Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Preliminary Investigation

Multi-Radiologist User Study for Artificial Intelligence-Guided Grading of COVID-19 Lung Disease Severity on Chest Radiographs

Matthew D. Li, MD, Brent P. Little, MD, Tarik K. Alkasab, MD, PhD, Dexter P. Mendoza, MD, MSc, Marc D. Succi, MD, Jo-Anne O. Shepard, MD, Michael H. Lev, MD, Jayashree Kalpathy-Cramer, PhD

Rationale and Objectives: Radiographic findings of COVID-19 pneumonia can be used for patient risk stratification; however, radiologist reporting of disease severity is inconsistent on chest radiographs (CXRs). We aimed to see if an artificial intelligence (AI) system could help improve radiologist interrater agreement.

Materials and Methods: We performed a retrospective multi-radiologist user study to evaluate the impact of an AI system, the PXS score model, on the grading of categorical COVID-19 lung disease severity on 154 chest radiographs into four ordinal grades (normal/minimal, mild, moderate, and severe). Four radiologists (two thoracic and two emergency radiologists) independently interpreted 154 CXRs from 154 unique patients with COVID-19 hospitalized at a large academic center, before and after using the AI system (median washout time interval was 16 days). Three different thoracic radiologists assessed the same 154 CXRs using an updated version of the AI system trained on more imaging data. Radiologist interrater agreement was evaluated using Cohen and Fleiss kappa where appropriate. The lung disease severity categories were associated with clinical outcomes using a previously published outcomes dataset using Fisher's exact test and Chi-square test for trend.

Results: Use of the AI system improved radiologist interrater agreement (Fleiss $\kappa = 0.40$ to 0.66, before and after use of the system). The Fleiss κ for three radiologists using the updated AI system was 0.74. Severity categories were significantly associated with subsequent intubation or death within 3 days.

Conclusion: An AI system used at the time of CXR study interpretation can improve the interrater agreement of radiologists.

Key Words: COVID-19; Chest radiograph; Artificial intelligence; Computer-assisted diagnosis.

Abbreviations: AI Artificial intelligence, **CXR** Chest x-ray, **COVID-19** Coronavirus disease 2019, **PXS** Pulmonary x-ray severity

INTRODUCTION

Radiographic assessment of coronavirus disease 2019 (COVID-19) pneumonia severity has been shown to correlate with clinical endpoints such as hospital admission, intubation, and death (1–4). Thus, reproducible evaluation of lung disease severity communicated in the radiology report may be helpful for clinical triage and treatment decisions. However, in routine clinical practice, there is no widely used standardized method for the communication of lung disease severity on COVID-19 chest radiographs

Acad Radiol 2021; 28:572–576

From the Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts. (M.D.L., J.K.-C.); Division of Thoracic Imaging and Intervention, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts (B.P.L., D.P.M., J.-A.O.S.); Division of Emergency Radiology, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts (T.K.A., M.D.S., M.H.L.); Medically Engineered Solutions in Healthcare Incubator, Innovation in Operations Research Center (MESH IO), Massachusetts General Hospital, Boston, Massachusetts (M.D.S.). Received September 30, 2020; revised January 12, 2021; accepted January 13, 2021. **Address:** Athinoula A. Martinos Center for Biomedical Imaging, 149 13th Street, Charlestown, MA, USA 02129. **Disclosures:**M.D.L. and J.K. report collaborating with Bayer Radiology on addressing regulatory requirements for potential clinical application of this technology (no funding or services provided for the work in this manuscript). M.D.L. reports funding from an RSNA R&E Fund Research Resident/Fellow Grant, outside of the submitted work. B.P.L. is a textbook associate editor and author for Elsevier, Inc. and receives royalties. J.O.S. reports book royalties from Elsevier, Inc. J.K. reports grants from GE Healthcare, non-financial support from AWS, and grants from Genentech Foundation, outside the submitted work. For the remaining authors none were declared. **Address correspondence to:** JK. e-mail: kalpathy@nmr.mgh.harvard.edu

https://doi.org/10.1016/j.acra.2021.01.016

(CXRs) in the radiology report. Widespread adoption of a manual severity scoring system at scale by radiologists is limited by the challenge of educating radiologists.

Artificial intelligence (AI) tools can potentially aid in the reproducible assessment of COVID-19 lung disease severity on chest x-rays (CXRs) by radiologists. One such tool is the Pulmonary X-ray Severity (PXS) score, which is a deep learning-based algorithm that provides a quantitative measure of COVID-19 lung disease severity on CXRs, extracted from the raw pixel data, which can be applied to patients with suspected or confirmed COVID-19 (5,6). A dockerized version of this algorithm has been deployed at a single institution, with a pipeline in place for automated inference on all frontal view DICOM files from CXR studies (7). In our envisioned radiologist workflow, the calculated PXS score (and a categorical grade based on the PXS score) will be delivered to the radiologist at the time of CXR study interpretation through the radiology information system, which can then help guide the report dictation. We hypothesized that radiologists will more consistently grade COVID-19 lung disease severity on CXRs when guided by this AI system.

To investigate this hypothesis, we performed a single institution retrospective multi-radiologist user study to evaluate the impact of the PXS score model-based (5,6) AI system on radiologist interrater agreement in the categorical grading of COVID-19 severity on chest radiographs (CXRs). We also analyzed the association between these severity categories and clinical outcomes in a previously published cohort of CXRs.

METHODS

The IRB of Mass General Brigham (Boston, Massachusetts) exempted this HIPAA-compliant retrospective study.

AI System for Assignment of Lung Disease Severity Categories

A published test cohort of 154 admission CXRs from 154 unique patients hospitalized with COVID-19 at Massachusetts General Hospital (Boston, Massachusetts) at least in part from March 27, 2020 to March 31, 2020 was used, with patient characteristics summarized in this previous work (5). Quantitative lung disease severity scores (PXS scores) were calculated for the frontal view pixel data DICOMs using previously reported deep learning-based models, PXS-original (5) and PXS-updated (6). The updated model is a version of the former tuned on additional outpatient CXR data, with improved performance compared to the original model that was trained on hospitalized patients. These models adapt a Siamese neural network-based approach to quantify disease severity in the CXR along a continuous spectrum (8). The reference standard for lung disease severity that PXS correlates with is the average of multiple radiologist ratings of CXR images using a modified version of the Radiographic Assessment of Lung Edema scoring system (mRALE), which depends on the density and extent of lung opacities (9). Technical details of the implementation are available in the previous reports, with code related to model training and inference shared on GitHub at https://github.com/QTIM-Lab/PXS-score.

For this study, these PXS scores were used to assign categorical grades for lung disease severity, where PXS $\leq$ 2.5 was normal/minimal, 2.5<PXS$\leq$5 was mild, 5<PXS$\leq$9 was moderate, and PXS>9 was severe. The PXS score thresholds were subjectively determined by one author (M.D.L, not a rater in the user study) who visually inspected the study CXRs together with their PXS scores. We chose to use these categorical descriptors because they are already variably used by radiologists in CXR reports and are commonly understood to indicate a severity assessment by both radiologists and other clinicians. Radiologists using the AI system viewed this categorical severity grade (derived from PXS-original or PXS-updated) and numeric PXS score at the time of image interpretation, as presented in a spreadsheet row matched to the CXR study accession. The radiologist would then independently grade the lung disease severity based on their own assessment of the CXR image, viewed in a diagnostic PACS workstation, in the context of the PXS score-assigned grade.

Multi-Radiologist AI User Study

Seven diagnostic radiologists from Massachusetts General Hospital participated in this study, including five thoracic radiologists and two emergency radiologists who routinely interpret CXRs of patients with suspected or confirmed COVID-19.

Four radiologists (two thoracic subspecialty-trained and two emergency non-thoracic subspecialty-trained) independently assigned "normal/minimal", "mild", "moderate", and "severe" grades to the 154 CXRs at PACS workstations without using the AI system. We used a washout time period of at least 10 days to allow the radiologists to forget the images (11, 13, 19, and 38 days, median 16 days). Then, each radiologist independently assessed the same 154 CXRs (randomly shuffled order of studies) while using the PXS-original AI system. Three additional thoracic subspecialty trained radiologists independently assessed the same 154 CXRs using the PXS-updated AI system.

Clinical Outcomes for Categorical Severity Grades

Of the 154 admission CXRs in this study, 142 did not have an endotracheal tube on the CXR image. These data are a subset of a previously published dataset of hospitalized patients with COVID-19 (5), with outcomes data for subsequent intubation or death within 3 days of hospital admission. We used this outcomes data to assess the association between the categorical severity grades and this clinical outcome.

Statistics

Interrater agreement was assessed using Fleiss κ for $\geq$ 3 raters or Cohen κ for 2 raters, as implemented in the *irr* package

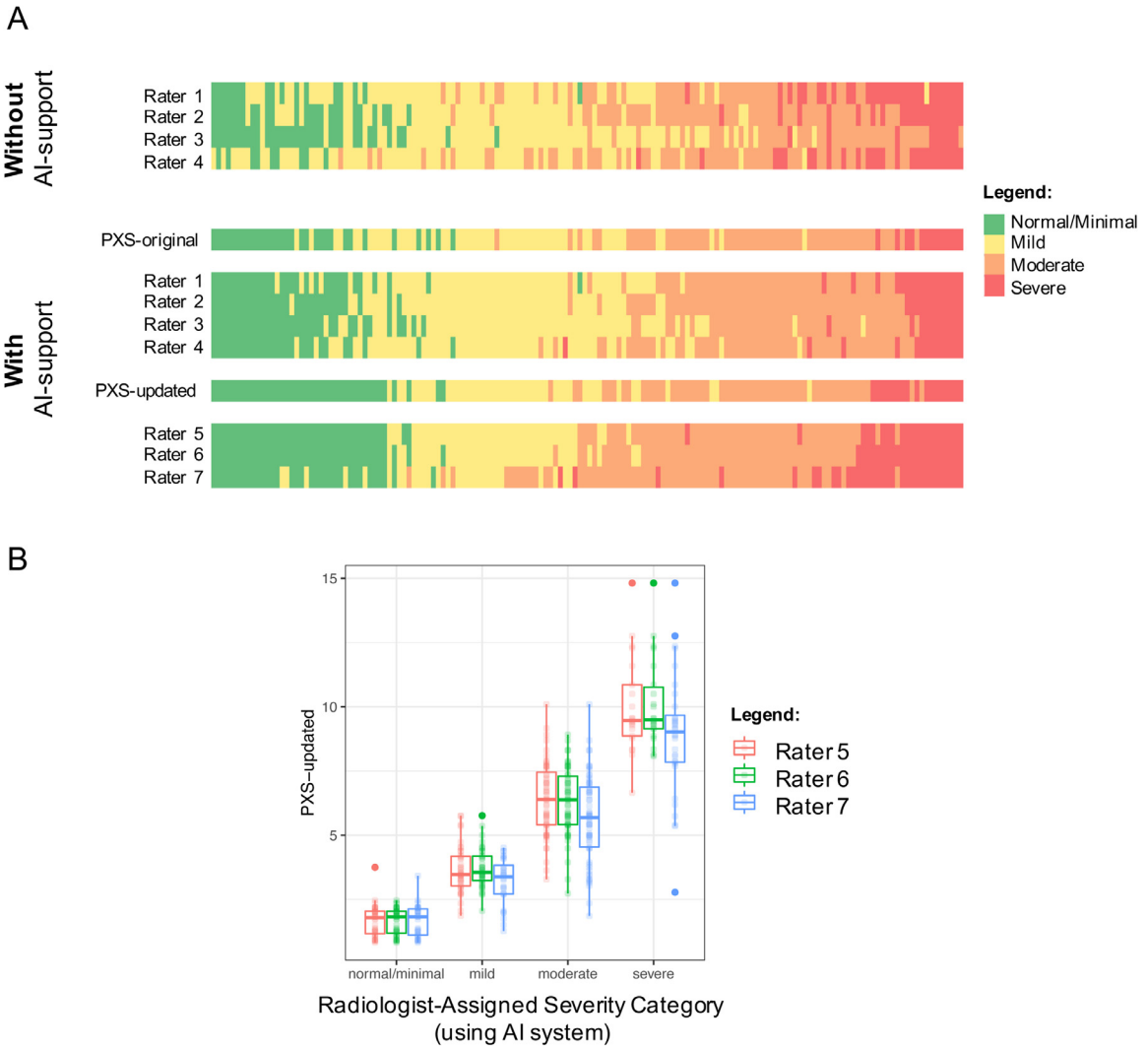

Figure 1. (A) Heatmaps show the categorical COVID-19 lung disease severity gradings on CXRs assessed by four radiologist users without and with AI system guidance (PXS-original) and three radiologist users with AI system guidance (PXS-updated). Each column corresponds to an admission CXR from 154 hospitalized patients with COVID-19. Studies are ordered by the average of the 7 rater assignments when all using AI system guidance. Raters 1, 2, and 5–7 were thoracic radiologists and raters 3 and 4 were emergency radiologists. The thresholds between categorical severity grades differ among the radiologists, but become more consistent with the use of the AI system. (B) Boxplots show PXS-updated scores for each categorical severity grade assigned by three thoracic radiologists for the 154 CXRs using the AI system (PXS-updated), separated by the three raters. (Color version of figure is available online.)

(version 0.84.1) in R (version 3.6.1). Bootstrap 95% confidence intervals were calculated for all of the κ values. For the calculation of odds ratios, Fisher's exact test was implemented in R. To evaluate the trend between AI severity categories and the proportions of patients with the outcome of intubation or death within 3 days, the Chi-square test for trend in proportions was implemented in R. Statistical significance was defined *a priori* as $p < 0.05$. Data visualizations were generated in Excel (version 16.40) and R.

RESULTS

For the four radiologists who assessed the 154 CXRs without and with the PXS-original AI system, Fleiss κ improved from 0.40 to 0.66. In Figure 1A, the thresholds between the different categorical severity grades appear more uniform when the AI system was used, in comparison to before the AI system was used. The Cohen κ between the two thoracic radiologists improved from 0.51 to 0.75, and the Cohen κ between the two emergency radiologists improved from 0.25 to 0.60. These findings are summarized with bootstrap 95% confidence intervals in Table 1.

The Fleiss κ for three thoracic radiologists using the PXS-updated AI system was 0.74 (Fig 1B). We found that these severity categories were significantly associated with subsequent intubation or death within 3 days of admission in 142 CXRs for patients (from the 154 CXR cohort) without an endotracheal tube on the admission CXR, which was a subset of previously published clinical outcomes data from hospitalized patients with COVID-19 (5) (Table 2). A CXR with a

TABLE 1. Summary of radiologist interrater agreement in categorical assessment of COVID-19 lung disease severity on chest radiographs

Radiologist User Cohort	κ Without Using AI System (95% CI)	κ Using AI System (95% CI)
PXS-original (four radiologists)	0.40 (0.34–0.46)	0.66 (0.59–0.72)
2 Thoracic radiologists	0.51 (0.39–0.61)	0.75 (0.66–0.83)
2 Emergency radiologists	0.25 (0.12–0.36)	0.60 (0.48–0.70)
PXS-updated (three thoracic radiologists)	n/a	0.74 (0.68–0.81)

κ, kappa (Fleiss or Cohen where appropriate), n/a, not assessed. Bootstrap 95% confidence intervals (CI) are shown for each κ in parentheses.

TABLE 2. Odds ratios for subsequent intubation or death within 3 days of hospital admission in 142 patients hospitalized at least in part from March 27, 2020 to March 31, 2020 at Massachusetts General Hospital who did not have an endotracheal tube on the admission CXR, using AI-assigned (PXS-original or PXS-updated) severity categories on the CXR

AI Severity Category	PXS-Original			PXS-Updated		
	Patients, *N*	Odds Ratio	*p*-value*	Patients, *N*	Odds Ratio	*p*-value*
Normal/Minimal	31	0.2	0.002	38	0.2	0.004
Mild	51	0.6	0.2	44	0.8	0.7
Moderate	54	2.4	0.03	50	1.9	0.09
Severe	6	4.4	0.09	10	5.5	0.01
Combined moderate/severe	60	4.1	<0.001	60	3.1	0.002

* Fisher exact test.

normal/minimal grade had an odds ratio significantly below 1, while CXRs with moderate or severe grades had odds ratio significantly greater than 1, for both the PXS-original and PXS-updated systems. There were significant trends for increasing AI severity category and increasing proportions of patients with intubation or death within 3 days ($p < 0.001$ for PXS-original and $p < 0.001$ for PXS-updated).

DISCUSSION

For artificial intelligence-based clinical decision support tools to be used in the real world, multi-user studies are important to assess the effects of such models on clinician performance (10). In this study, we show how an AI system can improve the consistency of such assessment, which may improve the value of the radiology report to frontline clinicians. Manual severity grading schemes could alternatively be used (2,3); however, teaching radiologists and promoting widespread usage is challenging, and even if taught, interrater variability remains a problem. The studied AI system is easy for radiologists to use (with no need for training on a customized grading system) and provides clinically relevant information to the reader of the report.

In addition to the PXS score model used for the basis of the AI-guided severity grading in this study, multiple deep learning-based models have been published that also automatically assess COVID-19 lung disease severity from CXR data (11–17). The next logical step in the evolution of such an AI system would be to report the AI-based severity score by itself to the frontline clinicians, potentially bypassing the diagnostic radiologist. However, there are serious concerns related to the potential clinical risks of autonomous AI systems applied to radiological imaging (18). In our AI system, the interpreting radiologist ultimately decides on the categorical severity grade and oversees the information that will be communicated to the frontline clinicians in the report. A more fine-grained severity score could also be checked for quality by a radiologist but implementing such a system is more challenging as the radiologist will then have to learn a more complex system of grading. We propose that a simple normal/minimal, mild, moderate, and severe grading scheme can still be helpful to the frontline clinician, while being easier to implement in practice.

There are multiple limitations to the clinical utility of this AI system. First, the radiologists participating in this study are from an academic medical center and are mostly thoracic radiology subspecialists. The impact that such an AI system would have on radiologists in the community is unclear. However, given that the interrater agreement between the two emergency radiologists also improved when using the AI system, the results are promising. Second, the generalizability of the deep learning model for different patient populations and x-ray machines needs to be assessed before the deployment of such an AI system in clinical practice. The underlying AI model does generalize to different test cohorts, though performance has been shown to vary (6). Third, while improved interrater agreement results in increased consistency, this does not necessarily reflect improved accuracy of assessment. Fourth, odds ratios for clinical outcomes associated with categorical severity grades may differ depending on the population prevalence of comorbidities associated with poor outcomes (e.g., heart failure, chronic kidney disease, obesity (19)) and as treatment options evolve (e.g., prone positioning (20), dexamethasone (21) and other therapeutics

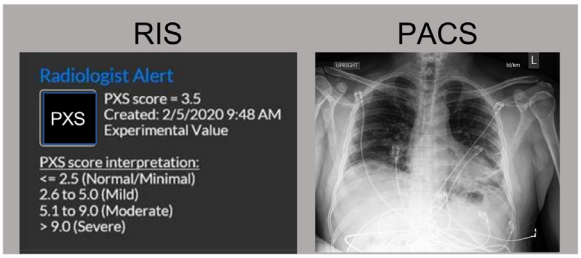

Figure 2. Prototype example for how the AI system information could be presented in the Radiology Information System (RIS), which would open up when a CXR study is opened in a PACS viewer. (Color version of figure is available online.)

in development for COVID-19). Thus, this caveat should be kept in mind when communicating to clinicians reading the radiologist CXR reports. Fifth, while the AI-based severity scores were presented in spreadsheet format for this user study, future implementation in clinical practice will require presentation of the information directly in the PACS or radiology information system. Informatics design considerations will be vital to ensuring that the information is communicated in a way that is easy and nondisruptive to the radiologist workflow (prototype example shown in Fig 2).

In conclusion, an AI system used at the time of CXR study interpretation can improve the interrater agreement of radiologists, which may help improve the communication of clinically relevant information to frontline clinicians. This system is relatively simple for radiologists to use but should be studied prospectively in future work.

ACKNOWLEDGMENTS

We thank anonymous attending radiologist raters from the Massachusetts General Hospital Department of Radiology for their contributions to this study. This study was supported by sundry funds to J.K. This research was carried out in whole or in part at the Athinoula A. Martinos Center for Biomedical Imaging at the Massachusetts General Hospital, using resources provided by the Center for Functional Neuroimaging Technologies, P41EB015896, a P41 Biotechnology Resource Grant supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB), National Institutes of Health. GPU computing resources were provided by the MGH and BWH Center for Clinical Data Science.

REFERENCES

1. Toussie D, Voutsinas N, Finkelstein M, et al. Clinical and chest radiography features determine patient outcomes in young and middle age adults with COVID-19. Radiology 2020:201754. [cited 2020 Sep 5]. http://pubs.rsna.org/doi/10.1148/radiol.2020201754.
2. Borghesi A, Maroldi R. COVID-19 outbreak in Italy: experimental chest X-ray scoring system for quantifying and monitoring disease progression. Radiol Medica 2020; 125:509–513. [cited 2020 Sep 12]. https://pubmed.ncbi.nlm.nih.gov/32358689/.
3. Joseph NP, Reid NJ, Som A, et al. Racial/ethnic disparities in disease severity on admission chest radiographs among patients admitted with confirmed COVID-19: a retrospective cohort study. Radiology 2020:202602. [cited 2020 Sep 5]. http://pubs.rsna.org/doi/10.1148/radiol.2020202602.
4. Kim HW, Capaccione KM, Li G, et al. The role of initial chest X-ray in triaging patients with suspected COVID-19 during the pandemic. Emerg Radiol 2020: 1. [cited 2020 Sep 6]. /pmc/articles/PMC7306559/?report=abstract.
5. 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. Radiol Artif Intell 2020; 2:e200079. [cited 2020 Sep 1]. http://pubs.rsna.org/doi/10.1148/ryai.2020200079.
6. Li MD, Arun NT, Aggarwal M, et al. Improvement and multi-population generalizability of a deep learning-based chest radiograph severity score for COVID-19. medRxiv 2020. doi:10.1101/2020.09.15.20195453. [cited 2020 Sep 18];2020.09.15.20195453.
7. Hashemian B, Manchanda A, Li MD, et al. Clinical deployment and validation of a radiology artificial intelligence system for COVID-19 . 2020 Aug [cited 2020 Sep 1]. https://www.researchsquare.com/article/rs-61220/v1
8. Li MD, Chang K, Bearce B, et al. Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging. npj Digit Med 2020; 3:48. [cited 2020 Mar 29]. http://www.nature.com/articles/s41746-020-0255-1.
9. Warren MA, Zhao Z, Koyama T, et al. Severity scoring of lung oedema on the chest radiograph is associated with clinical outcomes in ARDS. Thorax 2018; 73:840–846.
10. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. BMJ 2020: 368. [cited 2020 Sep 12]. https://www.bmj.com/content/368/bmj.m689.
11. Cohen JP, Dao L, Roth K, et al. Predicting COVID-19 pneumonia severity on chest X-ray with deep learning. Cureus 2020; 12(7). [cited 2020 Sep 11]Available from: /pmc/articles/PMC7451075/?report=abstract.
12. Zhu J, Shen B, Abbasi A, et al. Deep transfer learning artificial intelligence accurately stages COVID-19 lung disease severity on portable chest radiographs. Singh D, editor. Deep transfer learning artificial intelligence accurately stages COVID-19 lung disease severity on portable chest radiographs. PLoS One 2020; 15:e0236621. [cited 2020 Sep 6]. https://dx.plos.org/10.1371/journal.pone.0236621.
13. Blain M, Kassin MT, Varble N, et al. Determination of disease severity in COVID-19 patients using deep learning in chest X-ray images. Diagnostic Interv Radiol 2020. [cited 2020 Sep 11]. https://pubmed.ncbi.nlm.nih.gov/32815519/.
14. Amer R, Frid-Adar M, Gozes O, et al. COVID-19 in CXR: from detection and severity scoring to patient disease monitoring. 2020 [cited 2020 Sep 11]. http://arxiv.org/abs/2008.02150
15. Signoroni A, Savardi M, Benini S, et al. End-to-end learning for semi-quantitative rating of COVID-19 severity on chest X-rays. 2020 [cited 2020 Sep 8]. http://arxiv.org/abs/2006.04603
16. Barbosa EM, Gefter WB, Yang R, et al. Automated detection and quantification of COVID-19 airspace disease on chest radiographs: a novel approach achieving radiologist-level performance using a CNN trained on digital reconstructed radiographs (DRRs) from CT-based ground-truth. 2020 [cited 2020 Sep 11]. http://arxiv.org/abs/2008.06330
17. 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. Eur Radiol 2020: 1–10. doi:10.1007/s00330-020-07269-8. [cited 2020 Sep 30].
18. Fleishon H, Haffty B. Comments of the American College of Radiology regarding the evolving role of artificial intelligence in radiological imaging. [cited 2020 Sep 29]. https://www.acr.org/-/media/ACR/NOINDEX/Advocacy/acr_rsna_-comments_fda-ai-evolvingrole-ws_6-30-2020.pdf
19. Petrilli CM, Jones SA, Yang J, et al. Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: Prospective cohort study. BMJ 2020: 369. [cited 2020 Oct 19]. http://dx.doi.org/10.1136/bmj.m1966.
20. Coppo A, Bellani G, Winterton D, et al. Feasibility and physiological effects of prone positioning in non-intubated patients with acute respiratory failure due to COVID-19 (PRON-COVID): a prospective cohort study. Lancet Respir Med 2020; 8:765–774. [cited 2020 Sep 30]. www.thelancet.com/respiratory.
21. Dexamethasone in hospitalized patients with Covid-19 — preliminary report. N Engl J Med 2020. [cited 2020 Sep 30]. https://www.nejm.org/doi/full/10.1056/NEJMoa2021436.