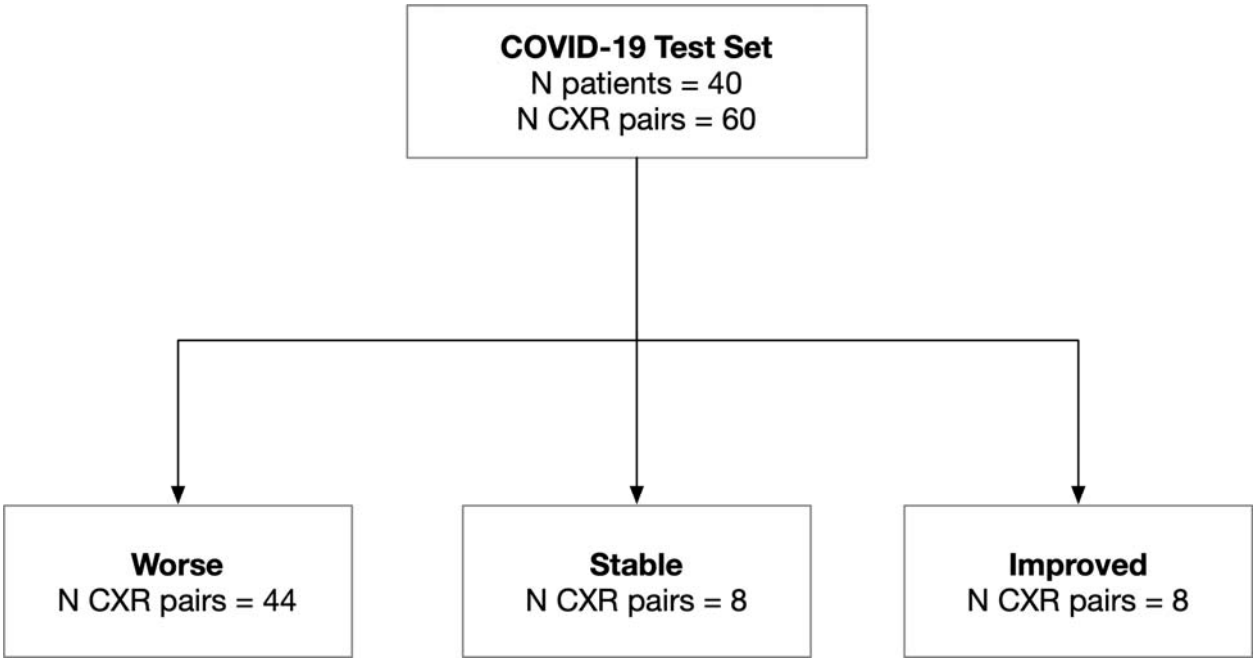

Figure 2 - Study flowchart. Some patients may have CXR pairs in more than one outcome category.

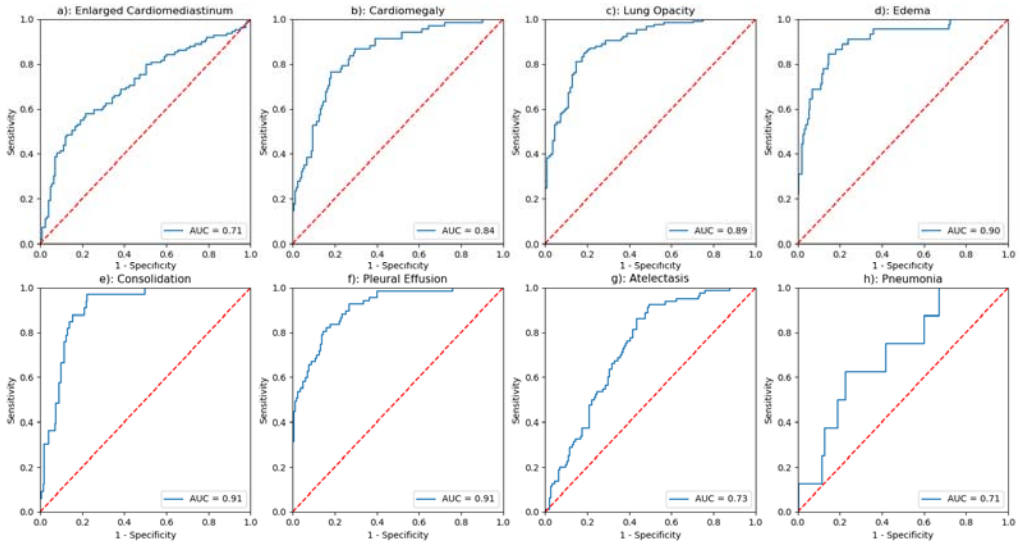

Figure 3 - Results of the deep learning architecture trained on CheXpert for seven radiological findings (**a** to **g**) and one radiological diagnosis (**h**) on a separate 234-cases test dataset, selected at random within the 500 test set studies of the CheXpert dataset (*cf*. Irvin et al. for details). The latter was composed of randomly sampled studies from the full dataset with no patient overlap. Three board-certified radiologists individually annotated each of the studies in this test set.

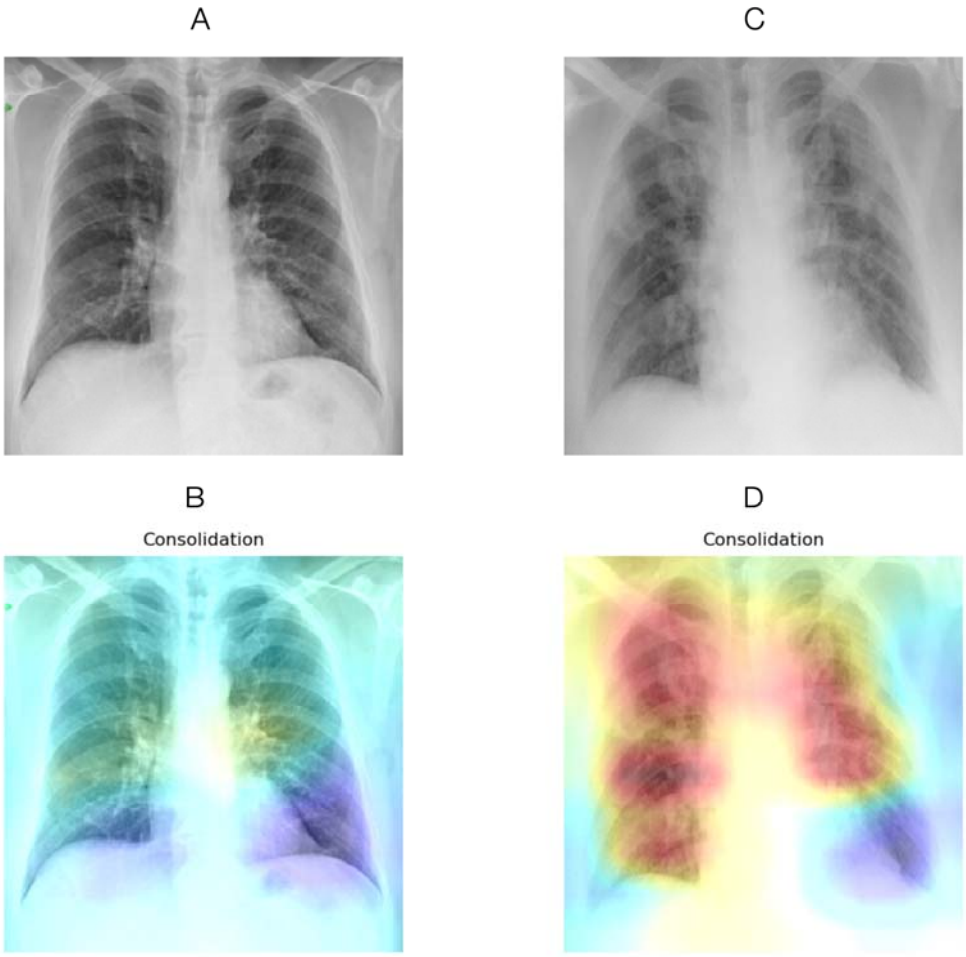

Figure 4 – Original CXRs and class activation maps for a random patient in the COVID-19 dataset. (**A**) CXR at admission, with (**B**) overlaid activation map for the most activated radiological finding ("Consolidation"). (**C**) CXR four days later, with a worsening radiological presentation. The regions activated for the same finding (**D**) now encompass a larger area as the disease has progressed.

Figure 5 – Boxplots show significant inter-group differences between Worse vs. Improving ($P < 0.05$, indicated by symbol *) in deep learning feature probabilities associated with three imaging findings: "Consolidation", "Pleural effusion", "Lung opacity", and diagnosis of pneumonia.

TABLES

Table 1 – Group demographics information

Group	N	Age (Mean ± SD)	*P* value	Sex (M/F/Unknown)	*P* value
CheXpert	223,414	60.4 ± 17.8	-	136,636/90,777	-
COVID-19	60	55.6/13.2	**0.030**	40/13/7	0.024
Worse	44	55.5/ 13.2	0.97	30/9/5	1.0
Stable	8	56.9/16.7	0.85	6/1/1	1.0
Improve	8	54.7/7.8	0.845	4/3/1	0.36

Group legend: COVID-19: all PCR-confirmed COVID-19 cases; Worse, Stable, Improved: primary outcome for each pair of sequential CXR from individual COVID-19 patients, defined based on the *radiological* case history (*Worse*: if any additional findings were noted in the second CXR of a sequential pair, such as new lung opacities, or increase in lung opacities already present; *Stable*: if no change was reported; and *Improved*: if improvements were reported)

Table 2 – Deep learning feature differences between outcome groups (Mann-Whitney)

	Worse vs. Improve	Worse vs. Stable	Stable vs. Improve
Rx sign/Diagnosis	***P* value** **Cohen's d**	***P* value** **Cohen's d**	***P* value** **Cohen's d**
Consolidation	**0.0171** **0.791**	0.184 0.533	0.0946 -0.503
Lung Opacity	**0.0278** **0.783**	0.1908 0.508	0.186 -0.417
Pleural effusion	**0.0412** **0.479**	**0.0109** **0.764**	0.3183 0.52
Pneumonia	**0.0232** **0.568**	0.3378 0.391	0.0946 -0.378