nature portfolio

https://doi.org/10.1038/s42256-021-00307-0

Supplementary information

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

In the format provided by the authors and unedited

Supplementary Discussion 1 – Rationale for CLAIM and RQS Mandatory Minimums

CLAIM. Motivated to consider only papers that we believe present reproducible research, we decided that some of the CLAIM criteria should be mandatory for a paper to be included in our review. In selecting these criteria, we were conscious that each mandatory criterion chosen would exclude papers, and could introduce bias into our review. Therefore, all ten reviewers performing the quality review met to discuss and agree those criteria that should be mandatory and identified 8 criteria. We believe that these are quite liberal mandatory criteria, being the minimum required for the methodology in a paper to be reproduced. They are the following:

- *Data sources [item 7]*. The data sources must be clearly identified to allow reproducible collection of the same datasets. If only a subset of particular datasets has been used, then it must be detailed how this subset was acquired and how this could be reproduced.

- *Data pre-processing steps [item 9]*. The data we consider is primarily imaging data, therefore we require that the paper details the pre-processing steps in sufficient detail to reproduce. This includes details on how the image intensities were manipulated before being input to the networks, we expect details about any rescaling of the image resolution and the method for rescaling, the number of colour channels in the input image.

- *How data were assigned to partitions; specify proportions [item 20]*. For the training to be reproducible, we expect not only the proportions (or number) of images included within each of the training, validation and holdout cohorts but also the number of images with the outcome.

- *Level at which partitions are disjoint (e.g. image, study, patient, institution) [item 21]*. If a paper has only one image for each patient all obtained from the same center, then we can safely assume these are disjoint at patient level. In the instance that it is clear that there are multiple images for some (or all) patients in the dataset, we expect detail to be given for how the authors mitigated against images appearing in the different partitions for the same patients.

- *Detailed description of model, including inputs, outputs, all intermediate layers and connections [item 22]*. The construction of the architecture must be reproducible to allow the training to be replicated. Therefore, if a paper uses a common architecture and cites it elsewhere, we then deem this to be satisfied so long as the output layers are all detailed. If a custom architecture is employed, then we expect to be able to reproduce this from the detail in the paper.

- *Details of training approach, including data augmentation, hyperparameters, number of models trained [item 25]*. The method for training the model must be discussed in enough detail to allow reproduction, this includes details such as the loss function used, the optimizer, the initial learning rate (and any decay used) along with utilization of data augmentation.

- *Method of selecting the final model [item 26]*. If the authors consider a model that is not trained for a fixed number of epochs, we require that the authors detail how the final model was selected. This could be, for example, use of early stopping criteria to stop training after the validation loss, accuracy or AUC stops improving.

- *Metrics of model performance [item 28].* The metrics used to assess the model performance must be commonly used metrics or defined clearly within the paper.

Any paper which does not satisfy all eight of these will be excluded from the review.

RQS. As for the CLAIM criteria above, we determined that we should only include papers in the review where the documentation and quality of the methodology are of sufficient quality. RQS assigns a score to each paper from a maximum of 36. We had a choice, to either insist on mandatory RQS items to be satisfied (as for CLAIM) or we have a minimum score for papers to be included. We believed the latter approach to be the most liberal and preferred, however the choice of a minimum threshold is potentially controversial. Between the ten reviewers we judged that a minimum score of 6 would lead to inclusion of a paper in the review. This is based on three criteria identified in the RQS listing that we would expect for a traditional machine learning paper, whose scores totalled 6. These are:

- *Image protocol quality - well-documented image protocols [item 1, score 1]*. There must be sufficient details about the images used in the study to allow a reader to immediately identify any potential sources of bias and potential image quality issues.

- *Feature reduction or adjustment for multiple testing - decreases the risk of overfitting [item 5, score 3]*. We expect the authors to consider feature reduction techniques as typical papers will have a large number of initial features and a small number of samples, therefore are extremely liable to overfitting.

- *Validation - the validation is performed without retraining and without adaptation of the cut-off value [item 12, score 2].* Authors should consider validation of their model on an internal holdout or independent external dataset.

Any paper with a score below 6, is excluded from the review. In addition to the minimum score, we also impose the weak criterion that the data sources (and any subsets of them used) must be detailed in a manner which would be reproducible.

Supplementary Discussion 2 – Central Dataset Review of Bias

This document contains the descriptions and PROBAST reviews for all papers in our review. In SECTION 1, we focus on those that rely on publicly available data only and in SECTION 2, we then focus on papers with private datasets. References are provided in Section 3.

SECTION 1. Public datasets

We first review in detail the commonly used public datasets in the literature for COVID-19 detection and prognostication with AI techniques. We use the following abbreviations:

1. COHEN
2. RSNA
3. KERMANY
4. MOONEY
5. CHOWDHURY
6. CHEST X-RAY8
7. CHEST X-RAY14
8. COVIDX
9. CHEXPERT
10. SIRM
11. RADIOPAEDIA
12. EURORAD
13. JSRT

SECTION 1.1. Summary of each public dataset

1. **COHEN**

   Reference: *Joseph Paul Cohen and Paul Morrison and Lan Dao. COVID-19 image data collection, arXiv:2003.11597, 2020 (https://github.com/ieee8023/covid-chestxray-dataset)* (63)

   This public dataset is a compilation of images from several other public sources and includes both CT and Chest X-Ray images of patients with COVID-19, SARS, MERS, ARDS along with other viral and bacterial pneumonias. Data can be submitted directly to the repository on GitHub but it is also retrieved from Radiopaedia (https://radiopaedia.org/), the Italian Society of Medical and Interventional Radiology (SIRM, https://www.sirm.org/category/senza-categoria/covid-19/), Eurorad (https://www.eurorad.org/), Coronacases (https://coronacases.org/). The images are typically JPEG or PNG format. There is ad hoc clinical metadata for many of the patients. The dataset is continually updated by contributors.

2. **RSNA**

   Reference: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge (64)

   The RSNA Pneumonia Detection Challenge was a Kaggle competition run in 2018. The dataset used for the competition is a subset of 30,000 images from the ChestX-Ray8 dataset from NIH (discussed in (6. CHEST X-RAY8)). The 30,000 images consist of 15,000 images that had pneumonia-related labels ('pneumonia', 'infiltration', 'consolidation'), a random 7,500 images with 'no findings' label and an additional random 7,500 scans without the pneumonia-related and 'no findings' labels. Six radiologists then relabelled the 30,000 images and annotated the image with a bounding box for the pneumonia. A test set, a subset of 4,500 images, were reviewed and annotated independently by two Society of Thoracic Radiology radiologists. Therefore, 4,500 of the 30,000 images had three reviews and the remainder had one review. The images are in DICOM format. Bounding box annotations are available in addition.

3. **KERMANY**

   Reference: Kermany, Daniel; Zhang, Kang; Goldbaum, Michael (2018), "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification", Mendeley Data, v2 http://dx.doi.org/10.17632/rscbjbr9sj.2(65)

   This dataset consists of paediatric chest X-rays (anterior-posterior) obtained for patients aged one to five years old at ngzhou Women and Children's Medical Center, Guangzhou, China. There are a total of 5,856 images, including 1,493 depicting viral pneumonia, 2,780 for bacterial pneumonia and 1,583 normal cases.

4. **MOONEY**

   Reference: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia (66)

   This Kaggle dataset is the same as the Kermany *et al.* dataset detailed in (**3. KERMANY**) but hosted on Kaggle.

5. **CHOWDHURY**

   Reference: M.E.H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M.A. Kadir, Z.B. Mahbub, K.R. Islam, M.S. Khan, A. Iqbal, N. Al-Emadi, M.B.I. Reaz, "Can AI help in screening Viral and COVID-19 pneumonia?" arXiv preprint, 29 March 2020, https://arxiv.org/abs/2003.13145. https://www.kaggle.com/tawsifurrahman/covid19-radiography-database (67)

   This public dataset is a collection of COVID-19, viral pneumonia and normal chest X-ray images. The COVID-19 images have been assembled using the Italian Society of Medical and Interventional Radiology (SIRM), the Cohen dataset **(1. COHEN)** and images extracted from 43 publications. The non-COVID images have been collected from the **(4. MOONEY)** Kaggle dataset. The images are in PNG format.

6. **CHEST X-RAY8**

   Reference: Wang, Xiaosong, et al. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. (68)

   This NIH dataset comprises 108,948 frontal chest X-ray images from 32,717 unique patients. Images are labelled using NLP techniques into either normal or one or more of 8 labels: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia and Pneumothorax.

7. **CHEST X-RAY14**

   Reference: https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community (69)

   This NIH dataset comprises 112,120 frontal chest X-ray images from 30,805 unique patients. Images are labelled as normal or have one or more of 14 labels: Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural Thickening, Pneumonia, Pneumothorax.

8. **COVIDX**

   Reference: Wang, Linda, and Alexander Wong. "COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images." arXiv preprint arXiv:2003.09871 (2020). https://github.com/lindawangg/COVID-Net (70)

   This dataset is a compilation of the data contained in many public repositories, namely:

   - **(1. COHEN)** dataset.
   - COVID-19 X-ray images submitted to the GitHub repository https://github.com/agchung/Figure1-COVID-chestxray-dataset via the upload portal https://figure1.typeform.com/to/lLrHwv. These are in JPEG format.
   - COVID-19 X-ray images submitted to the GitHub repository https://github.com/agchung/Actualmed-COVID-chestxray-dataset. These are in JPEG format.
   - **(2. RSNA)**
   - **(5. CHOWDHURY)**

9. **CHEXPERT**

   Reference: Irvin, Jeremy, et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019. (71)

   This public dataset consists of 244,316 chest x-rays of 65,240 unique patients from Stanford Hospital. Using an automated rule-based labelling tool based on the radiologist report, the images are labelled as one or more of: Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomegaly, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax, Support Devices.

10. **SIRM**

    Reference**:** https://www.sirm.org/category/senza-categoria/covid-19/ repository.

    Dataset of the Italian Society of Medical and Interventional Radiology (SIRM). It has associated annotations for ground glass opacity regions, consolidation and pleural effusion. N.B. The SIRM dataset is a subset of the (1. COHEN) dataset already included in this study. This is a public repository that any researcher can contribute to.

11. **RADIOPAEDIA**

    Reference: https://radiopaedia.org/

    This is an online repository (a subset of **(1. COHEN)**) that any researcher can contribute to.

12. **EURORAD**

    Reference: https://www.eurorad.org/

    This is an online repository (a subset of **(1. COHEN)**) that any researcher can contribute to.

13. **CORONACASES**

Reference: https://coronacases.org/

The images are typically JPEG or PNG format. There is ad hoc clinical metadata for many of the patients. The dataset is continually updated by contributors.

14. **JSRT**

Reference: J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-i. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, "Development of a digital image database for chest radiographs with and without a lung nodule," American Journal of Roentgenology, vol. 174, no. 1, pp. 71–74, Jan 2000. [Online]. Available: https://doi.org/10.2214/ajr.174.1.1740071

"This is a digital image database (www.macnet.or.jp/jsrt2/cdrom_nodules.html) of 247 chest radiographs with and without a lung nodule." "One hundred and fifty-four conventional chest radiographs with a lung nodule and 93 radiographs without a nodule were selected from 14 medical centers and were digitized by a laser digitizer with a 2048 × 2048 matrix size (0.175-mm pixels) and a 12-bit gray scale.

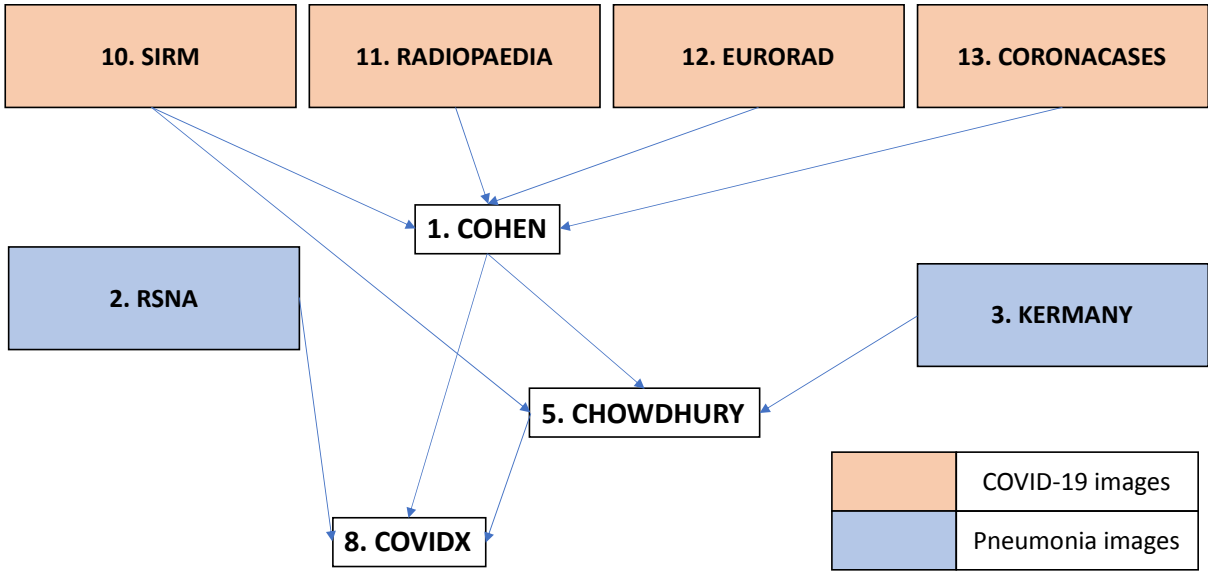

Figure 1: Diagram to show the sources for 'Frankenstein' datasets **(1. COHEN)**, **(5. CHOWDHURY)** and **(8. COVIDX)**. We see **(8. COVIDX)**, in particular, incorporates two datasets, one of which is a subset of the other already.

SECTION 1.2. Papers that use public datasets

In this section we will perform the Domain 1 (Participants) PROBAST review for the following papers which use entirely public datasets:

Update to 24 June 2020

A. (1) Acar 2020
B. (5) Ghoshal 2020
C. (6) Ezzat 2020
D. (7) Luz 2020
E. (8) Tartaglione 2020
F. (9) Gueguim Kana 2020
G. (46) Cohen 2020
H. (13) Heidari 2020
I. (14) Bassi 2020
J. (18) Chen 2020
K. (19) Li 2020
L. (20) Zokaeinikoo 2020

Update to 14 August 2020

M. (21) A 2020
N. (22) AakarshMalhotra 2020
O. (49) D 2020
P. (47) Elaziz 2020
Q. (56) Ghosh 2020
R. (26) Han 2020
S. (23) Rahaman 2020
T. (24) RulaAmer 2020
U. (27) Shah 2020
V. (25) Tsiknakis 2020

Update to 3 October 2020

W. (59) Tamal 2020
X. (30) Zhang 2020
Y. (31) Bararia 2020
Z. (32) Wang 2020
AA. (37) MikhailGoncharov 2020
BB. (62) Yip 2020
CC. (36) MuhammadFarooq 2020

Individual reviews of bias

Update to 24 June 2020

A. (1) **Acar 2020**

DOMAIN 1: Participants
A. Risk of Bias
Describe the sources of data and criteria for participant selection: The datasets used are publicly available datasets which are listed next: (I. COVID-CT-Dataset): This dataset was collected from Jan 19th to Mar 25th. For COVID-19 CT data, they obtain 349 CT images labelled as being positive for COVID-19. These CT images have different sizes. The minimum, average, and maximum height are 153, 491 and 1853. The minimum, average, and maximum width are 124, 383, and 1485. These images are from 216 patient cases. Whilst for the non-COVID-19 cases, they collected a set of non-COVID-19 CT images as negative training examples from four sources see S1, S2, S3 and S4 below (695 images). S1 - https://medpix.nlm.nih.gov/home S2 - https://luna16.grand-challenge.org/ S3 – **(11. RADIOPAEDIA)** S4 - https://www.ncbi.nlm.nih.gov/pmc/ This dataset can be found at: https://github.com/UCSD-AI4H/COVID-CT (II. CORD-19) The COVID-19 Open Research Dataset(CORD-19) is a growing resource of scientific papers on COVID-19 and related historical coronavirus research. Each paper is associated with bibliographic metadata, like title, authors, publication venue, etc, as well as unique identifiers such as a DOI, PubMed Central ID, PubMedID, the WHO Covidence , MAG identifier Link: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge ---------- When combined the two datasets, they got: The chest CT images in our dataset consist of 1,232

COVID-19 and 1,668 healthy images			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		UNCLEAR	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

Authors stated "Some of the CT images collected from the data sources are ignored due to repetition or high correlation problems in the CT images on the source" However, it is unclear how this was performed – there is not further insights in this regard. Moreover, one cannot know whether a PCR test has confirmed for all the COVID-19 positives. There is a wide distribution of sources/regions from which data has been sourced, but bias may exist.

* selected high bias, this particular based on CORD-19, it is a dataset for related historical coronavirus research. Authors when fusion the aforementioned two datasets, it is unclear the protocol followed to do this.

Reviewer 2:

Subset of data used but inclusion/exclusion criteria not given.

B. Applicability

Describe included participants, setting and dates: ***see description related*** *(*COVID-CT-Dataset*) and* (II. CORD-19)

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**UNCLEAR**	**-**

Rationale of applicability rating:

Reviewer 1:

There is a lack of details on the demographic of the datasets. Therefore, it is hard to assess how applicable is the model when this information is not discussed. Also, exclusions are not clear.

Reviewer 2:

No information to judge

B. (5) **Ghoshal 2020**

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: (1. COHEN) and (4. MOONEY).			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

The Cohen dataset contains generally older adults whereas the Mooney (Kermany) dataset contains only children aged 1 to 5. This is a huge bias in the inclusion of participants. We also have no evidence that patients are truly COVID-19 positive for the Cohen dataset.

Reviewer 2:

"We have selected 68 Posterior-Anterior (PA) X-ray images of lungs with COVID-19 cases from [Cohen]" – no description of how these were selected, so this could have introduced bias. Also unclear whether these two datasets represent similar types of cohort, and therefore whether they are directly comparable.

Reviewer 3 (from (20)**):**

The negative cases (4. MOONEY) are only paediatric.

Reviewer 4 (from (9)***):***

It is clear that many of their non-COVID examples are paediatric patients. "Not certain" is not a valid class (in the PROBAST checklist, I would put "high bias" for uniformity of outcomes due to this

dataset). Data augmentation was performed only for the COVID-19 class to ameliorate the class imbalance issue, but this produces clear bias. I rate this a high risk for bias dataset.

B. Applicability

Describe included participants, setting and dates: Discussed earlier.

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of applicability rating:

Reviewer 1:

Based on the significant differences in population, and that COVID-19 generally affects older people more, a model designed using these datasets is deemed to lack applicability.

Reviewer 2:

The purpose is to show that it is possible to develop uncertainty-aware systems for COVID-19 X-ray awareness, and does not claim that their system is ready for clinical use. However, the training data is so biased that it is uncertain whether this has been achieved.

Reviewer 3 (from (20)**):**

Negative cases are only paediatric.

Reviewer 4 (from (9)***):***

The demographics clearly differ significantly among the classes, and many non-COVID-19 examples are of infants. Besides this issue, there are no clear inclusion or exclusion criteria for the patients involved in this study. For these reasons, the training data are not representative of the target population, so the developed model has limited applicability.

C. (6) **Ezzat 2020**

Same as **(B.** (5) **Ghoshal 2020).**

D. (7) **Luz 2020**

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: (8. COVIDX)			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		UNCLEAR	-
1.2 Were all inclusions and exclusions of participants appropriate?		UNCLEAR	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**UNCLEAR**	**-**
Rationale of bias rating: ***Reviewer 1:*** We cannot know for the public data whether a positive PCR test has been received for the COVID-19 positive cases, similarly for other diagnoses which have not been independently confirmed. There is a wide distribution of countries from which data has been sourced and the dataset is heterogeneous for pathologies. But underlying biases may exist. ***Reviewer 2:*** *Different illnesses from different sources.*			
B. Applicability			
Describe included participants, setting and dates: Discussed earlier.			
Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**UNCLEAR**	**-**
Rationale of applicability rating: ***Reviewer 1:*** As before, we cannot be sure of the demographics of each dataset, as these are not discussed to be able to assess how applicable this model would be. ***Reviewer 2:*** Without demographics difficult to judge.			

E. (8) **Tartaglione 2020**

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: (1. COHEN), (2. RSNA), (3. KERMANY), CORDA, Montgomery County X-ray and Shenzhen Hospital X-ray dataset [https://lhncbc.nlm.nih.gov/publication/pub9931] CORDA: This dataset was collected by the radiology unit of Citta della Salute e della Scienza di Torino (CDSS) hospital in Turin from 16th March 2020 to 30th March 2020. It comprises 447 chest x-rays from 386 patients. Of these, 297 were COVID-19 positive cases and the remaining 150 were COVID-19 negative. Montgomery County X-ray dataset: This dataset was collected by the Department of Health and Human Services of Montgomery County, MD, USA as a part of the from Montgomery County's Tuberculosis screening Program. The dataset consists of 138 chest x-rays (posterior-anterior) of which 58 have abnormal tuberculosis manifestations and the remaining 80 are normal. Images are in DICOM format and radiology readings. [Link: http://openi.nlm.nih.gov/imgs/collections/NLM-MontgomeryCXRSet.zip] Shenzhen Hospital X-ray dataset: This dataset was collected by Shenzhen No. 3 Hospital in Shenzhen, China collected from outpatient clinics within a 1-month period, mostly in September 2012. The dataset consists of 662 frontal chest X-rays, of which 336 have manifestations of tuberculosis and 326 are normal. [Link: https://www.kaggle.com/yoctoman/shcxr-lung-mask.]		
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**
Rationale of bias rating: ***Reviewer 1:***			

There are several potential biases in the dataset. Firstly, for the (3. KERMANY) dataset we have only paediatric patients considered, for the Montgomery and Shenzhen dataset there is no detail as to how these subsets were obtained from the overall data collected by the screening trial or outpatient clinics respectively.

In addition, we cannot know for the public data whether a positive PCR test has been received for the COVID-19 positive cases, similarly for other diagnoses which have not been independently confirmed. There is a wide distribution of countries from which data has been sourced and the dataset is heterogeneous for pathologies. But underlying biases may exist.

Reviewer 2:

A highlight in the paper is the so-called CORDA dataset, however, this is unclear is they separated patient level for the training and the test set (it seems not to be available online yet). Moreover from Table I, one can see that some exclusions were done for the COVID+ -- that is, the dataset is described to have 297 COVID+, however, in the CORDA (table I) dataset composition, the COVID+ is train=126 + test= 90 = 216. It is unclear how the exclusion was performed. This is prevalent when combining CORDA with other datasets, the number of positive samples are reduced in other cases but it is unclear how they were excluded.

Moreover, authors stated in the paper "the CORDA dataset is unbalanced and some data balancing is possible, borrowing samples from publicly available non-COVID datasets." However, no further insights are given in this regard.

Besides the above arguments, there is not clear explanation of how the subsets of the datasets were performed when combined.

B. Applicability

Describe included participants, setting and dates: Discussed earlier.

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**UNCLEAR**	**-**

Rationale of applicability rating:

Reviewer 1:

As before, we cannot be sure of the demographics of each dataset, as these are not discussed to be able to assess how applicable this model would be.

Reviewer 2:

Relevant information in terms of patient exclusion is not included. Moreover, there is unclear the demographic of the aforementioned datasets.

F. (9) **Gueguim Kana 2020**

Same as **(B.** (5) **Ghoshal 2020).**

G. (46) **Cohen 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection: (1. COHEN), (2. RSNA), (6. CHEST X-RAY8), (9. CHEXPERT), MIMIC-CXR, PADCHEST and OPENI

MIMIC-CXR is a dataset of 227,835 Chest X-ray imaging studies for 65,379 patients presenting to the Beth Israel Deaconess Medical Center Emergency Department between 2011–2016. A total of 377,110 images are available in the dataset. Associated to these are the free text radiology reports. *[Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 6, 317 (2019). https://doi.org/10.1038/s41597-019-0322-0]*

PADCHEST is a dataset of 160,000 Chest X-ray images obtained from 67,000 patients that were interpreted and reported by radiologists at San Juan Hospital (Spain) from 2009 to 2017, covering six different views. Also provided is additional information regarding image acquisition and patient demographics. The labels for the dataset have been obtained by a supervised labelling, 27,593 reports were labelled and the rest automatically labelled. *[Bustos, Aurelia, et al. "Padchest: A large chest x-ray image dataset with multi-label annotated reports." arXiv preprint arXiv:1901.07441 (2019).]*

OPENI is a biomedical image search engine which researchers can upload images to. This contains both CT and X-ray images with any additional metadata that was uploaded with the image.

	Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?	HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?	HIGH	-

Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

The paper considers 88,079 Chest X-ray images for non-COVID patients, but it is unclear what subsets of the described datasets were used to obtain this figure. It is a potential source of bias if they have been chosen non-randomly.

In addition, we cannot know for the public data whether a positive PCR test has been received for the COVID-19 positive cases, similarly for other diagnoses which have not been independently confirmed. There is a wide distribution of countries from which data has been sourced and the dataset is heterogeneous for pathologies. But underlying biases may exist.

Reviewer 2:

There is a large degree of image overlap between the different datasets used, with RSNA dataset comprising of a subset of Chest X-RAY8 dataset. Automatic labelling, especially with the Chest X-RAY8 dataset, has been shown to be unreliable. Furthermore, image file types vary between the datasets, with the COHEN dataset using JPEG/PNG formats while the RSNA dataset uses DICOM.

B. Applicability

Describe included participants, setting and dates: Discussed earlier.

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of applicability rating:

Reviewer 1:

As before, we cannot be sure of the demographics of each dataset, as these are not discussed to be able to assess how applicable this model would be.

Reviewer 2:

Unclear with regards to the demographics and severity of the patients included in this dataset.

H. (13) **Heidari 2020**

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: The dataset consists of 8474 chest X-rays of which 415 are of COVID-19 patients, 2880 normal examinations and 5179 of patients with pneumonia. The dataset was assembled from: (3. KERMANY), (5. CHOWDHURY) and N. Chen, M.Zhou, X.Dong, "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study," The Lancet,395(2020),507-513. Kermany consist of AP chest x-rays in paediatric patients, the Chowdhury dataset consists of publications, the Italian Society of Radiology database and the Cohen dataset. The Chen dataset consists of images from five hospitals in China.			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**
Rationale of bias rating: ***Reviewer 1:*** Taking pictures from publications is inappropriate, the set of normal controls contains a paediatric dataset. ***Reviewer 2:*** Using the Kermany dataset as non-COVID-19 examples introduces clear bias because this dataset contains scans of paediatric patients (a population that is not represented in the COVID-19 examples or in the target population). Using the Chowdhury and Cohen datasets introduce images taken from publications, which follow no clear inclusion or exclusion criteria and can introduce bias due to inconsistent changes in resolution or contrast.			

B. Applicability			
Describe included participants, setting and dates: As discussed above.			
Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**LOW**	**-**

Rationale of applicability rating:

Reviewer 1:

The manuscript fits the review question.

Reviewer 2:

Because many of the non-COVID-19 examples are paediatric and many other examples are included without following inclusion or exclusion criteria, these data are not representative of the target population, and this model has limited applicability.

I. (14) **Bassi 2020**

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: (5. CHOWDHURY)			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

The majority of the (5. CHOWDHURY) dataset consists of scans from (1. COHEN) and (4. MOONEY). The biases for these datasets have been discussed in **(B.** (5) **Ghoshal 2020).** Therefore, this is scored high risk for the same reasons. No evidence COVID-19 positive patients are confirmed.

Reviewer 2:

As in **(B.** (5) **Ghoshal 2020).**

B. Applicability

Describe included participants, setting and dates: See **(B.** (5) **Ghoshal 2020).**

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of applicability rating:

Reviewer 1:

See **(B.** (5) **Ghoshal 2020).**

Reviewer 2:

See **(B.** (5) **Ghoshal 2020).**

J. (18) **Chen 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

Public datasets: COVID-CT-Dataset (Zhao et al. 2020) and **(10. SIRM)**

COVID-CT-Dataset:

COVID-19 positive cases: 349 COVID-19 images from 216 patients collected from preprints about COVID-19 from medRxiv and bioRxiv from January 19th to March 25th. First manually selected all CT images, and used associated caption to decide if it is COVID-19 positive.

COVID-19 negative cases: MedPix database (195 images from 55 patients), LUNA (36 images from 17 patients), PMC (202 images from 55 patients), **(11. RADIOPAEDIA)** (30 images from 2 patients)

leading to a total of 463 images from 55 patients.

Pretrained using DeepLesion (Yan et al. 2018) [over 32,000 lung CT images] and Lung Image Database Consortium Image Collection (LIDC-IDRI) [224,617 CT images]

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		UNCLEAR	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ **high**/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

There is a much larger number of images than the number of patients, with 30 images procured from radiopaedia from only 2 patients. Images collected from hundreds of preprints is difficult to determine how the diagnosis for each patient was determined. Using different sources for the positive and negative cases leads to a risk of source bias.

Reviewer 2:

Several possible sources of bias are present: It is unclear, whether the SIRM cases and the Zhao cases overlap, which might result in the same patient appearing both in training and testing (via internal cross-validation).

B. Applicability

Describe included participants, setting and dates:

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ **unclear**)*	**UNCLEAR**	**-**

Rationale of applicability rating:

Reviewer 1:

Demographics and severity unclear.

Reviewer 2:

Demographics information is incomplete.

K. (19) **Li 2020**

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: (1. COHEN) and (2. RSNA)			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		LOW	-
1.2 Were all inclusions and exclusions of participants appropriate?		UNCLEAR	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**UNCLEAR**	**-**
Rationale of bias rating:			

Reviewer 1:

The RSNA dataset can be regarded as high quality and we can have confidence in the labels for the data in this dataset. For Cohen, we cannot know for the public data whether a positive PCR test has been received for the COVID-19 positive cases, similarly for other diagnoses which have not been independently confirmed. There is a wide distribution of countries from which data has been sourced and the dataset is heterogeneous for pathologies. But underlying biases may exist.

Reviewer 2:

Authors used two well-known datasets (1. COHEN) and (2. RSNA), they sampled the images per patient for each split to avoid images from the same patient be included in both training and test sets. The protocol for the participants/dataset is appropriate. However, it is unclear several important factors from (1. COHEN) including PCR confirmation and a highly heterogeneous dataset.

B. Applicability			
Describe included participants, setting and dates: Discussed earlier			
Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**UNCLEAR**	**-**

Rationale of applicability rating:

Reviewer 1:

As before, we cannot be sure of the demographics of each dataset, as these are not discussed to be able to assess how applicable this model would be.

Reviewer 2:

There is a lack of details on the demographic of the datasets. Therefore, it is hard to assess how applicable is the model when this information is not discussed.

L. (20) **Zokaeinikoo 2020**

Same as **(B.** (5) **Ghoshal 2020).**

Update to 14 August 2020

M. (21) **A 2020**

Same as **(B.** (5) **Ghoshal 2020).**

N. (22) **AakarshMalhotra 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection: Combination of: **(1. COHEN)**, **(7. CHEST X-RAY14)**, **(10. CHEXPERT)**, **(10. SIRM)**, **(11. RADIOPAEDIA)**, BSTI, **(12. EURORAD)**, TWITTER.

BSTI: This is a repository of the British Society of Thoracic Imaging (https://bsticovid19.cimar.co.uk/worklist/?embedded=) containing CXR and CT imaging along with some unstructured metadata.

TWITTER: Images were extracted from a thread on the Twitter account 'ChestImaging', consisting of 110 COVID-19 CXRs from Spain.

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

Extensive use of public repositories with no validation for COVID-19 status, heavy overlap between datasets so likely for optimistic performance in model development. No inclusion / exclusion criteria specified.

Reviewer 2:

The authors have included multiple copies of the same dataset. Despite splitting data at a subject level into training and testing sets, this poses the risk of contaminating the testing set with known cases. With the data from twitter, it is unclear if all patient have had an independent confirmation of their diagnosis apart from the radiological assessment.

B. Applicability

Describe included participants, setting and dates: Discussed earlier.

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of applicability rating:

Reviewer 1:

Significant dataset overlaps and we cannot be sure of the demographics of each dataset, as these are not discussed to be able to assess how applicable this model would be.

Reviewer 2:

The dataset is at risk of reduced applicability given that not all patients will have had an

independent confirmation of their COVID infection.

O. (49) **D 2020**

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: Combination of: (1. COHEN), (7. CHEST X-RAY14). A random subsample of the (7. CHEST X-RAY14) was used for control images.			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**
Rationale of bias rating: ***Reviewer 1:*** The use of public datasets with no clear COVID-19 verification process and no clear inclusion/exclusion criteria introduces a high risk of bias. Random subsampling is not reproducible. ***Reviewer 2:*** It is not clear how reliable the Cohen dataset really is. Considering the images were gathered from publications about COVID-19, severe cases may be overrepresented (especially in comparison to asymptomatic cases).			
B. Applicability			
Describe included participants, setting and dates: Discussed earlier.			
Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of applicability rating:

Reviewer 1:

We cannot be sure of the demographics of each dataset, as these are not discussed to be able to assess how applicable this model would be.

Reviewer 2:

CXR is used in hospitals for assessment of COVID-19 patients, but the demographics of the included participants are unknown.

P. (47) **Elaziz 2020**

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: Combination of: (1. COHEN), (5. CHOWDHURY). Note that the dataset (1. COHEN) is a subset of (5. CHOWDHURY).			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

High risk of bias for (5. CHOWDHURY) dataset alone, see discussion in **(I.** (14) **Bassi 2020).**

Reviewer 2:

This paper inherits the biases of the (1. COHEN) dataset and includes overlapping datasets.

B. Applicability			
Describe included participants, setting and dates: Discussed earlier.			
Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**
Rationale of applicability rating: ***Reviewer 1:*** See discussion in **(I.** (14) **Bassi 2020).** ***Reviewer 2:*** High risk of issues over the applicability, see (5. CHOWDHURY) and (1. COHEN).			

Q. (56) **Ghosh 2020**

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: Uses the dataset of the Italian Society of Medical and Interventional Radiology **(10. SIRM)** repository. It has associated annotations for ground glass opacity regions, consolidation and pleural effusion.			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**
Rationale of bias rating:			

Reviewer 1:

There is no verification for patients being COVID-19 positive and as anyone can contribute to the repository there are no inclusion / exclusion criteria. This introduces a high risk of bias.

Reviewer 2:

There is no control over which cases are included in the SIRM database. There is also no confirmation that they have been diagnosed with COVID-19 using a PCR test. Therefore for 1.2 the best we can say is that the risk is unclear. Therefore, there is a high risk of bias in this model.

B. Applicability

Describe included participants, setting and dates: Discussed earlier.

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of applicability rating:

Reviewer 1:

We cannot know the demographics and whether patients are truly COVID-19 positive, therefore have a high concern over how applicable a model would be.

Reviewer 2:

As above.

R. (26) **Han 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

This paper considers 230 CT scans for 79 COVID-19 patients, 130 CT scans for 130 pneumonia patients and 130 CT scans for 130 'people without pneumonia'. The dataset was collected from the 'designated COVID-19 hospitals in Shandong Province'.

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

We do not know which hospitals this data was retrieved from, nor do we know the patients in the 'people without pneumonia' category.

Reviewer 2:

Unclear what the inclusion and exclusion criteria are, except that CT scans without image manifestations were excluded. This has potential bias especially towards detection of only severe disease.

B. Applicability

Describe included participants, setting and dates: Discussed earlier.

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of applicability rating:

Reviewer 1:

We cannot know the demographics and the diseases in the control group so high risk this is not applicable.

Reviewer 2:

As above.

S. (23) **Rahaman 2020**

Same as **(B.** (5) **Ghoshal 2020).**

T. (24) **RulaAmer 2020**

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: (1. COHEN), (2. RSNA), FIGURE1. **FIGURE1**: COVID-19 X-ray images submitted to the GitHub repository https://github.com/agchung/Figure1-COVID-chestxray-dataset via the upload portal https://figure1.typeform.com/to/lLrHwv. These are in JPEG format.			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**
Rationale of bias rating: ***Reviewer 1:*** The public repositories have no verification for COVID-19 cases and they are going to be subject to any biases in the images uploaded by the contributors to these datasets. ***Reviewer 2:*** CXR is used in hospitals for assessment of COVID-19 patients, but the demographics of the included participants are unknown.			
B. Applicability			
Describe included participants, setting and dates: Discussed earlier.			
Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**
Rationale of applicability rating:			

Reviewer 1:

We cannot know the demographics of these patients from the public datasets.

Reviewer 2:

Demographics and sources are unknown for conclusions on applicability.

U. (27) **Shah 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

This paper uses the dataset of Zhao, Jinyu and Zhang, Yichen and He, Xuehai and Xie, Pengtao, "COVID-CT-Dataset: a CT scan dataset about COVID-19", https://arxiv.org/abs/2003.13865 with dataset at https://github.com/UCSD-AI4H/COVID-CT

Dataset (1) is assembled using images obtained from papers uploaded to medRxiv, bioRxiv and others. The caption of the image is then used to judge whether the image relates to a COVID-19 or other diagnosis. The dataset used in the paper consists of CT scans for 347 COVID-19 patients along with 397 non-COVID-19 patients with different pathologies. Contributions are made to this dataset by emailing the authors with the images and associated metadata [It currently has 249 CT images from COVID-19 patients. There are 216 patients in the current dataset. It is therefore unclear what the data was at the time of this study.]

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

The dataset is subject to huge potential biases being extracted from papers. We cannot know for the public data whether a positive PCR test has been received for the COVID-19 positive cases, similarly for other diagnoses which have not been independently confirmed.

Reviewer 2:

This dataset is subject to very significant bias as there is no control over the cases included in the dataset; they are selected by a variety of paper authors to include in their papers.

B. Applicability

Describe included participants, setting and dates: Discussed earlier.

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of applicability rating:

Reviewer 1:

We cannot know the demographics of these patients from this dataset and the images appearing in publications will typically be biased for interesting or unusual cases.

Reviewer 2:

As above; there is no control over the images used, and they may well not be representative of the population of COVID-19 patients and non-COVID-19 patients.

V. (25) **Tsiknakis 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection: (1. COHEN), (2. RSNA), (3. KERMANY), QUIBIM, BSTI, RADIOGYAN.

QUIBIM: This dataset is an initiative for collecting radiological imaging data from various centers in Europe. Note: this includes (1. COHEN) already.

BSTI: This is a repository of the British Society of Thoracic Imaging (https://bsticovid19.cimar.co.uk/worklist/?embedded=) containing CXR and CT imaging along with some unstructured metadata.

RADIOGYAN: This is not cited in the paper and it is unclear what it refers to.

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

There are many biases using QUIBIM which already includes (1. COHEN), along with using the paediatric (3. KERMANY) patients as a control group. We also have no verification that patients have are COVID-19 confirmed. Finally, with unreferenced dataset RADIOGYAN we cannot conclude on the bias introduced by this, but must assume it is high.

Reviewer 2:

The QUIBIM dataset is mentioned separately from the Cohen dataset, despite the former including images derived from the Cohen dataset. Therefore, it is unclear whether the COVID-19 images are unique. Furthermore, the dataset 'Radiogyan' which is mentioned in the paper is not cited and therefore the risk of bias cannot be evaluated for this dataset. This also prevents any demographic comparisons between classes.

B. Applicability

Describe included participants, setting and dates: Discussed earlier.

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of applicability rating:

Reviewer 1:

We cannot know the demographics of these patients from this dataset.

Reviewer 2:

Use of datasets with overlapping images without explicit acknowledgement, as well as lack of citation for the 'Radiogyan' dataset prevents judgement of applicability for this model.

Update to 3 October 2020

W. (59) **Tamal 2020**

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: (1. COHEN), (3. KERMANY) and (14. JSRT)			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(**low**/ high/ unclear)*	**HIGH**	**-**
Rationale of bias rating: ***Reviewer 1:*** The bias is inherited from the issues with (1. COHEN) and (3. KERMANY). Use of paediatric patients is inappropriate and the public repositories have no positive COVID-19 verification. ***Reviewer 2:*** Overall high risk of bias associated with the use of Cohen and Kermany datasets, particularly regarding the inclusion of paediatric patients.			
B. Applicability			

Describe included participants, setting and dates: Discussed earlier.			
Concern that the included participants and setting do not match the review question	**CONCERN:** ***(low**/ high/ unclear)*	**HIGH**	**-**
Rationale of applicability rating: ***Reviewer 1:*** As discussed above, inappropriate use of paediatric patients limits the applicability. ***Reviewer 2:*** Inclusion of images from paediatric patients limits application.			

X. (30) **Zhang 2020**

See **B. Ghosal 2020**.

Y. (31) **Bararia 2020**

See **B. Ghosal 2020**.

Z. (32) **Wang 2020**

See **K. Li 2020**

AA. (37) **MikhailGoncharov 2020**

DOMAIN 1: Participants
A. Risk of Bias
Describe the sources of data and criteria for participant selection: Several public datasets: • For lung segmentation: LUNA16: "LUNA16 Jacobs et al. (2016) is a public dataset for cancerous lung nodules segmentation. It includes 888 annotated 3D thoracic CT scans from the LIDC/IDRI database Armato III et al. (2011). Scans widely differ by scanner manufacturers (17 scanner models), slice

thicknesses (from 0.6 to 5.0 mm), in-plane pixel resolution (from 0.461 to 0.977 mm), and other parameters. Annotations for every image contain binary masks for the left and right lungs, the trachea and main stem bronchi, and the cancerous nodules. The lung and trachea masks were originally obtained using an automatic algorithm van Rikxoort et al. (2009) and the lung nodules were annotated by 4 radiologists Armato III et al. (2011). During the preliminary experiments we excluded 7 cases with absent or completely broken lung masks and extremely noisy scans."

- For lung segmentation and triage model: NSCLC: "NSCLC-Radiomics dataset Kiser et al. (2020); Aerts et al. (2015) represents a subset of The Cancer Imaging Archive NSCLC Radiomics collection Clark et al. (2013). It contains left and right lungs segmentations annotated on 3D thoracic CT series of 402 patients with diseased lungs. Pathologies — tumors, atelectasis, and effusion — are included in the lungs volumes masks. Pleural effusion is also delineated separately, when present. However, we used only united lungs binary masks in our experiments."
- For triage model: MedSeg-29: "MedSeg web-site3 shares 2 publicly available datasets of annotated volumetric CT images. The first dataset consists of 9 volumetric CT scans from here4 that were be converted from JPG to Nifti format. The annotations of this dataset include lung masks and COVID-19 masks segmented by a radiologist. The second dataset consists of 20 volumetric CT scans shared by Jun et al. (2020). The left and rights lungs, and infections are labeled by two radiologists and verified by an experienced radiologist."
- For triage model: Mosmed-1110: "1110 CT scans from Moscow outpatient clinics were collected from 1st of March, 2020 to 25th of April, 2020, within the framework of outpatient computed tomography centers in Moscow, Russia Morozov et al. (2020a). Scans were performed on Canon (Toshiba) Aquilion 64 units in with standard scanner protocols and, particularly 0.8 mm inter-slice distance. However, the public version of the dataset contains every 10th slice of the original study, so the effective inter-slice distance is 8mm.
  The quantification of COVID-19 severity in CT was performed with the visual semi-quantitative scale adopted in the Russian Federation and Moscow in particular Morozov et al. (2020d). According to this grading the dataset contains 254 images without COVID-19 symptoms. The rest is split into 4 categories: CT1 (afected lung percentage 25% or below, 684 images), CT2 (from 25% to 50%, 125 images), CT3 (from 50% to 75%, 45 images), CT4 (75% and above, 2 images).
  Radiologists performed an initial reading of CT scans in clinics, after which experts from the advisory department of Center for Diagnostics and Telemedicine (CDT) independently conducted the second reading as a part of total audit targeting all CT studies with suspected COVID-19."
- Mosmed-Lung-Cancer-500: for test set: "In a public dataset Morozov et al. (2020b) containing 500 chest CT scans randomly selected from patients over 50 years of age, 63 CT scans were found to have no annotated nodules. After the second reading, 36 patients with pathological conditions not corresponding to the lung nodules (segmental and lobar pneumonia, lung atelectasis) were found. The remaining 27 studies were without pathological changes in the lungs."
- Mosmed-20: for test set: "It is a dataset Morozov et al. (2020c) of 42 CT studies collected from 20 patients in a infectious diseases hospital during the second half of February 2020, at the beginning of Russian outbreak. We removed 4 cases with movement artifacts. The remaining 38 cases were split into healthy (5) and COVID-19 (27 CT1, 2 CT2, 3 CT3, 1 CT4) cases. As we see, at the beginning of the outbreak the majority of cases have mild severity and the resulted structure represents a typical out-patient clinic during the pandemic.
  It also important to note that Mosmed-1100 was collected from a cloud PACS which connects all Moscow out-patient clinics. In-patient clinincs are not connected to this PACS. Finally, Mosmed-1100 collection were initiated 1-2 weeks after collection of Mosmed-20, so studies duplication is almost impossible."

The Mosmed-20 dataset no longer seems to be available. However, the Mosmed-1100 collection is. It says in the overview (README_EN_2.pdf):

"Please note: this distribution has been made based on radiologic findings only, neither on polymerase chain reaction (PCR) test results or clinical verification."

This does not affect this PROBAST domain but will affect others.

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

These public datasets may well be representative, but with the exception of Mosmed-Lung-Cancer-500, there is no indication that these are randomly-selected patients or a cohort or so on. Little further selection has been performed.

Reviewer 2:

We do not know the inclusion/exclusion criteria for these. Also only pre-processed scans provided with every tenth slice retained. Not PCR confirmed so not a verified COVID-19 positive cohort.

B. Applicability

Describe included participants, setting and dates: See above.

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**UNCLEAR**	**-**

Rationale of applicability rating:

Reviewer 1:

The test sets are solely from Russia, and there is the possibility that the trained model may not be more widely applicable. But the aim of the paper is to show that this methodology works when tested on a cohort matching the cohort on which the model is trained.

Reviewer 2:

We do not know the inclusion/exclusion criteria so cannot appreciate how widely applicable this is.

BB. (62) **Yip 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

MOSMEDDATA dataset. Data obtained between 1st March 2020 and 25th April 2020 provided by municipal hospitals in Moscow, Russia." "1110 CT scans from unique patients – 42% male, 56% female, median age = 47 (18-97)." ""Please note: this distribution has been made based on radiologic findings only, neither on polymerase chain reaction (PCR) test results or clinical verification."

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

Unclear how patients were identified as COVID-19 positive, unclear how patients were selected, and unclear how many patients were contributed by each hospital.

Reviewer 2:

Without PCR and no clear inclusion/exclusion criteria this is at high risk of bias. Patient selection on the basis of imaging alone introduces a high risk of bias.

B. Applicability

Describe included participants, setting and dates: Discussed earlier.

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of applicability rating:

Reviewer 1:

Very limited information on this dataset hinders application.

Reviewer 2:

We cannot know what the inclusion/exclusion criteria are and so the applicability is at high risk of bias.

CC. (36) **MuhammadFarooq 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

"We used the COVIDx dataset that was recently made public by the authors of the COVID-Net [7]. It consists of a total of 5941 posteroanterior chest radiography images from 2839 patients with 4 classes namely 1) Normal (no infections), 2) Bacterial (bacterial pneumonia) 3) Viral (non-COVID-19 viral pneumonia) 4) COVID-19. The dataset was curated by combining two publicly available datasets. The authors have made a pre-processed version of the dataset available at https://github.com/lindawangg/COVID-Net.

In the current version of the dataset, there are 68 COVID-19 radiographs from 45 COVID-19 patients. There were a total of 1203 patients with negative pneumonia (i.e. Normal class), 931 patients with a bacterial pneumonia and 660 patients with nonCOVID-19 viral pneumonia cases."

COVIDx dataset: compilation of data from COHEN (1), RSNA (2) and CHOWDHURY (5)

See #8 in the Central Dataset Review (COVIDX)

		Dev	Val
1.3 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.4 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

These images were chosen from those available to the dataset collators, and may well represent "interesting" or "extreme" cases rather than "typical" cases.

Also, the normal and bacterial datasets include paediatric cases.

Reviewer 2:

Bias is inherited from COHEN, RSNA and CHOWDHURY datasets, such as inappropriate incorporation of paediatric patients.

B. Applicability

Describe included participants, setting and dates:

See above.

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of applicability rating:

Reviewer 1:

It is unclear how generally applicable these images will be when in a hospital setting with a wider variety of cases. The bias inherent in the included participants also inhibits its ability to transfer beyond.

Reviewer 2:

High risk of bias as stated above limits applicability

SECTION 2. Papers which use private datasets

In this section we perform the Domain 1. Participants PROBAST for the following papers that use private datasets:

Update to 24 June 2020

DD. (2) Amyar 2020
EE. (3) Ardakani 2020
FF. (4) Bai 2020
GG. (45) Georgescu 2020
HH. (38) Chassagnon 2020
II. (39) Chen 2020

JJ. (40) Shi 2020
KK. (41) Guiot 2020
LL. (10) Jin 2020
MM. (11) Ko 2020
NN. (44) Lassau 2020
OO. (12) Mei 2020
PP. (15) Pu 2020
QQ. (42) Qi/Yue 2020
RR. (16) Wang 2020
SS. (17) Wang 2020
TT. (43) Zhu 2020

Update to 14 August 2020

UU. (55) Chen 2020
VV. (48) HanqingChao 2020
WW. (50) Qin 2020
XX. (52) Wei 2020
YY. (53) Wu 2020
ZZ. (54) Zheng 2020

Update to 3 October 2020

AAA. (57) Schalekamp 2020
BBB. (28) Li 2020
CCC. (58) Xie 2020
DDD. (34) Li 2020
EEE. (29) Wang 2020
FFF. (33) Zhang 2020
GGG. (35) Wang 2020
HHH. (60) Wang 2020
III. (61) Xu 2020
JJJ. (51) Ramtohul 2020

SECTION 2.1. Individual paper reviews

Update to 24 June 2020

DD. (2) **Amyar 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection: This paper uses three datasets:

(1) *Zhao, Jinyu and Zhang, Yichen and He, Xuehai and Xie, Pengtao, "COVID-CT-Dataset: a CT scan dataset about COVID-19", https://arxiv.org/abs/2003.13865 with dataset at https://github.com/UCSD-AI4H/COVID-CT*

(2) *http://medicalsegmentation.com/covid19/*

(3) *This dataset is not cited within the paper but reference is made "the hospital "Henri Becquerel Center" in Rouen city of France includes 100 CT of normal patients and 98 of lung cancer."*

Dataset (1) is assembled using images obtained from papers uploaded to medRxiv, bioRxiv and others. The caption of the image is then used to judge whether the image relates to a COVID-19 or other diagnosis. The dataset used in the paper consists of CT scans for 347 COVID-19 patients along with 397 non-COVID-19 patients with different pathologies. Contributions are made to this dataset by emailing the authors with the images and associated metadata [It currently has 249 CT images from COVID-19 patients. There are 216 patients in the current dataset. It is therefore unclear what the data was at the time of this study.]

Dataset (2) consists of 100 axial slices of CT scans from ">40 patients with COVID-19" that were taken from the Italian Society of Medical and Interventional Radiology (SIRM, https://www.sirm.org/category/senza-categoria/covid-19/) repository. It has associated annotations for ground glass opacity regions, consolidation and pleural effusion.

Dataset (3) is unreferenced and it is not possible to conclude what patients are in this dataset.

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

Datasets (1) and (2) are subject to huge potential biases being extracted from papers and being a

subset of those submitted to the SIRM repository. Dataset (3) is undocumented in the paper and they suggest it relates to cancer patients which would be an unexpected control group for COVID-19 diagnosis.

We cannot know for the public data whether a positive PCR test has been received for the COVID-19 positive cases, similarly for other diagnoses which have not been independently confirmed. There is a wide distribution of countries from which data has been sourced and the dataset is heterogeneous for pathologies. But underlying biases may exist.

Reviewer 2:

Sources of data in (1) are not specified; demographics are not specified; (1) and (2) may be biased in favour of clear COVID cases. No criteria for inclusion specified. Though not relevant to these questions, source of COVID-19 confirmation is not specified.

B. Applicability

Describe included participants, setting and dates: See section A

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of applicability rating:

Reviewer 1:

For all previously described reasons. We have no knowledge of the patient population for this paper or demographic makeup.

Reviewer 2:

Paper claims in the conclusion: "We have shown also that we can obtain very good sensitivity from CT images, which can tackle the need to detect infected people at an early stage to isolate them, and therefore, to limit the spreading of the disease." But we don't know what stage of illness the training data comes from, so this generalisation is not justified. (And this hope is vaguely matched in the abstract.)

EE. (3) **Ardakani 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection: This paper uses one dataset.

<u>For the COVID-19 patients</u>, "the entirety of patients representing flu-like symptoms with an initial diagnosis of the novel coronavirus, regardless of demographic values such as age and sex, were included in the study. Prior to enrolment, chest high-Resolution CT (HRCT) examination was obtained from all patients during the acute phase of the disease." "The inclusion criterion was the confirmation of diagnosis of COVID19 through RT-PCR performed on nasopharyngeal swabs samples. Patients with concurrent pulmonary infections, as confirmed by laboratory tests and negative RT-PCR, were excluded. In addition, patients with CT imaging suggestive of chronic lung diseases and subsequent pulmonary involvement were excluded. Imaging studies were performed between 3 and 6 days from the onset of flu-like symptoms."

<u>For the non-COVID patients</u> "we retrospectively analyzed the HRCT images of patients from September 2019 to December 2019 with other causes of atypical and viral pneumonia as adenoviral or H1N1 flu from the PACS of our university hospital"

The final dataset included 108 COVID-19 patients and 86 non-COVID patients. One presumes that the COVID-19 patients were from the same university hospital, but this is not explicitly stated.

There is a significant age imbalance in the dataset with the age of the COVID-19 patients aged 50.22±10.85 years whereas the non-COVID-19 patients are aged 61.45±15.04 years [p < 0.001].

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		UNCLEAR	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

<u>Reviewer 1:</u>

There is a significant age imbalance in the dataset with the age of the COVID-19 patients aged 50.22±10.85 years whereas the non-COVID-19 patients are aged 61.45±15.04 years. This bias can lead to misleading conclusions that age and outcome are linked.

Reviewer 2:

The whole cohort is from one centre for COVID, but unclear how patients selected for non-COVID.

B. Applicability

Describe included participants, setting and dates: Discussed above.

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**UNCLEAR**	**-**

Rationale of applicability rating:

Reviewer 1:

As discussed above, with such a large age imbalance in the dataset this is likely not applicable to a more general patient cohort.

Reviewer 2:

The review question was whether CNNs could be used to distinguish COVID-19 from non-COVID-19 cases, and this seems to be demonstrated for this hospital. They do not claim that their particular training can be used for this purpose.

FF. (4) **Bai 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection: The data is sourced from (1) Rhode Island Hospital, (2) the Hospital of the University of Pennsylvania, (3) Xiangya Hospital and (4) nine hospitals in Hunan, China [names in Table E1].

Sources (1)—(3) provided the pneumonia control chest CT scans with (1) and (4) providing the PCR confirmed COVID-19 cases. The dataset consists of 665 pneumonia cases and 521 COVID-19 CT scans. The data from (2) and 3 of the 9 hospitals from (4) were used as holdout cohorts for independent testing.

Figure 1 specifies the following:

Patients with COVID-19 definitely diagnosed by RT-PCR and available CT from RIH and 9 hospitals in Hunan Province, China (n = 699) -> Excluded patients (n = 178): Patients with no abnormal finding

on chest CT scans -> COVID-19 chest CT scans with abnormal finding (n = 521).			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

There are two immediate significant potential sources of bias. Firstly, the patients in the COVID-19 cohort are aged 46±16 years whereas the non-COVID cohort has patients aged 62±19 years. This is a bias which an algorithm could link to the outcome. Secondly, the Chinese hospitals have only provided COVID-19 data and this is again a potential source of bias for an algorithm to associate Chinese patients to COVID-19 outcome. Almost all Western data was for non-COVID patients.

Reviewer 2:

The patients were from different centres in different continents; they were a complete cohort from these centres and were appropriate for the research question. However, the majority of COVID-19 patients were from the Chinese hospitals and the majority of the non-COVID cases were from the American hospitals. This makes the dataset very biased.

B. Applicability

Describe included participants, setting and dates:

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of applicability rating:

Reviewer 1:

For the reasons previously described there is a high concern over the applicability of an algorithm developed using these datasets. It is unclear for example how this algorithm would perform on a cohort of non-COVID Chinese patients or whether it has leaned that younger patients are likely to be COVID-19 based on the data biases.

Reviewer 2:

The exclusion of COVID-19 cases without abnormalities on the CT scans is appropriate for the question of whether ML models can assist radiologists to distinguish between COVID-19 and non-COVID pneumonias that are visible on CT scans. Nevertheless, the earlier discussed issues mean the claims cannot be confidently asserted.

GG. (45) **Georgescu 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection: This paper uses one dataset.

This dataset was collected by the paper authors from 16 unspecified different centres in North America and Europe. The dataset consists of CT scans of 1150 patients who were positive for covid-19, 159 patients with non-covid-19 pneumonia, 177 patients with interstitial lung disease and 610 patients without any pathology. Images are in DICOM format and radiology readings. COVID-19 diagnosis confirmed either by RT-PCR test or diagnosed from "clinical symptoms, epidemiological exposure and radiological assessment". The ILD cohort consists of patients with various types of ILD exhibiting ground glass opacities, reticulation, honeycombing and consolidation to different degrees. Data does not seem to be publicly available. No details on participant selection given.

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		UNCLEAR	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

Whilst the COVID-19 cohort from North America have been confirmed by an RT-PCR test, the COVID-19 cohort from European has been either confirmed by an RT-PCR test or diagnosed based on clinical symptoms, epidemiological exposure and radiological assessment. It is unclear if based on "clinical symptoms" is a good measure for assign positive cases. Therefore, bias might exist.

Reviewer 2:

There was no breakdown of illness for each centre so it's unclear if there are biases due to one type of illness being from only one centre. Also, possibly bias due to ground truth measure in covid-19.

B. Applicability

Describe included participants, setting and dates: ***Range of illnesses, proportion of male:female, median and age range, and other details given for train/validation/test sets and broken down by illness.***

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**UNCLEAR**	**-**

Rationale of applicability rating:

Reviewer 1:

Although the authors indeed provide a table with the details on the demographic of their dataset, there are some missing information. Therefore, it is difficult to assess how applicable this model would be.

Reviewer 2:

Good points: range of illnesses in training/test data set and patient demographics broken down by illness and train/validate/test dataset, bad points: unclear if there are bias due to different illnesses being from different centres.

HH. (38) **Chassagnon 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection: Six hospitals [undisclosed]

Patients diagnosed with COVID-19 from March 4th to April 5th from eight large University Hospitals

were eligible if they had positive reverse transcription polymerase chain reaction (PCR-RT) and signs of COVID-19 pneumonia on unenhanced chest CT. A total of 693 patients formed the full dataset (321,360 CT slices). Only the CT examination performed at initial evaluation was included. Exclusion criteria were 1/ contrast medium injection and 2/ important motion artifacts. No patient was

intubated at the time of the CT acquisition.

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		LOW	-
1.2 Were all inclusions and exclusions of participants appropriate?		UNCLEAR	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**UNCLEAR**	**-**

Rationale of bias rating:

Reviewer 1:

Assembly of the dataset seems to be reasonable but potentially a bias due to requirement of both imaging and positive PCR.

Reviewer 2:

*There is a slight risk of bias from including only patients who were positive on PCR **AND** had imaging findings may have introduced bias.*

B. Applicability

*Describe included participants, setting and dates: **See above (A)***

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**LOW**	**-**

Rationale of applicability rating:

Reviewer 1:

No issues for applicability noted.

Reviewer 2:

Seems applicable to the papers' target population.

II. (39) **Chen 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

COVID-19 data

From January 1 to February 8, 2020, ***seventy consecutive patients*** *with COVID-19 admitted in 5 independent hospitals from 4 cities were enrolled in this study (mean age, 42.9 years; range, 16–69 years), including 41 men (mean age, 41.8 years; range, 16–69 years) and 29 women (mean age, 44.5 years; range, 16–66 years). All patients were confirmed with SARS-CoV-2 infection by real-time RT-PCR and next-generation sequencing.*

Of these patients:

- *24 were from Huizhou City*
- *25 from Shantou City*
- *15 from Yongzhou City*
- *6 from Meizhou City.*

Non-COVID-19 data

At the same period, another 66 pneumonia patients without COVID-19 from Meizhou People's Hospital were recruited as controls (mean age, 46.7 years; range, 0.3–93 years), including 43 men (mean age, 46.0 years; range, 0.3–93 years) and 23 women (mean age, 48.0 years; range, 1–86 years). All the controls were confirmed with consecutive negative RT-PCR assays.

Training data: *51 COVID-19 patients from Huizhou, Yongzhou, and Shantou cities and 47 controls from Meizhou City.*

Validation data: *A total of 19 COVID-19 patients from two hospitals (6 patients from Meizhou People's Hospital and 13 patients from the First Affiliated Hospital of Shantou University Medical College) and 19 randomly selected controls from Meizhou City were incorporated into the validation cohort.*

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		UNCLEAR	-
Risk of bias introduced by selection of participants	**RISK:**	**HIGH**	**-**

	(low/ high/ unclear)		

Rationale of bias rating:

Reviewer 1:

Patients selection is reasonable and no significant demographic biases between the training and validation cohorts. A risk would be that the pneumonia patients are all taken from one center and therefore the algorithm may learn to associate the center to the outcome.

Reviewer 2:

There could be potential concern that the negative controls were false negatives. As it is unclear how many consecutive PCR tests were carried out, it is difficult to judge the risk. I was hesitating between a low and an unclear rating.

B. Applicability

Describe included participants, setting and dates: ***As above (A)***

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**UNCLEAR**	**-**

Rationale of applicability rating:

Reviewer 1:

The pneumonia patients are from a single center and the sample size is very small so it is very uncertain whether this approach would generalise to a much wider population and where pneumonia patients are from different centers.

Reviewer 2:

Pneumonia only from a single center.

JJ. (40) **Feng Shi 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

CT images of a total of 2685 participants were retrospectively collected. Three hospitals were involved, including Tongji Hospital of Huazhong University of Science and Technology, Shanghai Public Health Clinical Center of Fudan University, and China-Japan Union Hospital of Jilin University.

COVID-19 data

In this dataset, 1658 cases were the confirmed COVID-19 cases diagnosed by positive nucleic acid testing with conformation by national CDC. The age of COVID-19 subjects is 49±14 years.

Non-COVID-19 data

The other 1027 cases were community acquired pneumonia (CAP) patients. The ages of the CAP subjects are 56±14 years.

Other demographic data for each cohort is not given. It is not described which hospital provided each dataset and how many of each cohort.

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		UNCLEAR	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

It is unclear the quantity of scans each hospital has provided and the underlying biases in this. There is a significant age difference between the COVID-19 and non-COVID-19 patient cohorts and therefore the algorithm could associate age with outcome.

Reviewer 2:

How were CAP patients diagnosed? I assume from imaging only. How were patients selected? No inclusion/exclusion criteria provided. I would rate this a high risk of bias.

B. Applicability

Describe included participants, setting and dates: ***As described in A.***			
Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**
Rationale of applicability rating: ***Reviewer 1:*** It is unclear whether the algorithm has associated the ages to the outcome. With no knowledge of how many scans were provided by each hospital we cannot know whether each cohort of data came from multiple centers. ***Reviewer 2:*** Because we do not know the source hospital for each example, it is unclear if the validation data comes from a different source than the training data. Validating a model on internally rather than externally collected data can significantly affect results, limiting this model's applicability.			

KK. (41) **Guiot 2020**

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: *We analysed the data from 181 RT-PCR confirmed COVID-19 patients as well as 1200 other non-COVID-19 control patients to build and assess the performance of the model. The datasets were collected from 2 different hospital sites of the CHU Liège, Belgium.* *In the COVID-19 cohort, patient age distributions are 64.4 ± 15.8 and in the non-COVID group the age distribution is* 63.8 ± 14.4. It is not detailed how many images were provided from each of the hospitals.			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		UNCLEAR	-
Risk of bias introduced by selection of participants	**RISK:**	**HIGH**	**-**

	(low/ high/ unclear)		

Rationale of bias rating:

Reviewer 1:

As described above, there is no discussion of which center the COVID and non-COVID images from. If the COVID dataset is from one center then it is possible for the algorithm to learn to associate center with outcome.

Reviewer 2:

The non-COVID patients were "consecutive patients" (do not know what that means) that underwent chest CT imaging. It is unclear how these patients were selected for chest imaging, but it is reasonable to assume that these patients are representative of the population suspected of having COVID, which is the authors' target population. The two classes are "COVID" and "not COVID"; the class "not COVID" is ill-defined, I would have put "high bias" in uniformity of outcomes in the PROBAST checklist due to this quality in the dataset. However, I think it is reasonable to say that these data are at a relatively low risk of bias.

B. Applicability

Describe included participants, setting and dates: As discussed above.

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of applicability rating:

Reviewer 1:

As we do not know from the paper where the datasets were obtained, we cannot conclude how applicable the models developed will be on new cohorts. There is a high concern that they would not be applicable.

Reviewer 2:

Because we do not know which examples were sourced from which of the two hospitals, we cannot determine whether the model was validated externally or internally. This limits the model's applicability.

LL. (10) **Jin 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

There are a total of 877 COVID-19 images in the dataset from 850 unique patients. There are 541 non-COVID-19 images from 541 unique patients. The non-COVID images are primarily pneumonia (252), "old lesions" (103), healthy (91) and "tumour" (44).

The data was acquired from 5 hospitals in China:

- Zhongnan Hospital of Wuhan University; 141 (115) COVID, 116 (116) non-COVID
- Wuhan's Leishenshan Hospital; 165 (165) COVID, 0 (0) non-COVID
- Beijing Tsinghua Changgung Hospital; 5 (5) COVID, 284 (284) non-COVID
- Wuhan No. 7 Hospital; 188 (187) COVID, 141 (141) non-COVID
- Tianyou Hospital Affiliated to Wuhan University of Science & Technology; 378 (378) COVID, 0 (0) non-COVID.

Our positive samples were all collected from confirmed patients, following China's national diagnostic and treatment guidelines at the time of the diagnosis, which required positive results in nucleic acid test.

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		LOW	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

Three of the centres have provided imbalanced datasets, specifically Wuhan's Leishenshan Hospital and Tianyou Hospital Affiliated to Wuhan University of Science & Technology only provided COVID-19 patient data whereas Beijing Tsinghua Changgung Hospital provided data which was almost all non-COVID-19 (284 non COVID-19 vs. 5 COVID-19 images). Therefore this bias could lead to an association between data source and outcome.

Reviewer 2:

There is a large and balanced number of COVID-19 and non-COVID-19 images, with the majority from unique patients. There is a risk of source bias, since half the COVID-19 cases came from two hospitals where no non-COVID-19 cases were used. This is further compounded by the use of different models of CT equipment, which increases the likelihood of source bias.

B. Applicability

Describe included participants, setting and dates:

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of applicability rating:

Reviewer 1:

It Is unclear how this algorithm would perform on patients at the centers for which imbalanced data was provided, and whether if provided with an image from the minority class, it would simply predict the dominant class.

Reviewer 2:

Risk of source bias as described above.

MM. (11) **Ko 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection: WKUH, CNUH, **(10. SIRM)** and ZHAO.

Authors combined two sources their own dataset (coming from 2 different regions Wonkwang University Hospital (WKUH) and Chonnam National University Hospital (CNUH).) and the publicly available dataset from the Italian Society of Medical and Interventional Radiology (SIRM) public database, which characteristic are as follows:

[Their own dataset] For the COVID-19 data group, we used a total of 1,194 chest CT images: 673 chest CT images (13 patients) from CNUH, 421 images (7 patients) from WKUH, and 100 images (60 patients) from the Italian Society of Medical and Interventional Radiology (SIRM) public database.

The 20 patients from CNUH and WKUH included 9 males and 11 females, with an average age of 59.6±17.2 years. Regarding the COVID-19 data from WKUH and CNUH, all of the COVID-19 patients were confirmed as positive for the virus by RT-PCR viral detection and were acquired between December 31, 2019 and March 25, 2020. The median period from symptom onset to the first chest CT exam was 8 days, with the range from 2 to 20 days.

+ 1 external dataset for testing:

The dataset was acquired from January 19 and March 25, 2020, 264 low-quality chest CT images (LQI)

were used as additional testing data.

(I. COVID-CT-Dataset) This dataset was collected from Jan 19th to Mar 25th. For COVID-19 CT data, they obtain 349 CT images labelled as being positive for COVID-19. These CT images have different sizes. The minimum, average, and maximum height are 153, 491 and 1853. The minimum, average, and maximum width are 124, 383, and 1485. These images are from 216 patient cases. Whilst for the non-COVID-19 cases, they collected a set of non-COVID-19 CT images as negative training examples from four sources see S1, S2, S3 and S4 below (695 images).

S1 - https://medpix.nlm.nih.gov/home

S2 - https://luna16.grand-challenge.org/

S3 - https://radiopaedia.org/articles/covid-19-3

S4 - https://www.ncbi.nlm.nih.gov/pmc/

This dataset can be found in:

Link: https://github.com/UCSD-AI4H/COVID-CT

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		UNCLEAR	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**
Rationale of bias rating:			

Reviewer 1:

"The data from WKUH, CNUH, and SIRM were randomly split". They did not use the official partition from SIRM, so it is unclear if the exclusion was performed followed a correct protocol e.g. at patient level. Therefore, bias might exist.

WKUH and CNUH, all of the COVID-19 patients were confirmed as positive for the virus by RT-PCR. However, the SIRM is not clear how the COVID-19 positive cases were confirmed

Reviewer 2:

Demographics given, data from multiple sources with a mixture of illnesses spread across the centres. SIRM dataset doesn't specify how covid-19 was confirmed.

B. Applicability

Describe included participants, setting and dates: ***see description from A***

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**LOW**	**-**

Rationale of applicability rating:

Reviewer 1:

It is one of the few papers that offer a strong demographic of their dataset.

Reviewer 2:

Basic demographics given.

NN.(44) **Lassau 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

Data were collected at two French hospitals (Kremlin Bicêtre Hospital (KB), APHP, Paris, and Gustave

Roussy Hospital (IGR), Villejuif. There are a total of 796 patients from KB and 131 patients from IGR with imaging, clinical and biological data.

Inclusion criteria were (1) date of admission at hospital (from the 12th of February to the 20th of March at Kremin Bicêtre and from the 2nd of March to the 24th of April at Institut Gustave Roussy) and (2) a positive diagnosis of COVID-19. Patients were considered positive either because of a positive RT-PCR (real-time fluorescence polymerase chain reaction) based on nasal or lower respiratory tract specimens or a CT scan with a typical appearance of COVID-19 as defined by the ACR criteria for negative RT-PCR patients. Children and pregnant women were excluded from the study.

The models were trained using five-fold cross-validation on 646 KB patients and tested on 150 Kremlin-Bicêtre KB patients. External validation was performed on the independent Institut Gustave Roussy (IGR) dataset of 137 patients.

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		LOW	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

Performing the development on a single center data and validating on an external dataset is a reasonable approach. The demographics of each cohort are reasonably well matched.

Using two sources for the outcome variable introduces high potential bias due to the negative PCR patients being assessed using the CT scan features.

Reviewer 2:

Positive cases included those with negative PCR but with a typical appearance of COVID-19 on CT scan. Since the appearance of COVID-19 is indistinguishable from other viral pneumonias on CT, there is a low specificity and therefore a risk of false positive inclusions as part of the COVID-19 positive group.

B. Applicability			
Describe included participants, setting and dates:			
Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**
Rationale of applicability rating:			

Reviewer 1:

See discussions above. The study design does match the review question reasonably well but the issues over sources for outcome lead to potential issues for applicability.

Reviewer 2:

Concern with the inclusion criteria for the positive cases.

OO.(12) **Mei 2020**

DOMAIN 1: Participants		
A. Risk of Bias		
Describe the sources of data and criteria for participant selection:		

This dataset contains chest CT scans for 905 patients with 419 COVID-19 positive cases and 486 COVID-19 negative cases. The data was acquired between 17th Jan 2020 and 3rd March 2020 from 18 medical centers in 13 provinces in China. A total of 419 patients (46.3%) tested positive for SARS-CoV-2 by laboratory-confirmed real-time RT– PCR assay and next-generation sequencing, whereas 486 patients (53.7%) tested negative (confirmed by at least two additional negative RT–PCR tests and clinical observation).

In the COVID-19 cohort, patient age distributions are 43.0 ± 16.4 and in the non-COVID group the age distribution is 38.6 ± 16.3 (p = 0.00086). It is not detailed how many images were provided from each of the hospitals.

	Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?	HIGH	-

1.2 Were all inclusions and exclusions of participants appropriate?		LOW	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

It is unknown which centres have provided the COVID-19 and non-COVID-19 data. This leads to a potential bias and the algorithm associating data source to outcome. Combined with the statistically significant age disparity between the cohorts leads to a high risk of bias.

Reviewer 2:

There is a large number of balanced images for both COVID-19 positive and negative cases. Inclusion criteria for both positive and negative cases is low bias. However, since the number of images per hospital was not specified, it is unclear whether or not there is source bias resulting from an imbalance of positive cases or negative controls provided by each hospital.

B. Applicability

Describe included participants, setting and dates: ***Discussed above***

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**UNCLEAR**	**-**

Rationale of applicability rating:

Reviewer 1:

As above, without knowing the sources of each scan, we cannot know the wider applicability of the algorithm.

Reviewer 2:

Balanced number of positive cases and negative controls. However, there is a significant difference in age distributions between the two groups.

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

Retrospectively collected 498 CT scans from 151 subjects positive for COVID-19 by RT-PCR and chest CT imaging findings. Subjects all had either close contacts with individuals from Wuhan or had a travel history to Wuhan. Ages 45.7 ± 15.7, males 55%.

Retrospectively collected 497 CT scans acquired on different subjects with other types of pneumonias (majority caused by influenza A and B virus, human parainfluenza virus (types I, II and III), human rhinovirus and adenovirus). Ages 45.9 ± 17.7, males 52.8%

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		UNCLEAR	-
1.2 Were all inclusions and exclusions of participants appropriate?		UNCLEAR	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/* ***unclear****)*	**UNCLEAR**	**-**

Rationale of bias rating:

Reviewer 1:

Balanced number of images for both positive and negative controls. Appropriate criteria were used for selection of positive cases but unclear for negative controls. However, it is unclear how the participants were selected in both groups. Also unclear when the participants were selected.

Reviewer 2:

It is unclear, whether the cases of two classes were collected in the same or in different hospitals. Furthermore, the selection criteria for CAP-positive cases are not named.

B. Applicability

Describe included participants, setting and dates:

Concern that the included participants and setting do not match the review question	**CONCERN:** *(****low****/ high/*	**LOW**	**-**

	unclear)		

Rationale of applicability rating:

Reviewer 1:

Balanced demographics and appropriate inclusion criteria.

Reviewer 2:

Chest-CT imaging data is routinely gathered for COVID-19 suspected cases. While there are doubts about the composition of the dataset, the general setup is applicable for the posed research question.

QQ. (42) **Qi 2020**

DOMAIN 1: Participants		
A. Risk of Bias		
Describe the sources of data and criteria for participant selection: A total of 52 patients with laboratory-confirmed SARS-CoV-2 infection and their initial CT images were enrolled from 5 designated hospitals in Ankang, Lishui, Zhenjiang, Lanzhou, and Linxia between January 23, 2020 and February 8, 2020. As of February 20, patients remained in hospital or with non-findings in CT were excluded. Therefore, 31 patients with 72 lesion segments were included in the final analysis. **Development cohort:** this comprised 26 patients (12 from Ankang, 8 from Lishui, 4 from Lanzhou, and 2 from Linxia) with 59 lesion segments. **Test cohort:** this comprised 5 patients from Zhenjiang with 13 lesion segments. There are no significant demographic differences between the groups of patients with short-term and long-term stays.		
	Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control	HIGH	-

study data?			
1.2 Were all inclusions and exclusions of participants appropriate?		UNCLEAR	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

It is unclear whether (and unlikely that) the 52 patients studied are all COVID-19 patients at the 5 hospitals. Therefore, there could be a bias in how these patients have been selected. Also, the study of multiple lesions in the same patients is also a potential bias.

Reviewer 2:

Are these the only COVID-19 positive patients attending these centres between 23rd of Jan and 8th of Feb? If not, no detail on how they have been selected from a larger cohort is given.

B. Applicability

Describe included participants, setting and dates:

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of applicability rating:

Reviewer 1:

With such a small sample it is not possible to conclude on the applicability of this model. The test set of only 5 patients would give high concern, along with an unknown bias due to multiple lesions for a subset of patients.

Reviewer 2:

Only 5 patients in the test set and unknown probable subset of the COVID-19 patients.

RR. (16) **Wang 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

This dataset consists of 1,065 CT images of pathogen-confirmed COVID-19 cases (325 images) along with those previously diagnosed with typical viral pneumonia (740 images). The dataset used images was assembled from three centers:

- Center 1: Xi'an Jiaotong University First Affiliated Hospital
- Center 2: Nanchang University First Hospital
- Center 3: Xi'an No.8 Hospital of Xi'an Medical College

These were utilised in the following ways:

- 320 images (160 images from COVID-19 negative and 160 images from COVID-19 positive) from Center 1 were obtained to construct the model.
- To test the stability and generalization of the model, 455 images (360 COVID-19 negative images and COVID-19 positive 95 images) were obtained for internal validation from Center 1.
- 290 images (COVID-19 negative 220 images and COVID-19 positive 70 images) were obtained from Center 2 and 3 for external validation.

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		UNCLEAR	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

The paper does not detail the demographics of the COVID-19 and non-COVID-19 cohorts which could be a significant source of bias. There is no discussion of how the patients were selected from each of the hospitals and confusion over the number of COVID-19, this is also a potentially significant source of bias.

Reviewer 2:

There are 79 COVID-19 positive patients, + 15 patients who had a multiple negative PCR before a positive one. These same patient's scans were segmented, and then it appears slices were separated to create 160 COVID-19 positive training images, and 95 COVID-19 positive images for

testing – therefore it seems there is a risk that slices from the same patient are located in the training and testing sets the dataset is biased.

B. Applicability

Describe included participants, setting and dates: ***Discussed above***

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of applicability rating:

Reviewer 1:

For all the previously discussed reasons, and no thorough understanding of how the dataset was constructed, we must have high concern over the applicability of an algorithm developed using this dataset and how representative this dataset is of the intended population.

Reviewer 2:

Dataset is not fully explained and understood.

SS. (17) **Wang 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

In this study, two datasets are utilised:

- *COVID-19 dataset:1,266 CT scans for 1,266 COVID-19 patients from hospitals in Wuhan city, Henan, Anhui, Heilongjiang, Beijing and Huangshi city.*
- *CT-EGFR dataset: 4,106 CT scans for 4,106 lung cancer patients and associated gene information (specifically EGFR mutation status). This is used to 'pre-train' a network by predicting the EGFR mutation status. This data is from Sichuan province, with 2096 patients having EGFR mutant lung cancer.*

In the COVID-19 dataset, 1,266 patients were finally included who met the following inclusion criteria: (i) RT-PCR confirmed COVID-19; (ii) lab-confirmed other types of pneumonia before Dec. 2019; (iii) have non-contrast enhanced chest CT at diagnosis time. Pneumonia patients were

collected before Dec. 2019 to ensure they were non-COVID-19.

In the COVID-19 dataset there are images from several centers:

- *Wuhan city and Henan province: 560 COVID-19 and 149 other pneumonia.* ***[Development]***
- *Anhui province: 102 COVID-19 and 124 other pneumonia.* ***[External validation]***
- *Heilongjiang province: 92 COVID-19 and 69 other pneumonia.* ***[External validation]***
- *Beijing: 53 COVID-19.* ***[External validation]***
- *Huangshi city: 117 COVID-19.* ***[External validation]***

There are no clear differences in the demographics of each cohort. It is not explained in the paper how these patients were selected from each center.

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		UNCLEAR	-
1.2 Were all inclusions and exclusions of participants appropriate?		LOW	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**UNCLEAR**	**-**

Rationale of bias rating:

Reviewer 1:

It is unclear what subsets were selected from each center, e.g. data range and exclusion criteria. This selection process could potentially introduce biases.

Reviewer 2:

Large but not balanced number of positive and negative cases, with more positive cases than negative cases. Inclusion criteria for selection of COVID-19 positive and non-COVID-19 pneumonia cases were appropriate. However, it is unclear how participants were selected, either all included within a certain time period, or as a subset.

B. Applicability

Describe included participants, setting and dates: ***As discussed.***

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**UNCLEAR**	**-**

Rationale of applicability rating:

Reviewer 1:

For the reasons discussed above.

Reviewer 2:

Unclear how the partitions were selected.

TT. (43) **Zhu 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

The dataset consists of 422 CT scans for 422 COVID-19 patients from Shanghai Public Health Clinical Center and Sichuan University West China Hospital. All patients were confirmed by the national centers for disease control (CDC) based on the positive new Coronavirus nucleic acid antibody. The paper aims to predict severe from non-severe COVID-19 along with the conversion time from non-severe to severe COVID-19.

The male/female breakdown of the severe cases is 35/51 whereas for the non-severe cases it is 160/162.

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

It is unclear what subsets were selected from each center, e.g. data range and exclusion criteria. This selection process could potentially introduce biases. The method for labelling a patient as severe/non-severe is not discussed.

Reviewer 2:

Patients were confirmed covid-19 with a PCR test but severe/non-severe method of determining ground truth not given.

B. Applicability

Describe included participants, setting and dates: ***As discussed.***

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**LOW**	**-**

Rationale of applicability rating:

Reviewer 1:

No issues noted.

Reviewer 2:

The demographics approximately match (or at least aren't unbalanced).

Update to 14 August 2020

UU.(55) **Chen 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

"All COVID-19 patients treated in Chengdu Public Health Center between Jan 20, 2020 and Mar 31, 2020 were enrolled in our study. The diagnosis of COVID-19 was based on a positive result high throughput sequencing or real-time reverse transcriptase–polymerase-chain-reaction (RT-PCR) assay of nasal and pharyngeal swab specimens. After collecting the CT imaging and clinical management data, a subset of patients were excluded according to the following criteria: (i) age < 18 years old; (ii) incomplete medical records; (iii) cases with no arterial blood analysis result corresponding to

respective CT images."

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		LOW	-
1.2 Were all inclusions and exclusions of participants appropriate?		LOW	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**LOW**	**-**

Rationale of bias rating:

Reviewer 1:

Appropriate datasets with reasonable inclusion/exclusion criteria.

Reviewer 2:

Full cohort study with clear and appropriate inclusion and exclusion criteria.

B. Applicability

Describe included participants, setting and dates: ***As discussed.***

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of applicability rating:

Reviewer 1:

Uses data from only one center.

Reviewer 2:

The use of only one hospital means that the patients may not be representative of other cases, and so the results may not be more widely applicable. There is no external test ("validation") set

VV. (48) **HanqingChao 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

The following datasets are included in this paper:

- *<u>113 CT images from Firoozgar Hospital (Tehran, Iran)</u>*

  'Medical records of adult patients admitted with known or suspected COVID-19 pneumonia in Firoozgar Hospital (Tehran, Iran) between February 23, 2020 and March 30, 2020. Among the 117 patients with positive RT-PCR assay for COVID-19, three patients were excluded due to presence of extensive motion artifacts on their chest CT. With one patient who neither admitted to ICU nor discharged, 113 patients are used in this study.'

- *<u>125 CT images from Massachusetts General Hospital (Boston, MA, USA)</u>*

  'Medical records of adult patients admitted with COVID-19 symptom in MGH between March 11 and May 3, 2020. 125 RT-PCR positive admitted patients underwent unenhanced chest CT are selected to form this dataset.'

- *<u>57 CT images from University Hospital Maggiore della Carita (Novara, Piedmont, Italy)</u>*

  We reviewed medical records of adult patients admitted with COVID-19 pneumonia in the Novara Hospital (Piedmont, Italy) between March 4, 2020 and April 6, 2020. We collected clinical and outcome information of 57 patients with positive RT-PCR assay for COVID-19.

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		LOW	-
1.2 Were all inclusions and exclusions of participants appropriate?		LOW	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**LOW**	**-**

Rationale of bias rating:

<u>Reviewer 1:</u>

Appropriate datasets with reasonable inclusion/exclusion criteria.

Reviewer 2:

Patients seem to be consecutive with PCR tests to identify positive patients and reasonable inclusion/exclusion criteria.

B. Applicability			
Describe included participants, setting and dates: ***As discussed.***			
Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**LOW**	**-**

Rationale of applicability rating:

Reviewer 1:

No issues.

Reviewer 2:

Using datasets from USA, Iran and Italy gives a lower risk of applicability issues for the problem statement.

WW. (50) **Qin 2020**

DOMAIN 1: Participants
A. Risk of Bias
Describe the sources of data and criteria for participant selection:

The paper is unclear precisely on the data source, mentioning this is a bi-center study. We therefore assume, based on the author affiliations that this refers to: Shanghai Jiao Tong University Medical School Affiliated Ruijin Hospital and Ruian People's Hospital.

The dataset description says that "from January 19 to February 6, 2020, 311 patients were enrolled at the fever observation department. Inclusion criteria for patients suspected of COVID-19 were set according to the sixth edition of the Diagnosis and Treatment Program of COVID-19 proposed by The National Health Commission of the People's Republic of China: (1) epidemiological history: history of travel to Wuhan or history of residence in Wuhan or other areas with continuous transmission of

local cases within 14 days before the onset of the disease, history of contact with COVID-19 patients within 14 days before the onset of the disease, and clustering or epidemiological association with COVID-19, and (2) clinical features: fever and/or disorder of the respiratory system, imaging manifestations of COVID-19 pneumonia, normal or reduced leukocyte count, or reduced lymphocyte count. Patients with epidemiological history and any two of the three above-mentioned clinical features and patients without epidemiological history and with all the three clinical features were classified by the multiple-disciplinary expert group as suspected.

"All the suspected patients were tested by RT-PCR. Throat and nose swab specimens were obtained. A total of 106 patients with positive results of RT-PCR tests conducted at Shanghai Municipal CDC were included and considered as COVID-19-positive cases. The RT-PCR tests were repeated for 21 COVID-19 patients because results of the first RT-PCR tests were negative. Patients with negative chest CT manifestation (n = 12, 11.3%), missing data (n = 5, 4.7%), and poor quality of CT images (n = 1, 0.9%) were excluded. Cases with the negative results of RT-PCR tests at least twice consecutively were considered as non-COVID-19. For 205 with negative results of RT-PCR, pneumonia was diagnosed based on the Infectious Diseases Society of America/American Thoracic Society (IDSA/ATS) guideline. In brief, patients with at least one of the clinical symptoms of cough, sputum, fever, dyspnea, and pleuritic chest pain, plus at least one finding of coarse crackles on auscultation or elevated inflammatory biomarkers, in addition to a new pulmonary infiltration on chest CT, were diagnosed to have pneumonia. Patients with poor quality of medical images (n = 6, 2.9%), negative chest CT manifestation (n = 78, 38.0%), lung cancer (n = 3, 1.5%), and missing data (n = 38, 18.5%) were excluded. Throat and nose swab specimens of COVID-19-negative patients were tested by IgM antibody and influenza viruses A and B as appropriate for the detection of etiology. Bacterial infection was also diagnosed according to the IDSA/ATS guidelines. Finally, a total of 168 patients, including 88 COVID-19-positive and 80 COVID-19-negative subjects, were included in the present study."

Demographics of the dataset are reported. Differences in age and sex are not statistically significant at p = 0.05 level.

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		LOW	-
1.2 Were all inclusions and exclusions of participants appropriate?		LOW	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**LOW**	**-**

Rationale of bias rating:

Reviewer 1:

Appropriate datasets with reasonable inclusion/exclusion criteria.

Reviewer 2:

Given the dates referenced, it is likely that the patients were consecutive and therefore low bias with respect to patient selection. The inclusion criteria for positive cases was based on a positive RT-PCR, and exclusion for negative cases based on at least 2 negative RT-PCR. However, the exclusion of patients with no visible evidence of CT changes despite positive RT-PCR may lead to a bias towards selection of severe cases. There are no issues with cohort demographics.

B. Applicability

Describe included participants, setting and dates: ***As discussed.***

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**LOW**	**-**

Rationale of applicability rating:

Reviewer 1:

No issues.

Reviewer 2:

Appropriate inclusion and exclusion criteria, with matched demographics between cohorts.

XX. (52) **Wei 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

'The patients' data were collected from the First Affiliated Hospital of University of Science and Technology of China and the Affiliated Infectious Disease Hospital. During the period between January 20, 2020 and February 20, 2020, patients were included if they met the following criteria: (1) exhibiting positive results of 2019-nCoV nucleic acids and (2) having undergone chest CT examination during the initial diagnosis (within 3 days after admission). Excluded were those who had no obvious lung CT abnormalities or had pneumonia caused by other common bacterial or viral pathogens. According to the clinical classification criteria, 81 patients were enrolled (60 common cases and 21 severe cases).'

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		LOW	-
1.2 Were all inclusions and exclusions of participants appropriate?		LOW	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**LOW**	**-**

Rationale of bias rating:

Reviewer 1:

Appropriate datasets with reasonable inclusion/exclusion criteria.

Reviewer 2:

No issues.

B. Applicability

Describe included participants, setting and dates: ***As discussed.***

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**LOW**	**-**

Rationale of applicability rating:

Reviewer 1:

No issues.

Reviewer 2:

Data only from China, not clear if applicable outside of this demographic.

YY. (53)**Wu 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

'The institutional review board of Renmin Hospital of Wuhan University (Centre 1), Huangshi Central Hospital (Centre 2), Henan Provincial People's Hospital (Centre 3), and Beijing Youan Hospital (Centre 4) approved this multi-regional retrospective study, and the informed consent was waived. From November 29, 2019 to February 19, 2020, a total of 492 patients diagnosed with COVID-19 by etiological evidence of reverse transcriptase– polymerase chain reaction (RT-PCR) test were retrospectively collected. To relieve the impact of different durations from symptom onsets to the first CT scanning, we designated the 492 patients into two groups: 1) the early-phase group: CT scans were performed within one week after symptom onset (0-6 days, n = 317); and 2) the late-phase group: CT scans were performed one week later after symptom onset (≥7 days, n = 175). Here, day 0 was defined as the initial day of symptom onset, which was self-reported by patients on admission. In the early phase group: 212 patients from Center 1 (n = 106) and Center 2 (n = 106) comprised the training cohort since they all came from Hubei Province (the hardest-hit region). 105 patients from Center 3 (n = 65) and Center 4 (n = 40) comprised the validation cohort. In the late-phase group: 139 patients from Center 1 (n = 125) and Center 2 (n = 14) comprised the training cohort, and 36 patients from Center 3 (n = 23) and Center 4 (n = 13) comprised the validation cohort. All the included patients had regular follow-up for at least five days. The end-points of this study was the poor outcome, which was defined as death, need for mechanical ventilation, or ICU admission [6,32,33]. The follow-up durations were assessed from CT evaluation to poor outcome.'

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		LOW	-
1.2 Were all inclusions and exclusions of participants appropriate?		UNCLEAR	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**UNCLEAR**	**-**

Rationale of bias rating:

Reviewer 1:

They say 'evidence of RT-PCR' for inclusion but don't mention the results of those PCR tests, though we assume positive for all.

Reviewer 2:

The authors don't mention how many patients tested positive and how many of them had a CT and were therefore included. This may introduce bias if less symptomatic patients do not receive a CT and only have a PCR test +/- x-Ray. Additionally, the authors don't mention if there were further

exclusion criteria.

B. Applicability

Describe included participants, setting and dates: ***As discussed.***

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**UNCLEAR**	**-**

Rationale of applicability rating:

Reviewer 1:

No issues.

Reviewer 2:

Due to the selection of only patients with CT, it is unclear how well the algorithm would generalise to other hospitals where CT scans may be offered to patients more liberally or restrictively. However, the question of the study is well within the scope of our review.

ZZ. (54) **Zheng 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

'Data of patients diagnosed with COVID-19 who were admitted to the Jingzhou Central Hospital, Wuhan between January 21st and March 3rd, 2020 were reviewed. Inclusion criteria were as follows: 1) patients with laboratory-confirmed SARS-CoV-2 infection; 2) patients who underwent chest CT and laboratory tests on admission; and 3) patients with a minimum hospital stay of 7 days. Patients were excluded if any of the following conditions were met: 1) patients who were admitted to the intensive care unit (ICU) or underwent mechanical ventilation on admission (n = 8); 2) patients who were transferred or hospitalized before (n = 16); or 3) motion artefacts interfered with imaging diagnosis (n = 1). All patients were confirmed with COVID-19 infection using gene-sequencing or real time reverse-transcriptase polymerase chain reaction (RT-PCR) assays. Ultimately, 166 consecutive patients (103 males and 63 females; age 43.8 ± 12.3 years) were eligible and allocated to the training cohort.

'Patients from FuYang No.2 People's Hospital, Anhui employed the same inclusion and exclusion criteria, and 72 consecutive patients (38 males and 34 females; age 45.1 ± 15.8 years) were enrolled and assigned to the validation cohort.'

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		LOW	-
1.2 Were all inclusions and exclusions of participants appropriate?		LOW	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**LOW**	**-**

Rationale of bias rating:

Reviewer 1:

Appropriate datasets with reasonable inclusion/exclusion criteria.

Reviewer 2:

Appropriate data sources were used, and the demographics between the test and training sets are similar. It is a whole cohort.

B. Applicability

Describe included participants, setting and dates: ***As discussed.***

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**UNCLEAR**	**-**

Rationale of applicability rating:

Reviewer 1:

No issues, though only trained on China data.

Reviewer 2:

It is unclear whether this study is intended to be applicable beyond the geographically narrow population included in this study.

Update to 3rd October

AAA. (57) **Schalekamp 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

Data is from two large community hospitals in the Netherlands; Meander Medical Center, Amersfoort and Isala Hospitals, Zwolle.

"All consecutive patients between March 7th and April 24th 2020 suspected of COVID-19 on admission at the emergency department were derived. Patients that did not have a positive real-time reverse transcription polymerase chain reaction (RT-PCR) proven COVID-19 or patients that were not hospitalized were excluded. First RT-PCR tests were taken within 24 hours of hospital admission. If the first test was negative but clinical suspicion remained subsequent tests were performed.

"Other exclusion criteria were no chest radiography on admission, transferred patients with uncertain

onset of symptoms, status after pneumonectomy, and children (<18 years)."

		Dev	Val
1.5 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		LOW	-
1.6 Were all inclusions and exclusions of participants appropriate?		LOW	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**LOW**	**-**

Rationale of bias rating:

Reviewer 1:

The inclusion and exclusion criteria are reasonable, and the patient population is also reasonable.

Reviewer 2:

A cohort study consisting of all consecutive patients.

B. Applicability			
Describe included participants, setting and dates: Discussed earlier.			
Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**UNCLEAR**	**-**

Rationale of applicability rating:

Reviewer 1:

There is no perceived bias in how the patients were selected as all are included. The demographics are also similar between the critical and non-critical cohorts.

Reviewer 2:

This is one local demographic, and it is unclear how widely applicable it will be to other populations.

BBB. (28) **Li 2020**

DOMAIN 1: Participants
A. Risk of Bias
Describe the sources of data and criteria for participant selection

Data sources

- Publicly available CheXpert for pretraining and validation
- COVID-19 internal CXR dataset: CXR from Massachusetts General hospital
- COVID-19 external CXR dataset: Newton-Wellesley hospital

Internal validation set

- COVID-19 training set: 314 CXR from consecutive, unique patients hospitalised in part between April 1st-10th 2020.
- COVID-19 test set: 154 CXR from consecutive, unique patients hospitalised in part between March 27th-31st 2020. 92 underwent follow-up CXR used for longitudinal analysis.
- Excluded one patient due to pneumonectomy.
- Intubation and mortality data collected from medical records by two blinded investigators

External validation set

- External data set: 113 CXR of consecutive, unique patients hospitalised in part on April 15th 2020 at community hospital.

		Dev	Val
1.3 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		LOW	-
1.4 Were all inclusions and exclusions of participants appropriate?		LOW	-
Risk of bias introduced by selection of participants	**RISK:** *(**low**/ high/ unclear)*	**LOW**	**-**

Rationale of bias rating:

Reviewer 1:

Overall, a low risk of bias associated with the data source, using consecutive unique patients from two separate hospitals confirmed by PCR. Inclusion and exclusion criteria appropriate, excluding single patient with pneumonectomy.

Reviewer 2:

Seems a very reasonable dataset with consecutive patients and reasonable inclusion/exclusion criteria.

B. Applicability

Describe included participants, setting and dates: Discussed earlier.

Concern that the included participants and setting do not match the review question	**CONCERN:** ***(low**/ high/ unclear)*	**LOW**	**-**

Rationale of applicability rating:

Reviewer 1:

Appropriate inclusion and exclusion criteria for development of a prognostic model.

Reviewer 2:

No issues.

CCC. (58) **Xie 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

This study was approved by multiple Institutional Review Boards of Queen Mary hospital (QM), The University of HK PET CT unit (HKU) and Pamela Youde Nethersole Eastern (PYNE) hospital.

Case group: data screened between 24th January and 31st March 2020 from QM and PYNE hospitals. Patients with laboratory-confirmed COVID-19 by reverse transcription polymerase chain reaction were included, and their initial CT scans were retrieved.

"Control group: data were screened between 04/02/2012 and 31/03/2020 from HKU. Patients with reported GGOs in the radiological report were included in this study. A board-certified radiologist with fellowship training in cardiothoracic imaging (V.V., with 10 years' experience) then reviewed cases and included cases that have similar ground glass opacity appearances. For patients in the control group collected after December 2019, underwent strict clinical +/− laboratory assessment prior to entering the unit to exclude potential infection with COVID-19. "

Clinical details such as history and clinical assessment was obtained as standard for diagnosis, in conjunction with histological and laboratory tests if they were available. Patients with incomplete data were excluded. A total of 301 patients (age mean ± SD: 64 ± 15 years; male: 52.8 %) were enrolled in this study."

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		HIGH	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(**low**/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

It is not specified what is meant by 'strict clinical +/- laboratory assessment' and therefore the bias with respect to the control group containing COVID-19 positive patients cannot be evaluated. However, there is a risk of source bias resulting from the collection of images of control and case

groups from different hospitals.

Reviewer 2:

Since the case group and control group come from different hospitals and populations, it is likely that a considerable bias is introduced.

B. Applicability

Describe included participants, setting and dates: Discussed earlier.

Concern that the included participants and setting do not match the review question	**CONCERN:** ***(low**/ high/ unclear)*	**UNCLEAR**	**-**

Rationale of applicability rating:

Reviewer 1:

More information is required regarding the inclusion criteria for the control group.

Reviewer 2:

Appropriate criteria for a COVID19 diagnosis model.

DDD. (34) **Li 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

The hospitals involved in this study include Massachusetts General Hospital (Hospital 1) (Boston, USA), Hospital Santa Paula (Hospital 2) (São Paulo, Brazil), and Newton Wellesley Hospital (Hospital 3) (Newton, USA), Hospitals 1 and 2 are large academic medical centers, while Hospital 3 is a community hospital in the Boston metropolitan area.

Hospital 1 Outpatient Dataset. This dataset was composed of 358 CXRs from 349 unique patients who presented for outpatient imaging at urgent care or respiratory illness clinics associated with

Hospital 1 and tested positive for COVID-19 by nasopharyngeal swab RT-PCR obtained at their outpatient visit from March 15, 2020 to April 15, 2020.

Hospital 2 Emergency Test Set. This dataset was composed of 303 CXRs from 242 unique patients who presented to the emergency department with suspected COVID-19 at Hospital 2. These CXRs were sampled from patients from February 1, 2020 to May 30, 2020 with at least one COVID-19 RT-PCR result within ±3 days of the CXR. Sampling was stratified on RT-PCR test results, so that 70% of CXRs in the dataset would have at least one positive associated test and 30% would have all negative tests.

In addition to these two data sets that were created for this study, we also used previously published data sets for model testing, including 154 admission CXRs from 154 unique patients hospitalized for COVID-19 at Hospital 1 (Hospital 1 Inpatient Test Set) and 113 admission CXRs from 113 unique patients hospitalized for COVID-19 at Hospital 3 (Hospital 3 Inpatient Test Set).

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		LOW	-
1.2 Were all inclusions and exclusions of participants appropriate?		LOW	-
Risk of bias introduced by selection of participants	**RISK:** *(**low**/ high/ unclear)*	**LOW**	**-**

Rationale of bias rating:

Reviewer 1:

Inclusion and exclusion criteria were consistent and clearly communicated. All labels were obtained using RT-PCR. Data were collected from multiple sources.

Reviewer 2:

Overall low risk of bias. However, there were some patients for whom multiple chest x-rays were included in the study. By splitting the datasets at the image level, this may have lead to a contamination of the testing data set.

B. Applicability

Describe included participants, setting and dates: Discussed earlier.

Concern that the included participants and setting do not match the review question	**CONCERN:**	**LOW**	**-**

	***(low**/ high/ unclear)*		

Rationale of applicability rating:

Reviewer 1:

The patients represented in the training and testing sets are representative of the target population.

Reviewer 2:

No concern about the applicability.

EEE. (29) **Wang 2020**

DOMAIN 1: Participants
A. Risk of Bias

Describe the sources of data and criteria for participant selection:

"Patients with RT-PCR-confirmed COVID-19 who were admitted to Tongji Hospital (Wuhan, China) between Feb 1, 2020, and March 3, 2020, were identified and their unenhanced chest CT scans were retrieved from the Picture Archiving and Communication System of Tongji Hospital. The scans were obtained using a variety of scanner models and manufacturers. We also collected patient demographic information and RT-PCR test results from electronic medical records."

"Unenhanced CT chest scans for 2,191 adult patients (aged >14 years) with COVID-19 and 1,000 adult patients without COVID-19 who were admitted to Tongji Hospital during the same time period and had double negative RT-PCR test results were selected for algorithm development. The patients in the non-COVID-19 group might or might not have had positive CT findings. For patients who had undergone multiple CT scans, we used the first scan that had COVID-19 imaging manifestation for algorithm development."

"The dataset was randomly split into a development set (1,674 patients with COVID-19; 800 patients without COVID-19) and an internal validation set (439 patients with COVID-19; 200 patients without COVID-19) in a ratio of 8:2."

"Positive cases in the development set were annotated by radiologists. 105 cases were excluded due to difficulty with annotation. After data annotation, the development set was randomly split into a training set (1,318 patients with COVID-19; 640 patients without COVID-19) and a testing set

(329 patients with COVID-19; 160 patients without COVID-19) with a ratio of 8:2."

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		LOW	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(**low**/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

A very different selection of patients in training and test set. In and exclusion criteria disagree between training and testing data.

Reviewer 2:

The development set and internal validation set have difference inclusion and exclusion criteria. Also, it is unclear if imaging features informed the COVID-19/non-COVID-19 labels, and it is unclear how multiple scans from non-COVID-19 patients were handled.

B. Applicability

Describe included participants, setting and dates: Discussed earlier.

Concern that the included participants and setting do not match the review question	**CONCERN:** ***(low**/ high/ unclear)*	**LOW**	**-**

Rationale of applicability rating:

Reviewer 1:

Low level of concerns in terms of applicability.

Reviewer 2:

The patients in the development set are representative of the target population.

FFF. (33) **Zhang 2020**

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: CXR from patients with and without COVID-19 pneumonia from five hospitals. Pneumonia findings found using natural language processing searching radiologist reports. Non-COVID-19 pneumonia was selected based on pneumonia finding in report and date of study. Patients with pneumonia from COVID-19 timeframe were cross-referenced with lists of patients positive for COVID-19. Inclusion for non-COVID-19 pneumonia: frontal CXR, pneumonia diagnosis and imaging between October and December 2019. Patients under 18 excluded. Inclusion for COVID-19 group: frontal CXR, RT-PCR positive with diagnosis between February and May 2020. Excluded if CXR was performed more than 5 days prior or 14 days after RT-PCR confirmation. "The inclusion criteria for the COVID-19 positive group were patients that underwent frontal view CXR, with RT-PCR positive test for SARS-CoV-2 with a diagnosis of pneumonia between February 1, 2020 and May 31, 2020. Patients were excluded if CXR was performed more than 5 days prior or 14 days after RT-PCR confirmation" "The resulting datasets consisted of 5805 CXRs with RT-PCR confirmed COVID-19 pneumonia from 2060 patients and 5300 CXRs with non-COVID-19 pneumonia from 3148 patients for use in this study"			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		LOW	-
1.2 Were all inclusions and exclusions of participants appropriate?		LOW	-
Risk of bias introduced by selection of participants	**RISK:** *(**low**/ high/ unclear)*	**LOW**	**-**
Rationale of bias rating: ***Reviewer 1:*** Appropriate inclusion and exclusion criteria, notably selecting only frontal CXRs and excluding paediatric patients. Dataset is large and classes balanced.			

Reviewer 2:

The timing of the acquisition of the non-COVID19 patients ensures that these do not have COVID19. The high number of participating hospitals further implies a low risk of bias.

B. Applicability

Describe included participants, setting and dates: Discussed earlier.

Concern that the included participants and setting do not match the review question	**CONCERN:** ***(low**/ high/ unclear)*	**LOW**	**-**

Rationale of applicability rating:

Reviewer 1:

Large and balanced dataset, with appropriate inclusion and exclusion criteria. Cases and control demographics similar for age and sex.

Reviewer 2:

The appropriate choice of patients for the COVID19 and patients with non-COVID19 related pneumonia patients make this a suitable choice of participants for the distinguishing the two cases.

GGG. (35)**Wang 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

"We collected CT scans of 4657 patients (F/M, 1946/2711; mean age: 46 ± 17 years) from several cooperative hospitals, including a total of 936 normal scans, 2406 scans with ILD caused by viruses, and 1315 scans with COVID-19. All the pneumonia diseases were confirmed as positive by RT-PCR or serum antibody test besides COVID-19. The ILD patient inclusion or exclusion criteria was executed based on "An official American Thoracic Society/European Respiratory Society statement" by two experienced respiratory physicians (HL with 10 years of experience and FX with 15 years of experience). All the ILD CT images were independently reviewed by two experienced radiologists in CT diagnostics (XL with 8 years of experience and CL with 10 years of experience). The ILD CT images must have the pulmonary fibrosis features. In clinical practice, there were patients who underwent several scans. For each of these patients, we selected only the scan that was firstly reconstructed with the thinnest slice-thickness for building the dataset."

The geographical location of these hospitals is not stated, but one presumes from the authors' affiliations that they are probably all in China, or possibly even in a small region of China.

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		UNCLEAR	-
1.2 Were all inclusions and exclusions of participants appropriate?		UNCLEAR	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**UNCLEAR**	**-**

Rationale of bias rating:

Reviewer 1:

The choice of which patients' scans to pass on to the study authors is not stated. The inclusion/exclusion criteria are "based on" an official statement, but the details are not specified. It seems likely that there is little bias here, but there is no way of being confident of this. Unclear how many hospitals were involved, and how many patients contributed by each hospital.
Inclusion/exclusion criteria was mentioned for interstitial lung disease, but not for other cases. Also unclear whether patients were consecutive, and over what time period.

Reviewer 2:

Unclear how many hospitals were involved, and how many patients contributed by each hospital.

B. Applicability

Describe included participants, setting and dates:

Neither the setting not the dates are specified, and neither are any demographics provided.

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of applicability rating:

Reviewer 1:

Without details of the demographics or geographical location, there is a significant question about the wider applicability of the results. It may well be, though, that the model is more widely applicable.

Reviewer 2:

Unclear information about hospital sources limits application.

HHH. (60) **Wang 2020**

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: Multi-centre retrospective cohort study RT-PCR confirmed COVID-19 patients from three hospitals in Hubei province. Total of 161 patients (89/161 male and 72/161 female) with at least two CT scans included.			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		UNCLEAR	-
1.2 Were all inclusions and exclusions of participants appropriate?		UNCLEAR	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**UNCLEAR**	**-**
Rationale of bias rating: ***Reviewer 1:*** Low risk of bias associated with COVID-19 patients diagnosed based on WHO criteria followed by RT-PCR of throat swab specimens. However, unclear how many patients were derived from each hospital, unclear whether patient selection was consecutive and what exclusion criteria, if any, were used. ***Reviewer 2:*** Unclear which hospitals are being used. 3 designated in Hubei for training and one more for testing. "We included patients with confirmed COVID-19 pneumonia who were admitted to one of three hospitals in Hubei, all designated for COVID-19, and who underwent at least 2 chest CT scans after admission. The diagnosis of COVID-19 at the three hospitals was initially based on the criteria published by WHO on Jan 12, 2020, and all cases were later confirmed by real-time RT-PCR analysis of throat swab specimens according to a previously published protocol."			
B. Applicability			
Describe included participants, setting and dates: Discussed earlier.			
Concern that the included participants and setting do not	**CONCERN:**	UNCLEAR	**-**

match the review question	*(low/ high/ unclear)*		

Rationale of applicability rating:

Reviewer 1:

Unclear inclusion and exclusion criteria limits applicability of this dataset.

Reviewer 2:

As we cannot fully understand the datasets used, we cannot conclude on applicability.

III. (61) **Xu 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

"We recruited a total of 362 confirmed COVID-19 patients with two independent qRT-PCR tests from Wuhan Union Hospital between January 2020 and March 2020 in Wuhan, Hubei Province, China". Patients were categorised as "148 severe, 214 non-severe COVID-19 129"

Additionally, the authors "recruited 129 confirmed non-COVID viral infection participants from Kunshan Hospital, Suzhou, China".

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		UNCLEAR	-
1.2 Were all inclusions and exclusions of participants appropriate?		HIGH	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of bias rating:

Reviewer 1:

It is unclear how participants for non-COVID viral pneumonia were confirmed to be COVID-19 negative and also what the inclusion criteria are. As all non-COVID patients are from a separate

center this introduces a high risk of bias.

Reviewer 2:

Unclear how participants for non-COVID viral pneumonia were confirmed to be COVID-19 negative, and what inclusion/exclusion criteria were used for non-COVID viral infections.

B. Applicability

Describe included participants, setting and dates: Discussed earlier.

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**HIGH**	**-**

Rationale of applicability rating:

Reviewer 1:

As all non-COVID-19 patients are from one center and the COVID-19 patients are from another, both in China, we cannot conclude on the applicability.

Reviewer 2:

Unclear information about inclusion/exclusion criteria limits applicability.

JJJ. (51) **Ramtohul 2020**

DOMAIN 1: Participants

A. Risk of Bias

Describe the sources of data and criteria for participant selection:

"This prospective study was conducted at the Institut Curie Hospitals (ICH) in Paris and St Cloud (France)."

"All consecutive patients treated for cancer with chest CT evidence of COVID19 pneumonia were prospectively included during the European

"COVID-19 epidemic outbreak between March 15, 2020, and April 20, 2020. Based on the COVID-19 reporting and data system, mandatory features for chest CT COVID-19 pneumonia were based on multifocal ground-glass opacities, with or without consolidations, reticular thickening, or subpleural bands. Chest CT examinations were requested for either clinical suspicion of COVID-19 pneumonia, history of exposure to confirmed COVID-19 cases, RT-PCR-positive swab, suspicion of pulmonary

embolism, or routine cancer follow-up examination. Patients with no lung abnormalities were not included. Also, patients with preexisting equivocal findings before March 2020, such as GGO, were not included after comparison of the study chest CT scan with a previous CT scan (all imaging records for cancer patients treated at ICH are locally centralized).”

		Dev	Val
1.3 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		UNCLEAR	-
1.4 Were all inclusions and exclusions of participants appropriate?		UNCLEAR	-
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	UNCLEAR	**-**

Rationale of bias rating:

Reviewer 1:

Patients with pre-existing equivocal findings before March 2020, such as GGO, were not included after comparison of the study chest CT scan with a previous CT scan (all imaging records for cancer patients treated at ICH are locally centralized).

Chest CT examinations were requested for either clinical suspicion of COVID-19 pneumonia, history of exposure to confirmed COVID-19 cases, RT-PCR-positive swab, suspicion of pulmonary embolism, or routine cancer follow-up examination. Patients with no lung abnormalities were not included.

Reviewer 2:

All consecutive patients treated for cancer with chest CT evidence of COVID- 19 pneumonia were prospectively included during the European COVID-19 epidemic outbreak between March 15, 2020, and April 20, 2020.

Unclear how patients were determined to be COVID-19 positive

B. Applicability

Describe included participants, setting and dates: Discussed earlier.

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	UNCLEAR	**-**

Rationale of applicability rating:

Reviewer 1:

Only one location so unclear if it would be applicable outside this population.

Reviewer 2:

Unclear how patients were included so not sure about applicability.

SECTION 3. References

1. Acar E, ŞAhİN E, Yilmaz İ. Improving effectiveness of different deep learning-based models for

detecting COVID-19 from computed tomography (CT) images. medRxiv [Internet]. Cold Spring Harbor Laboratory Press; 2020 [cited 2020 Jul 29]; Available from: https://doi.org/10.1101/2020.06.12.20129643

2. Amyar A, Modzelewski R, Li H, Ruan S. Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation. Comput Biol Med [Internet]. Elsevier Ltd; 2020 [cited 2020 Dec 7];126. Available from: https://pubmed.ncbi.nlm.nih.gov/33065387/

3. Ardakani AA, Kanafi AR, Acharya UR, Khadem N, Mohammadi A. Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. Comput Biol Med. Elsevier Ltd; 2020;121:103795.

4. Bai HX, Wang R, Xiong Z, Hsieh B, Chang K, Halsey K, et al. AI Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Etiology on Chest CT. Radiology. Radiological Society of North America; 2020;201491.

5. Ghoshal B, Tucker A. Estimating Uncertainty and Interpretability in Deep Learning for Coronavirus (COVID-19) Detection - v2 [Internet]. arXiv. 2020 [cited 2020 Jul 29]. Available from: http://arxiv.org/abs/2003.10769

6. Ezzat D, Hassanien AE, Ella HA. An optimized deep learning architecture for the diagnosis of COVID-19 disease based on gravitational search optimization. Appl Soft Comput J [Internet]. Elsevier Ltd; 2020 [cited 2020 Dec 7]; Available from: https://pubmed.ncbi.nlm.nih.gov/32982615/

7. Luz E, Silva PL, Silva R, Silva L, Moreira G, Menotti D. Towards an Effective and Efficient Deep Learning Model for COVID-19 Patterns Detection in X-ray Images - v4 [Internet]. arXiv. 2020 [cited 2020 Jul 29]. Available from: http://arxiv.org/abs/2004.05717

8. Tartaglione E, Barbano CA, Berzovini C, Calandri M, Grangetto M. Unveiling COVID-19 from chest x-ray with deep learning: A hurdles race with small data. Int J Environ Res Public Health [Internet]. MDPI AG; 2020 [cited 2020 Dec 7];17:1–17. Available from: /pmc/articles/PMC7557723/?report=abstract

9. Gueguim Kana EB, Zebaze Kana MG, Donfack Kana AF, Azanfack Kenfack RH. A web-based Diagnostic Tool for COVID-19 Using Machine Learning on Chest Radiographs (CXR). medRxiv [Internet]. Cold Spring Harbor Laboratory Press; 2020 [cited 2020 Jul 29]; Available from: https://doi.org/10.1101/2020.04.21.20063263

10. Jin S, Wang B, Xu H, Luo C, Wei L, Zhao W, et al. AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system in four weeks. medRxiv [Internet]. Cold Spring Harbor Laboratory Press; 2020 [cited 2020 Jul 29];1–22. Available from: https://doi.org/10.1101/2020.03.19.20039354

11. Ko H, Chung H, Kang WS, Kim KW, Shin Y, Kang SJ, et al. COVID-19 Pneumonia Diagnosis Using a Simple 2D Deep Learning Framework With a Single Chest CT Image: Model Development and Validation. J Med Internet Res. NLM (Medline); 2020;22:e19569.

12. Mei X, Lee HC, Diao K yue, Huang M, Lin B, Liu C, et al. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nat Med. Nature Research; 2020;1–5.

13. Heidari M, Mirniaharikandehei S, Khuzani AZ, Danala G, Qiu Y, Zheng B. Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms. Int J Med Inform [Internet]. Elsevier Ireland Ltd; 2020 [cited 2020 Dec 7];144. Available from: https://pubmed.ncbi.nlm.nih.gov/32992136/

14. Bassi PRAS, Attux R. A Deep Convolutional Neural Network for COVID-19 Detection Using Chest X-Rays - v2 [Internet]. arXiv. 2020 [cited 2020 Jul 29]. Available from: http://arxiv.org/abs/2005.01578

15. Pu J, Leader J, Bandos A, Shi J, Du P, Yu J, et al. Any unique image biomarkers associated with COVID-19? Eur Radiol. Springer; 2020;1–7.

16. Wang S, Kang B, Ma J, Zeng X, Xiao M, Guo J, et al. A deep learning algorithm using CT images to screen for corona virus disease (COVID-19) - v5. medRxiv [Internet]. Cold Spring Harbor Laboratory Press; 2020 [cited 2020 Jul 29];1–19. Available from: https://doi.org/10.1101/2020.02.14.20023028

17. Wang S, Zha Y, Li W, Wu Q, Li X, Niu M, et al. A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. Eur Respir J [Internet]. European Respiratory Society; 2020 [cited 2020 Dec 7];56. Available from: https://pubmed.ncbi.nlm.nih.gov/32444412/

18. Chen X, Yao L, Zhou T, Dong J, Zhang Y. Momentum Contrastive Learning for Few-Shot COVID-19 Diagnosis from Chest CT Images [Internet]. arXiv. 2020 [cited 2020 Jul 29]. Available from: http://arxiv.org/abs/2006.13276

19. Li X, Li C, Zhu D. COVID-MobileXpert: On-Device COVID-19 Screening using Snapshots of Chest X-Ray - v3 [Internet]. arXiv. 2020 [cited 2020 Jul 30]. Available from: http://arxiv.org/abs/2004.03042

20. Zokaeinikoo M, Mitra P, Kumara S, Kazemian P. AIDCOV: An Interpretable Artificial Intelligence Model for Detection of COVID-19 from Chest Radiography Images - v3. medRxiv [Internet]. Cold Spring Harbor Laboratory Press; 2020 [cited 2020 Jul 29]; Available from: https://doi.org/10.1101/2020.05.24.20111922

21. Sayyed AQMS, Saha D, Hossain AR. CovMUNET: A Multiple Loss Approach towards Detection of COVID-19 from Chest X-ray. 2020 [cited 2020 Oct 12]; Available from: http://arxiv.org/abs/2007.14318

22. Malhotra A, Mittal S, Majumdar P, Chhabra S, Thakral K, Vatsa M, et al. Multi-Task Driven Explainable Diagnosis of COVID-19 using Chest X-ray Images. 2020 [cited 2020 Oct 12]; Available from: http://arxiv.org/abs/2008.03205

23. Rahaman MM, Li C, Yao Y, Kulwa F, Rahman MA, Wang Q, et al. Identification of COVID-19 samples from chest X-Ray images using deep learning: A comparison of transfer learning approaches. J Xray Sci Technol [Internet]. IOS Press; 2020 [cited 2020 Oct 12];28:821–39. Available from: https://pubmed.ncbi.nlm.nih.gov/32773400/

24. Amer R, Frid-Adar M, Gozes O, Nassar J, Greenspan H. COVID-19 in CXR: from Detection and Severity Scoring to Patient Disease Monitoring. 2020 [cited 2020 Oct 12]; Available from: http://arxiv.org/abs/2008.02150

25. Tsiknakis N, Trivizakis E, Vassalou E, Papadakis G, Spandidos D, Tsatsakis A, et al. Interpretable artificial intelligence framework for COVID‐19 screening on chest X‐rays. Exp Ther Med [Internet]. Spandidos Publications; 2020 [cited 2020 Oct 12];20:727–35. Available from: https://pubmed.ncbi.nlm.nih.gov/32742318/

26. Han Z, Wei B, Hong Y, Li T, Cong J, Zhu X, et al. Accurate Screening of COVID-19 Using Attention-Based Deep 3D Multiple Instance Learning. IEEE Trans Med Imaging [Internet]. Institute of Electrical and Electronics Engineers Inc.; 2020 [cited 2020 Oct 12];39:2584–94. Available from: https://pubmed.ncbi.nlm.nih.gov/32730211/

27. Shah V, Keniya R, Shridharani A, Punjabi M, Shah J, Mehendale N. Diagnosis of COVID-19 using CT scan images and deep learning techniques. medRxiv [Internet]. Cold Spring Harbor Laboratory Press; 2020 [cited 2020 Oct 12]; Available from: https://doi.org/10.1101/2020.07.11.20151332

28. Li MD, Arun NT, Aggarwal M, Gupta S, Singh P, Little BP, et al. Improvement and Multi-Population Generalizability of a Deep Learning-Based Chest Radiograph Severity Score for COVID-19. medRxiv [Internet]. Cold Spring Harbor Laboratory Press; 2020 [cited 2020 Dec 7]; Available from: https://doi.org/10.1101/2020.09.15.20195453

29. Wang M, Xia C, Huang L, Xu S, Qin C, Liu J, et al. Deep learning-based triage and analysis of lesion burden for COVID-19: a retrospective study with external validation. Lancet Digit Heal [Internet]. Elsevier Ltd; 2020 [cited 2020 Dec 7];2:e506–15. Available from: https://pubmed.ncbi.nlm.nih.gov/32984796/

30. Zhang R, Guo Z, Sun Y, Lu Q, Xu Z, Yao Z, et al. COVID19XrayNet: A Two-Step Transfer Learning Model for the COVID-19 Detecting Problem Based on a Limited Number of Chest X-Ray Images. Interdiscip Sci Comput Life Sci [Internet]. Springer Science and Business Media Deutschland GmbH; 2020 [cited 2020 Dec 7];12:555–65. Available from: /pmc/articles/PMC7505483/?report=abstract

31. Bararia A, Ghosh A, Bose C, Bhar D, Author C. Network for subclinical prognostication of COVID 19 Patients from data of thoracic roentgenogram: A feasible alternative screening technology. medRxiv [Internet]. Cold Spring Harbor Laboratory Press; 2020 [cited 2020 Dec 7]; Available from: https://doi.org/10.1101/2020.09.07.20189852

32. Wang Z, Xiao Y, Li Y, Zhang J, Lu F, Hou M, et al. Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest X-rays. Pattern Recognit [Internet]. Elsevier Ltd; 2021 [cited 2020 Dec 7];110:107613. Available from: /pmc/articles/PMC7448783/?report=abstract

33. Zhang R, Tie X, Qi Z, Bevins NB, Zhang C, Griner D, et al. Diagnosis of COVID-19 Pneumonia Using Chest Radiography: Value of Artificial Intelligence. Radiology [Internet]. Radiological Society of North America (RSNA); 2020 [cited 2020 Dec 7];202944. Available from: https://pubmed.ncbi.nlm.nih.gov/32969761/

34. Li MD, Arun NT, Gidwani M, Chang K, Deng F, Little BP, et al. Automated Assessment and Tracking of COVID-19 Pulmonary Disease Severity on Chest Radiographs using Convolutional Siamese Neural Networks. Radiol Artif Intell [Internet]. Radiological Society of North America (RSNA); 2020 [cited 2020 Dec 7];2:e200079. Available from: http://pubs.rsna.org/doi/10.1148/ryai.2020200079

35. Wang J, Bao Y, Wen Y, Lu H, Luo H, Xiang Y, et al. Prior-Attention Residual Learning for More Discriminative COVID-19 Screening in CT Images. IEEE Trans Med Imaging. Institute of Electrical and Electronics Engineers Inc.; 2020;39:2572–83.

36. Farooq M, Hafeez A. COVID-ResNet: A Deep Learning Framework for Screening of COVID19 from Radiographs. arXiv [Internet]. arXiv; 2020 [cited 2020 Dec 7]; Available from: http://arxiv.org/abs/2003.14395

37. Goncharov M, Pisov M, Shevtsov A, Shirokikh B, Kurmukov A, Blokhin I, et al. CT-based COVID-19 Triage: Deep Multitask Learning Improves Joint Identification and Severity Quantification - v1. arXiv [Internet]. arXiv; 2020 [cited 2020 Dec 7]; Available from: http://arxiv.org/abs/2006.01441

38. Chassagnon G, Vakalopoulou M, Battistella E, Christodoulidis S, Hoang-Thi T-N, Dangeard S, et

al. AI-Driven CT-based quantification, staging and short-term outcome prediction of COVID-19 pneumonia - v2. medRxiv [Internet]. 2020 [cited 2020 Jul 29]; Available from: https://doi.org/10.1101/2020.04.17.20069187

39. Chen XX, Tang Y, Mo Y, Li S, Lin D, Yang ZZ, et al. A diagnostic model for coronavirus disease 2019 (COVID-19) based on radiological semantic and clinical features: a multi-center study. Eur Radiol. Springer; 2020;1–10.

40. Shi F, Xia L, Shan F, Wu D, Wei Y, Yuan H, et al. Large-Scale Screening of COVID-19 from Community Acquired Pneumonia using Infection Size-Aware Classification [Internet]. arXiv. 2020 [cited 2020 Jul 29]. Available from: http://arxiv.org/abs/2003.09860

41. Guiot J, Vaidyanathan A, Deprez L, Zerka F, Danthine D, Frix A-N, et al. Development and validation of an automated radiomic CT signature for detecting COVID-19. medRxiv [Internet]. Cold Spring Harbor Laboratory Press; 2020 [cited 2020 Jul 29]; Available from: https://doi.org/10.1101/2020.04.28.20082966

42. Yue H, Yu Q, Liu C, Huang Y, Jiang Z, Shao C, et al. Machine learning-based CT radiomics method for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: a multicenter study. Ann Transl Med [Internet]. AME Publishing Company; 2020 [cited 2020 Dec 7];8:859–859. Available from: /pmc/articles/PMC7396749/?report=abstract

43. Zhu X, Song B, Shi F, Chen Y, Hu R, Gan J, et al. Joint Prediction and Time Estimation of COVID-19 Developing Severe Symptoms using Chest CT Scan [Internet]. arXiv. 2020 [cited 2020 Jul 29]. Available from: http://arxiv.org/abs/2005.03405

44. Lassau N, Ammari S, Chouzenoux E, Gortais H, Herent P, Devilder M, et al. AI-based multi-modal integration of clinical characteristics, lab tests and chest CTs improves COVID-19 outcome prediction of hospitalized patients - v2. medRxiv [Internet]. Cold Spring Harbor Laboratory Press; 2020 [cited 2020 Jul 29]; Available from: https://doi.org/10.1101/2020.05.14.20101972

45. Georgescu B, Chaganti S, Aleman GB, Barbosa EJM, Cabrero JB, Chabin G, et al. Machine Learning Automatically Detects COVID-19 using Chest CTs in a Large Multicenter Cohort - v2 [Internet]. arXiv. 2020 [cited 2020 Jul 29]. Available from: http://arxiv.org/abs/2006.04998

46. Cohen JP, Dao L, Morrison P, Roth K, Bengio Y, Shen B, et al. Predicting covid-19 pneumonia severity on chest x-ray with deep learning [Internet]. arXiv. arXiv; 2020 [cited 2020 Dec 7]. Available from: /pmc/articles/PMC7451075/?report=abstract

47. Elaziz MA, Hosny KM, Salah A, Darwish MM, Lu S, Sahlol AT. New machine learning method for imagebased diagnosis of COVID-19. PLoS One [Internet]. Public Library of Science; 2020 [cited 2020 Oct 12];15. Available from: https://pubmed.ncbi.nlm.nih.gov/32589673/

48. Chao H, Fang X, Zhang J, Homayounieh F, Arru CD, Digumarthy SR, et al. Integrative Analysis for COVID-19 Patient Outcome Prediction. ArXiv [Internet]. 2020 [cited 2020 Oct 12]; Available from: http://arxiv.org/abs/2007.10416

49. Gil D, Díaz-Chito K, Sánchez C, Hernández-Sabaté A. Early Screening of SARS-CoV-2 by Intelligent Analysis of X-Ray Images. 2020 [cited 2020 Oct 12]; Available from: http://arxiv.org/abs/2005.13928

50. Qin L, Yang Y, Cao Q, Cheng Z, Wang X, Sun Q, et al. A predictive model and scoring system combining clinical and CT characteristics for the diagnosis of COVID-19. Eur Radiol [Internet]. Springer; 2020 [cited 2020 Oct 12]; Available from: https://pubmed.ncbi.nlm.nih.gov/32607634/

51. Ramtohul T, Cabel L, Paoletti X, Chiche L, Moreau P, Noret A, et al. Quantitative CT Extent of Lung Damage in COVID-19 Pneumonia Is an Independent Risk Factor for Inpatient Mortality in a Population of Cancer Patients: A Prospective Study. Front Oncol [Internet]. Frontiers Media S.A.; 2020 [cited 2021 Jan 21];10. Available from: https://pubmed.ncbi.nlm.nih.gov/33014804/

52. Wei W, Hu X wen, Cheng Q, Zhao Y ming, Ge Y qiong. Identification of common and severe COVID-19: the value of CT texture analysis and correlation with clinical characteristics. Eur Radiol [Internet]. Springer; 2020 [cited 2020 Oct 12]; Available from: https://pubmed.ncbi.nlm.nih.gov/32613287/

53. Wu Q, Wang S, Li L, Wu Q, Qian W, Hu Y, et al. Radiomics analysis of computed tomography helps predict poor prognostic outcome in COVID-19. Theranostics [Internet]. Ivyspring International Publisher; 2020 [cited 2020 Oct 12];10:7231–44. Available from: https://pubmed.ncbi.nlm.nih.gov/32641989/

54. Zheng Y, Xiao A, Yu X, Zhao Y, Lu Y, Li X, et al. Development and validation of a prognostic nomogram based on clinical and ct features for adverse outcome prediction in patients with covid-19. Korean J Radiol [Internet]. Korean Radiological Society; 2020 [cited 2020 Oct 12];21:1007–17. Available from: https://pubmed.ncbi.nlm.nih.gov/32677385/

55. Chen Y, Wang Y, Zhang Y, Zhang N, Zhao S, Zeng H, et al. A quantitative and radiomics approach to monitoring ards in COVID-19 patients based on chest CT: A retrospective cohort study. Int J Med Sci [Internet]. Ivyspring International Publisher; 2020 [cited 2020 Oct 12];17:1773–82. Available from: /pmc/articles/PMC7378656/?report=abstract

56. Ghosh B, Kumar N, Sadhu AK, Ghosh N, Mitra P, Chatterjee J. A Quantitative Lung Computed Tomography Image Feature for Multi-Center Severity Assessment of COVID-19 [Internet]. medRxiv. Cold Spring Harbor Laboratory Press; 2020 [cited 2020 Oct 12]. Available from: https://doi.org/10.1101/2020.07.13.20152231

57. Schalekamp S, Huisman M, van Dijk RA, Boomsma MF, Freire Jorge PJ, de Boer W., et al. Model-based Prediction of Critical Illness in Hospitalized Patients with COVID-19. Radiology [Internet]. Radiological Society of North America (RSNA); 2020 [cited 2020 Dec 7];202723. Available from: http://pubs.rsna.org/doi/10.1148/radiol.2020202723

58. Xie C, Ng MY, Ding J, Leung ST, Lo CSY, Wong HYF, et al. Discrimination of pulmonary ground-glass opacity changes in COVID-19 and non-COVID-19 patients using CT radiomics analysis. Eur J Radiol Open [Internet]. Elsevier Ltd; 2020 [cited 2020 Dec 7];7. Available from: https://pubmed.ncbi.nlm.nih.gov/32959017/

59. Tamal M, Alshammari M, Alabdullah M, Hourani R, Alola HA, Hegazi TM. An Integrated Framework with Machine Learning and Radiomics for Accurate and Rapid Early Diagnosis of COVID-19 from Chest X-ray. medRxiv [Internet]. Cold Spring Harbor Laboratory Press; 2020 [cited 2020 Dec 7]; Available from: https://doi.org/10.1101/2020.10.01.20205146

60. Wang X, Hu X, Tan W, Mazzone P, Mireles-Cabodevila E, Han X-Z, et al. Multi-Center Study of Temporal Changes and Prognostic Value of a CT Visual Severity Score in Hospitalized Patients with COVID-19. Am J Roentgenol [Internet]. American Roentgen Ray Society; 2020 [cited 2020 Dec 7]; Available from: https://pubmed.ncbi.nlm.nih.gov/32903056/

61. Xu M, Ouyang L, Gao Y, Chen Y, Yu T, Li Q, et al. Accurately Differentiating COVID-19, Other Viral Infection, and Healthy Individuals Using Multimodal Features via Late Fusion Learning. medRxiv [Internet]. Cold Spring Harbor Laboratory Press; 2020 [cited 2020 Dec 7]; Available from: https://doi.org/10.1101/2020.08.18.20176776

62. Yip SSF, Klanecek Z, Naganawa S, Kim J, Studen A, Rivetti L, et al. Performance and Robustness of Machine Learning-based Radiomic COVID-19 Severity Prediction. medRxiv [Internet]. Cold Spring Harbor Laboratory Press; 2020 [cited 2020 Dec 7]; Available from: https://doi.org/10.1101/2020.09.07.20189977

63. Cohen JP, Morrison P, Dao L. COVID-19 Image Data Collection [Internet]. arXiv. 2020 [cited 2020 Jul 29]. Available from: http://arxiv.org/abs/2003.11597

64. RSNA Pneumonia Detection Challenge | Kaggle [Internet]. [cited 2020 Jul 29]. Available from: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge

65. Kermany, Daniel; Zhang, Kang; Goldbaum M. Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification [Internet]. Mendeley Data, v2. Mendeley; 2018 [cited 2020 Aug 12]. Available from: http://dx.doi.org/10.17632/rscbjbr9sj.2

66. Chest X-Ray Images (Pneumonia) | Kaggle [Internet]. [cited 2020 Jul 29]. Available from: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

67. COVID-19 Radiography Database | Kaggle [Internet]. [cited 2020 Jul 29]. Available from: https://www.kaggle.com/tawsifurrahman/covid19-radiography-database

68. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017. Institute of Electrical and Electronics Engineers Inc.; 2017. page 3462–71.

69. Chest X-RAY14 [Internet]. [cited 2020 Aug 12]. Available from: https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community

70. Wang L, Wong A. COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images [Internet]. arXiv. 2020 [cited 2020 Aug 12]. Available from: http://arxiv.org/abs/2003.09871

71. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. Proc AAAI Conf Artif Intell [Internet]. Association for the Advancement of Artificial Intelligence (AAAI); 2019 [cited 2020 Aug 12];33:590–7. Available from: www.aaai.org

Supplementary Table 1 – PRISMA Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	N/A
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	1
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	2
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	2
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	16
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	16
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	16
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	16
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	16
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	16
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	16

Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	16
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	N/A
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	N/A

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	N/A
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	N/A
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	2
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	2/3
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	3
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Table 2
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	N/A
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	N/A
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	N/A
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	5--14
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	15

Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	16
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	19

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097