Journal of Digital Imaging
https://doi.org/10.1007/s10278-022-00706-8

# The 2021 SIIM-FISABIO-RSNA Machine Learning COVID-19 Challenge: Annotation and Standard Exam Classification of COVID-19 Chest Radiographs

Paras Lakhani[1] · J. Mongan[2] · C. Singhal[3] · Q. Zhou[3] · K. P. Andriole[4] · W. F. Auffermann[5] · P. M. Prasanna[5] · T. X. Pham[5] · Michael Peterson[5] · P. J. Bergquist[6] · T. S. Cook[7] · S. F. Ferraciolli[8] · G. C. A. Corradi[8] · MS Takahashi[8] · C. S. Workman[9] · M. Parekh[1] · S. I. Kamel[1] · J. Galant[10] · A. Mas-Sanchez[10] · E. C. Benítez[10] · M. Sánchez-Valverde[10] · L. Jaques[10] · M. Panadero[10] · M. Vidal[10] · M. Culiañez-Casas[10] · D. Angulo-Gonzalez[11] · S. G. Langer[12] · María de la Iglesia-Vayá[13] · G. Shih[14]

Received: 21 October 2021 / Revised: 15 September 2022 / Accepted: 16 September 2022

**Abstract**
We describe the curation, annotation methodology, and characteristics of the dataset used in an artificial intelligence challenge for detection and localization of COVID-19 on chest radiographs. The chest radiographs were annotated by an international group of radiologists into four mutually exclusive categories, including “typical,” “indeterminate,” and “atypical appearance” for COVID-19, or “negative for pneumonia,” adapted from previously published guidelines, and bounding boxes were placed on airspace opacities. This dataset and respective annotations are available to researchers for academic and noncommercial use.

**Keywords** Machine Learning · Artificial Intelligence · COVID-19 · Pneumonia · Radiography · Thorax

## Introduction

COVID-19 is a respiratory disease caused by a novel coronavirus (severe acute respiratory syndrome coronavirus-2, or SARS-CoV-2) [1]. Since its discovery in December of 2019, COVID-19 has become an ongoing global pandemic. It is known to be highly infectious [2], more deadly than influenza in adults [3], and has taken a tremendous toll on those affected, having caused over 4.6 million deaths worldwide currently and rising [4].

COVID-19 is diagnosed by detection of genetic viral material, commonly using real-time polymerase chain reaction (RT-PCR) [5]. Imaging studies including chest radiography (CXR) have long been used as part of a standard work-up for patients presenting with respiratory distress [6, 7] and suspected pulmonary infection [8]. Regarding COVID-19, CXR is indicated in patients with moderate to severe features of COVID-19, those with worsening respiratory status or at risk for disease progression [9]. Additionally, CXR may be useful to evaluate other diagnoses in patients with

✉ Paras Lakhani
paras.lakhani@jefferson.edu

[1] Department of Radiology, Thomas Jefferson University, Sidney Kimmel Jefferson Medical College, 111 S 11th St, Philadelphia, PA 19107, USA

[2] University of California San Francisco, San Francisco, CA, USA

[3] MD.AI, New York, NY, USA

[4] Mass General Brigham and Harvard Medical School, Boston, MA, USA

[5] University of Utah Health, Salt Lake City, UT, USA

[6] Medstar Georgetown University Hospital, Washington DC, USA

[7] University of Pennsylvania, Philadelphia, PA, USA

[8] DASA, Alphaville, Barueri, SP, Brazil

[9] Vanderbilt University Medical Center, Nashville TN, USA

[10] Hospital Universitario San Juan de Alicante, San Juan de Alicante, Alicante, Spain

[11] Virgen del Rocio University Hospital, Seville, Spain

[12] Mayo Clinic, Rochester, MN, USA

[13] The Foundation for the Promotion of Health and Biomedical Research of Valencia Region, Valencia, Spain

[14] Weill Cornell Medicine, New York, NY, USA

Published online: 28 September 2022

pulmonary symptoms, such as bacterial pneumonia, pulmonary edema, pleural effusion, and pneumothorax [9].

Artificial intelligence (AI) has been used to facilitate diagnosis of thoracic diseases. In chest radiography, AI has shown promise in detection of pulmonary tuberculosis [10], pneumonia [11], pneumothorax [12], lung cancer [13], and recently COVID-19 [14]. Additionally, AI and quantitative grading of CXRs have been proposed for predicting patient prognosis and assessing treatment response in COVID-19 [15–17], as studies have shown that the degree of parenchymal involvement on CXRs in patients with COVID-19 correlates with outcomes [18–20]. AI approaches for automated CXR severity grading have shown to be precise [21] and can be obtained rapidly.

Since the onset of the COVID-19 pandemic, international groups have been curating and releasing public medical imaging datasets, which may prove useful in creation of AI algorithms, and for teaching and education. This has included the Valencian Region Medical ImageBank (BIMCV) COVID-19 and Medical Imaging Data Resource Center (MIDRC)-RSNA International COVID-19 Open Radiology Database (RICORD) [22, 23]. Many prior medical imaging datasets have been released that have been associated with public AI challenges [24–27]. Such competitions have been important to help advance the state of the art and create a community regarding AI in medical imaging [28].

Our goal was to add value to the existing public BIMCV and MIDRC-RICORD COVID-19 datasets by annotating their respective chest radiographs in a standard fashion for AI models to categorize CXR findings as negative for pneumonia, or typical, indeterminate, or atypical for COVID-19, adapted from previously published guidelines [29]. Additionally, bounding boxes were drawn over the opacities to aid localization and assess disease extent by size of the boxes. The latter may help facilitate development of future models aimed at prognosticating patient outcomes. The annotated dataset is available for public use as part of a machine learning COVID-19 challenge (https://www.kaggle.com/c/siim-covid19-detection) hosted by the Society for Imaging Informatics in Medicine (SIIM), the Foundation for the Promotion of Health and Biomedical Research of Valencia Region (FISABIO), and the Radiological Society of North America (RSNA). The annotation effort is from an international group of radiologists. This paper describes the dataset characteristics, annotation methodology, and rationale.

## Materials and Methods

A total of 10,178 CXRs were used in this annotation effort and challenge, which were obtained from two public sources, MIDRC-RICORD and BIMCV [22, 23]. This included 1000 CXRs from RICORD with only COVID-19 positive patients. The remainder were extracted from the BIMCV database, obtained in December 2020 (22,709 CXRs), consisting of COVID + (16,840 CXRs) and COVID − (5869 CXRs) exams. The BIMCV data then underwent processing including removal of lateral view radiographs and images without associated Digital Imaging and Communications in Medicine (DICOM) tags, which left 16,214 CXRs, which comprised COVID + (12,363 CXRs) and COVID − (3851 CXRs) exams. From this, 9178 random CXRs were extracted and combined with the RICORD data (Fig. 1).

In the final combined dataset of 10,178 images, there are 8042 (79%) COVID-19 positive and 2136 (21%) COVID-19 negative CXRs (Fig. 1). The controls (COVID-19 negative) from the BIMCV dataset include normal chest radiographs, and various pulmonary pathologies other than COVID-19, including imaging findings of bacterial pneumonia, cardiogenic pulmonary edema, pleural effusion, atelectasis, nodule/mass, interstitial lung disease, and pneumothorax. Each of these pathologies was noted among the control cases by the annotating radiologists, but the number of cases exhibiting each pathology was not determined.

The CXRs in this dataset consist of both PA and AP frontal views, obtained from both computed radiography (CR) and digital radiography (DX) devices. All medical imaging data and metadata in the MIDRC-RICORD and BIMCV databases had already been de-identified prior to being reviewed by the radiology annotators.

### Annotation Process

Annotators consisted of non-thoracic radiologists (13/22) and thoracic subspecialty radiologists (9/22) and were from a mix of institutions in North America, South America, and Europe. Nineteen out of the twenty-two annotators were staff practicing radiologists and had fully completed their training; 3/22 were senior radiology residents in training. Radiology annotators were recruited via membership outreach from SIIM and FISABIO.

The annotators were given access to an online web platform (MD.ai, New York, New York) and instructions for its use via a live teleconference. Additionally, the radiologists were provided written annotation instructions, reference materials, and multiple example cases for each category. Twenty-five practice cases were selected from the dataset, and then annotated by a practicing cardiothoracic radiologist (PL; 15 years' experience in radiology, 10 years' experience in thoracic radiology), which constituted the "ground truth." The annotators were then required to independently label these 25 practice cases and their annotations were compared with the "ground truth" label. A minimum threshold of 60% agreement with the "ground truth" labels was required to participate in labeling the full dataset; all 22 annotators met this requirement.

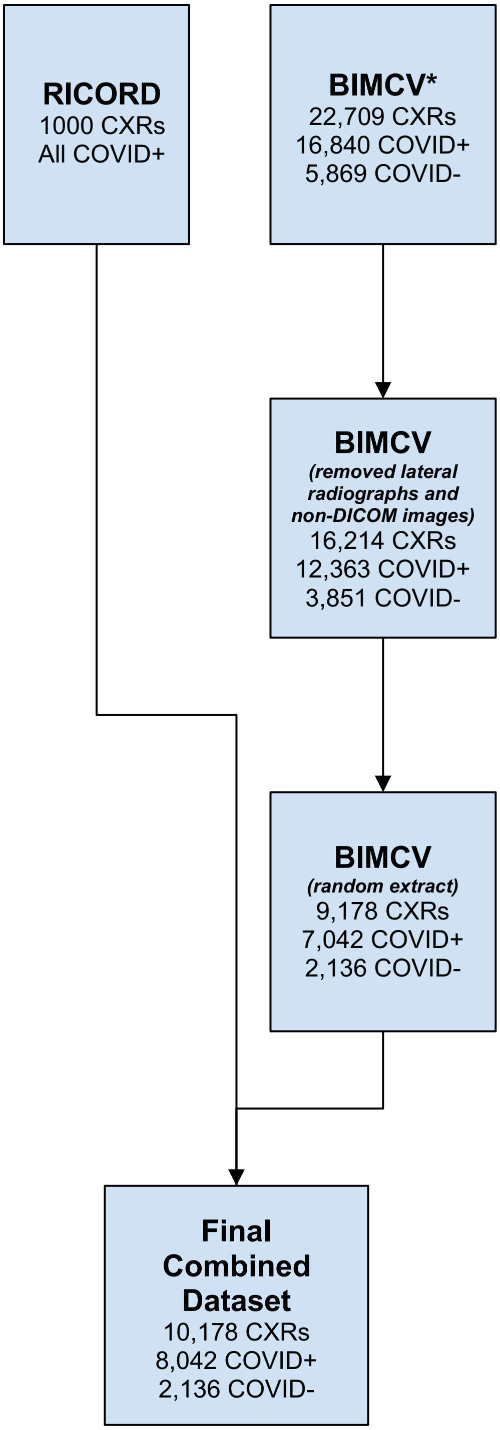

**Fig. 1** The distribution of cases obtained from the Valencian Region Medical ImageBank (BIMCV) and RSNA International COVID-19 Open Radiology Database (RICORD). *The BIMCV data was obtained on 12/2020, and the current dataset is larger

The annotators were required to score the radiographs at an exam-level into one of four categories (Table 1), which was adapted from previously published reporting guidelines [29].

Bounding boxes were placed on pulmonary airspace opacities, whether the exam was scored as typical, indeterminate, or atypical pattern for COVID-19. However, bounding boxes were not placed on pleural effusions, masses/nodules, or pneumothoraces. No bounding boxes were placed for the "negative for pneumonia" category.

In cases where the opacities were in proximity or near-confluent, the annotators were instructed to place one encompassing box rather than multiple separate boxes (Fig. 2), which was intended to improve the standardization of the annotations and decrease variability.

Representative examples of the annotated CXRs that were classified as typical, indeterminate, atypical, and negative for pneumonia are provided on Figs. 3, 4 and 5.

## Inter-rater Reliability

For the 25 practice cases, the median percent agreement and interquartile range (IQR) among the 22 radiologists was calculated. In addition, the intraclass correlation coefficient (ICC) and 95% confidence intervals (CI) were computed using the "irr" package in R (irr package 0.84.1, R version 3.6.2, R Core Team (2020)) [30] to assess for inter-rater reliability, because this was treated as ordinal data; for example, "typical appearance" was considered closer to "indeterminate appearance" than "negative for pneumonia." For this, "negative for pneumonia" was assigned a value of 0, "atypical for pneumonia" a value of 1, "indeterminate for pneumonia" a value of 2, and "typical for pneumonia" a value of 3. The ICC was calculated using a two-way mixed-effects model with absolute agreement. Regarding the ICC values, less than 0.40 was considered poor, 0.40–0.59 as fair, 0.60–0.74 as good, and greater than 0.75 as excellent, per guidelines by Cicchetti [31].

## Results

The distribution of CXR categories (negative for pneumonia, and typical, indeterminate, and atypical appearances of COVID-19) for the entire dataset, and for COVID-19 + and COVID-19 – CXRs are provided on Table 2.

For COVID + patients, most CXRs (5835/8038, 73%) had typical or indeterminate appearances (Table 2). Additionally, 1757/8038 (22%) of COVID + patients had no lung opacities and were "negative for pneumonia." On the other hand, for the COVID-19 – patients, only 359/2140 (17%) of the CXRs had a "typical appearance" of COVID-19, and 1029/2140 (48%) of the COVID – patients were graded as a "negative for pneumonia" (Table 2).

For the 25 practice cases, the median percent agreement among the radiologists was 86% (IQR: 64%, 91%), and the ICC was 0.70 (95% CI: 0.57, 0.83).

**Table 1** COVID-19 exam label categories and corresponding CXR findings

| CXR COVID-19 classification | CXR findings |
|---|---|
| **Typical appearance** | • Bilateral, peripheral, multifocal predominant opacities<br>• Diffuse bilateral opacities including both central and peripheral (e.g., "ARDS pattern")<br>• Diffuse bilateral opacities with fibrosis/reduced lung volumes (long standing ARDS/COVID-19 patients) |
| **Indeterminate appearance** | ***Absence of typical findings and:***<br>• Upper lung zone predominant opacities (e.g., mycobacterial infection, sarcoid, radiation therapy)<br>• Unilateral opacities, even if multifocal<br>• Central opacities with relative peripheral sparing ("batwing appearance", unlike diffuse ARDS) – e.g. cardiogenic edema, PCP pneumonia |
| **Atypical appearance** | ***Absence of Typical or Indeterminate findings and:***<br>• *Pneumothorax* without features of pneumonia<br>• Pleural effusion without features of pneumonia<br>• Mass(es) or nodule(s)<br>• Lobar Pneumonia (e.g., community acquired pneumonia)<br>• Scarring/fibrosis |
| **Negative for pneumonia** | • No lung opacities |

## Discussion

We describe the curation and annotation process of a multi-institutional international COVID-19 CXR dataset for the purposes of a SIIM-FISABIO-RSNA COVID-19 Kaggle AI challenge (https://www.kaggle.com/c/siim-covid19-detection). The labels are adapted from previously published guidelines for reporting COVID-19, in which CXRs are categorized into mutually exclusive categories of negative for pneumonia, or either typical, indeterminate, or atypical appearances of COVID-19 [29]. The rationale for using these guidelines is to improve the consistency in radiology reporting and is based on prior knowledge of radiographic manifestations of COVID-19 [29]. The reporting system should be used in context with the prevalence of COVID-19, as other infectious and non-infectious thoracic diseases can manifest with typical and indeterminate appearances of COVID-19 as outlined in this reporting system [28]. That being said, in areas of high prevalence, typical and indeterminate appearances are more likely to correspond to COVID-19 infection and would warrant communication with the referring clinician as to next steps in patient management.

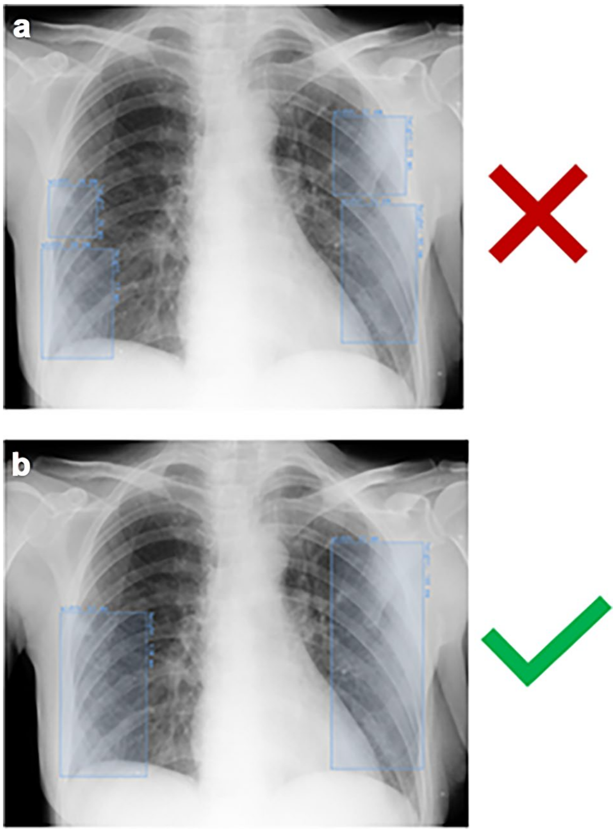

**Fig. 2** For near-confluent opacities, annotators were instructed to draw one larger more encompassing bounding box (**b**), rather than multiple smaller boxes (**a**) to improve standardization and decrease variability in the annotations. The radiographs above demonstrate bilateral airspace opacities in a COVID-19 positive patient

In annotating this dataset, our goal was to have no significant imbalance in the categories "atypical," "indeterminate," and "typical" appearances for COVID-19, and "negative for pneumonia." Prior to annotating the dataset, we decided not to include negative chest radiographs from other public sources that could have balanced the COVID-19 negative and positive studies, as this would have resulted in a higher percentage of "negative for pneumonia" cases. In the end, "atypical" had the lowest representation, followed by "indeterminate," "negative for PNA," and "typical" (Table 2).

While CXRs are not recommended for routine COVID-19 screening [32], they are often obtained in the emergency department, urgent care, and hospital setting in the work-up and management of patients with respiratory complaints including those with suspected or possible COVID-19 [29]. Often, the CXR results are available sooner than PCR, which can take up to 2–3 h. As such, CXR has the potential to influence decision making prior to a PCR result.

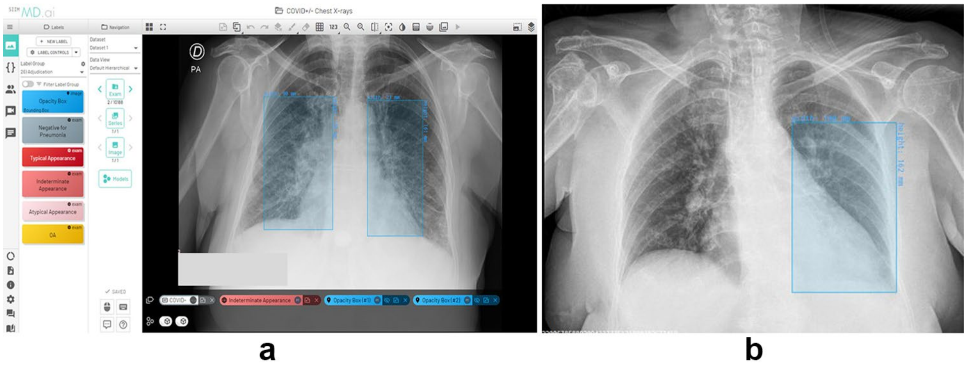

**Fig. 3** Panel **a** shows an indeterminate appearance of COVID-19. There are bilateral central opacities, which are outlined by the blue bounding boxes. This pattern is compatible with cardiogenic pulmonary edema. The image also depicts the web-based annotation platform (MD.ai) used by the radiologists, and the exam-level annotation options. Panel **b** shows an indeterminate appearance of COVID-19. There are unilateral opacities outlined by the blue bounding box

In one prior multi-reader study, AI had comparable performance to that of radiologists for detecting COVID-19 on chest radiographs [13] using PCR as the reference standard.

However, another study indicated that COVID-19 AI algorithms trained in this fashion can learn irrelevant features or "shortcuts" and may not generalize well when presented with external datasets [33]. As such, one of the motivations in this challenge was to add specific annotations, in which radiologists placed bounding boxes over opacities of interest and provided exam-level labels, which may result in models that can better generalize. Additionally, the localization of pulmonary opacities via bounding boxes may prove useful to frontline clinicians, ensure that the algorithm is evaluating the appropriate parts of the image, and provide a visual guide to the distribution of opacities. That said, it would be worthwhile to further evaluate the performance of CXR models trained to simply predict COVID-19 status versus the annotation method used in this challenge, which

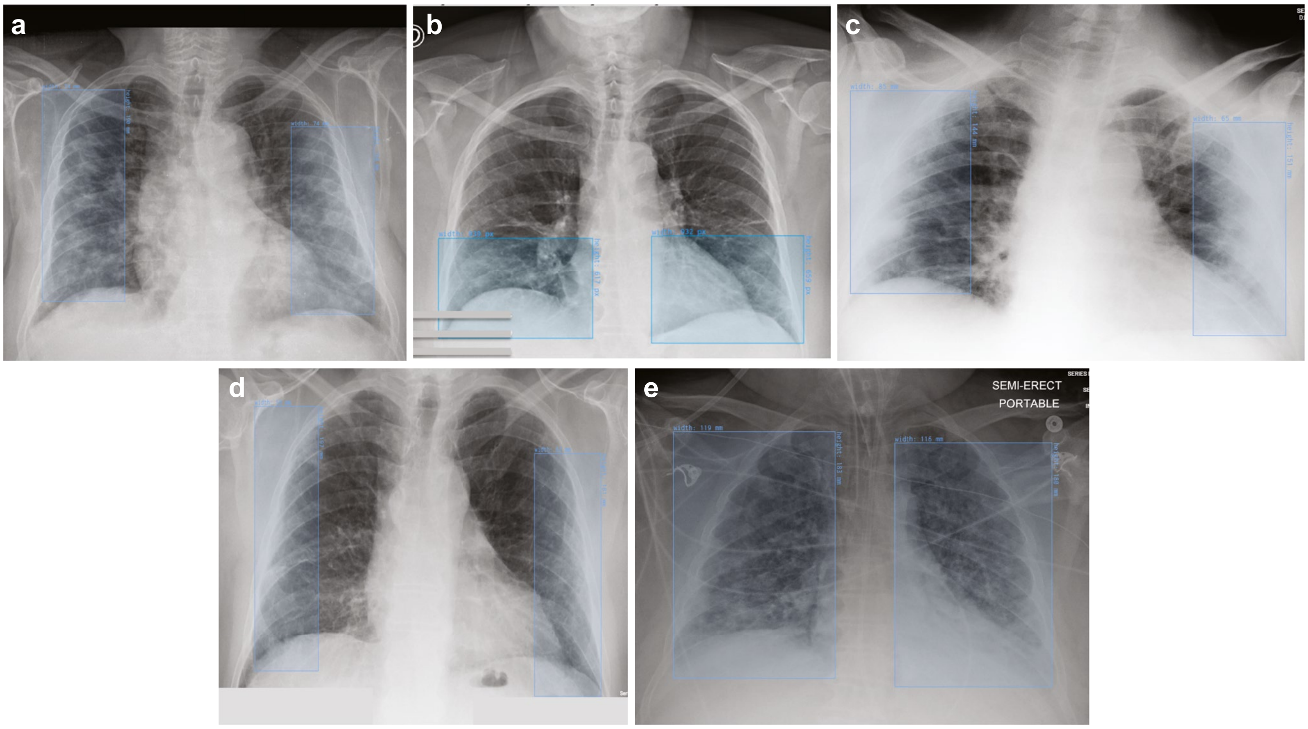

**Fig. 4** Typical appearances of COVID-19. Sample images (Panels **a**, **b**, **c**, **d**) of four CXRs demonstrating typical appearances of COVID-19, manifested by peripheral bilateral airspace opacities, which are outlined by blue bounding boxes. Panel **e** shows diffuse bilateral airspace opacities, both central and peripheral, as outlined by the blue bounding boxes. There is also mild reduction in lung volumes. These findings are commonly seen in severe COVID-19 in hospitalized patients with acute respiratory distress syndrome (ARDS)

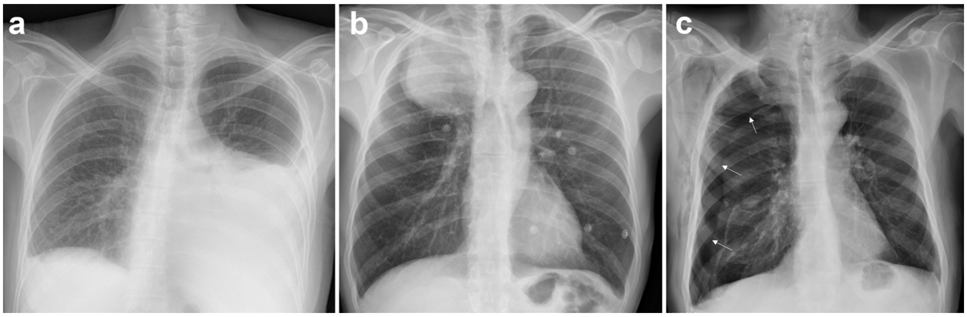

**Fig. 5** Atypical appearances of COVID-19. Panel **a** shows a left pleural effusion. Panel **b** shows a right upper lobe mass. Panel **c** shows a right pneumothorax (demarcated by the white arrows)

**Table 2** Distribution of atypical, indeterminate, typical, and negative for pneumonia exam categories for COVID + and COVID – patients in this dataset

| Exam category | Total exams | COVID + exams | COVID – exams |
|---|---|---|---|
| **Atypical** | 811 | 446 | 365 |
| **Indeterminate** | 1753 | 1365 | 387 |
| **Typical** | 4830 | 4470 | 359 |
| **Negative for pneumonia** | 2784 | 1757 | 1029 |
| **Total** | 10,178 | 8038 | 2140 |

predicts the imaging pattern and location of the opacities on CXR. It should be noted that various external data sources make up the data cohort, potentially leading to better model performance and generalization.

Another rationale for using bounding boxes, in addition to localizing the opacities, is that they could also be used to assess disease extent, as larger bounding boxes indicate a greater disease burden. Polygonal or freehand segmentations of the pulmonary opacities would provide greater accuracy of the disease extent, as boxes may overestimate such in some cases; however, we opted for bounding boxes for efficiency and inter-reader consistency. Secondly, the density of the opacities (e.g., mild, moderate, or severe) could have been annotated, as studies have shown that both extent and density of opacities correlate to disease severity and patient outcomes [16, 19, 20]. These proposed improvements in the annotations would be worthwhile efforts to pursue in the future.

In our data, 22% of CXRs in COVID-19 + patients had no lung opacities and were "negative for pneumonia," which is concordant with prior publications that have shown that CXRs may be normal in a significant number of COVID-19 + patients [9, 34], up to 58% in one study [35]. Similarly, typical appearances were most common in COVID-19 + patients, and negative for pneumonia were most common in COVID-19 – patients. However, it should be noted that the association of these CXR categories with COVID-19 status is dependent on the construct of the dataset, and disease prevalence at the time of imaging.

Regarding COVID-19 and other causes of ARDS, CXR may play a role in predicting patient outcomes, quantifying disease extent, and monitoring disease progression and response to therapy [15]. It should be noted that COVID-19-based AI algorithms that can quantify disease extent on CXR may be repurposed for other pulmonary infectious diseases. While human-based semi-quantitative scoring systems for CXR have been developed in COVID-19 [18–20], AI approaches can be more precise, and are significantly faster [21].

Another limitation of this annotation effort is that only one radiologist annotated each image, which may result in greater variability of the annotations. That being said, all annotators were required to undergo an initial training process, as well as score a minimum threshold to be able to participate in the annotation of the full dataset. For the 25 practice cases, the inter-rater reliability among the radiologists was considered good, with an ICC of 0.70 and a median percent agreement of 86%.

We hope that the annotations from this multi-institutional dataset will aid in the creation of AI CXR models that may help to facilitate diagnosis or predict outcomes in those with suspected COVID-19. This dataset and respective annotations are available to researchers for academic and noncommercial use.

**Acknowledgements** The authors would like to thank and acknowledge the contributions of Anna Zawacki of SIIM, Mariade la Iglesia Vaya PhD of FISABIO, Chris Carr of RSNA, Phil Culliton, and Julia Elliot of Kaggle.

**Author's Contribution** The annotation platform used for this work was provided by MD.ai at no cost to the annotators. Three authors (Singhal C, Zhou Q, Shih, G) serve as stakeholders for MD.ai. Langer SG is a founder of FlowSigma, which is unrelated to this work. Langer SG is a co-principal investigator of the NSF Convergence Accelerator grant (OIA Award Number 2040462). Lakhani P, Langer SG, and Shih G serve on the Machine Learning Steering Committee at SIIM. Shih G, Mongan J, and Andriole KP serve on the Machine Learning

Subcommittee at RSNA. Shih G and Andriole KP serve on the RSNA Radiology Informatics Committee and the RSNA ML Data Standards Subcommittee. Andriole KP serves on the RSNA R & E Education Grant Oversight Committee. Lakhani P, Mongan J, Andriole KP, and Shih G serve on the editorial board of Radiology: Artificial Intelligence. María de la Iglesia-Vayá is a member of FISABIO, which provided some of the data used in this challenge.

**Funding** The SIIM-FISABIO-RSNA COVID-19 Kaggle Challenge was funded via the NSF Convergence Accelerator grant (OIA Award Number 2040462), Hewlett-Packard (HP) and Intel.

**Data Availability** https://www.kaggle.com/c/siim-covid19-detection/data.

**Code Availability** N/A.

## Declarations

**Ethics Approval** This was a retrospective study using non-institutional public HIPAA compliant de-identified imaging data.

**Consent to Participate** N/A.

**Consent for Publication** N/A.

**Competing Interests** No other relevant conflicts of interest.

## References

1. Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. New England Journal of Medicine. 2020 Jan 24.
2. Liu Y, Gayle AA, Wilder-Smith A, et al. The reproductive number of COVID-19 is higher compared to SARS coronavirus. Journal of Travel Medicine. 2020 Mar 13
3. Piroth L, Cottenet J, Mariet AS, et al. Comparison of the characteristics, morbidity, and mortality of COVID-19 and seasonal influenza: a nationwide, population-based retrospective cohort study. The Lancet Respiratory Medicine. 2021 Mar 1;9(3):251-9.
4. World Health Organization, Coronavirus Dashboard. https://covid19.who.int/ Accessed on 9/15/2021.
5. Goudouris ES. Laboratory diagnosis of COVID-19. Jornal de Pediatria. 2021 Feb 22;97:7-12.
6. Hedlund LW, Putman CE. Methods for detecting pulmonary edema. Toxicology and Industrial Health. 1985 Apr;1(2):59-68.
7. Ferguson ND, Fan E, Camporota L, et al. The Berlin definition of ARDS: an expanded rationale, justification, and supplementary material. Intensive Care Medicine. 2012 Oct;38(10):1573-82.
8. Mandell LA, Wunderink RG, Anzueto A, et al. Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. Clinical Infectious Diseases. 2007 Mar 1;44(Supplement_2):S27–72.
9. Rubin GD, Ryerson CJ, Haramati LB, 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 Jul;296(1):172-80.
10. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology. 2017 Aug;284(2):574-82.
11. Rajpurkar P, Irvin J, Zhu K, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225. 2017 Nov 14.
12. Taylor AG, Mielke C, Mongan J. Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: A retrospective study. PLoS Medicine. 2018 Nov 20;15(11):e1002697.
13. Ausawalaithong W, Thirach A, Marukatat S, et al. Automatic lung cancer prediction from chest X-ray images using the deep learning approach. In2018 11th Biomedical Engineering International Conference (BMEICON) 2018 Nov 21 (pp. 1–5). IEEE.
14. Murphy K, Smits H, Knoops AJ, et al. COVID-19 on chest radiographs: a multireader evaluation of an artificial intelligence system. Radiology. 2020 Sep;296(3):E166-72.
15. Kundu S, Elhalawani H, Gichoya JW, et al. How might AI and chest imaging help unravel COVID-19's mysteries?. Radiology: Artificial Intelligence. 2020 May;2(3).
16. Li MD, Arun NT, Gidwani M, et al. Automated assessment and tracking of COVID-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks. Radiology: Artificial Intelligence. 2020 Jul 22;2(4):e200079.
17. Fridadar M, Amer R, Gozes O, et al. COVID-19 in CXR: From detection and severity scoring to patient disease monitoring. IEEE Journal of Biomedical and Health Informatics. 2021 Mar 26.
18. Mushtaq J, Pennella R, Lavalle S et al. Initial chest radiographs and artificial intelligence (AI) predict clinical outcomes in COVID-19 patients: analysis of 697 Italian patients. European Radiology. 2021 Mar;31(3):1770-9.
19. Reeves RA, Pomeranz C, Gomella AA, et al. Performance of a severity score on admission chest radiography in predicting clinical outcomes in hospitalized patients with coronavirus disease (COVID-19). American Journal of Roentgenology. 2021;217(3):623-632.
20. Borghesi A, Zigliani A, Masciullo R, et al. Radiographic severity index in COVID-19 pneumonia: relationship to age and sex in 783 Italian patients. La Radiologia Medica. 2020 May;125(5):461-4.
21. Li MD, Little BP, Alkasab TK, et al. Multi-radiologist user study for artificial intelligence-guided grading of COVID-19 lung disease severity on chest radiographs. Academic Radiology. 2021 Apr 1;28(4):572-6.
22. Vayá MD, Saborit JM, Montell JA, et al. BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients. arXiv preprint arXiv:2006.01174. 2020 Jun 1.
23. Tsai EB, Simpson S, Lungren MP, et al. The RSNA International COVID-19 Open Radiology Database (RICORD). Radiology. 2021 Apr;299(1):E204-13.
24. Halabi SS, Prevedello LM, Kalpathy-Cramer J et al. The RSNA pediatric bone age machine learning challenge. Radiology. 2019 Feb;290(2):498-503.
25. Filice, RW, Stein, A, Wu, CC, et al. Crowdsourcing pneumothorax annotations using machine learning annotations on the NIH chest X-ray dataset. J Digit Imaging 33, 490–496 (2020). https://doi.org/10.1007/s10278-019-00299-9
26. Flanders AE, Prevedello LM, Shih G, et al. Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. Radiology: Artificial Intelligence. 2020 Apr 29;2(3):e190211.
27. Colak E, Kitamura FC, Hobbs SB, et al. The RSNA Pulmonary Embolism CT Dataset. Radiology: Artificial Intelligence. 2021 Jan 20;3(2):e200254.
28. Prevedello LM, Halabi SS, Shih G, et al. Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions. Radiology: Artificial Intelligence. 2019 Jan 30;1(1):e180031.
29. Litmanovich DE, Chung M, Kirkbride RR, et al. Review of chest radiograph findings of COVID-19 pneumonia and suggested

reporting language. Journal of Thoracic Imaging. 2020 Nov 14;35(6):354-60.

30. Gamer M, Lemon J, Fellows I et al. IRR: Various coefficients of interrater reliability and agreement. R package version 0.84.1. https://CRAN.R-project.org/package=irr
31. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess. 1994;6(4):284-290.
32. American College of Radiology. ACR recommendations for the use of chest radiography and computed tomography (CT) for suspected COVID-19 infection. Available at: www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection. Accessed August 3, 2021.
33. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Machine Intelligence. 2021 May 31:1-0.
34. Wong HY, Lam HY, Fong AH, et al. Frequency and distribution of chest radiographic findings in patients positive for COVID-19. Radiology. 2020 Aug;296(2):E72-8.
35. Weinstock MB, Echenique AN, Russell JW, et al. Chest x-ray findings in 636 ambulatory patients with COVID-19 presenting to an urgent care center: a normal chest x-ray is no guarantee. J Urgent Care Med. 2020;14(7):13-8.

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