nature machine intelligence

ANALYSIS

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

OPEN

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

Michael Roberts[1,2] ✉, Derek Driggs[1], Matthew Thorpe[3], Julian Gilbey[1], Michael Yeung[4], Stephan Ursprung[4,5], Angelica I. Aviles-Rivero[1], Christian Etmann[1], Cathal McCague[4,5], Lucian Beer[4], Jonathan R. Weir-McCall[4,6], Zhongzhao Teng[4], Effrossyni Gkrania-Klotsas[7], AIX-COVNET*, James H. F. Rudd[8,36], Evis Sala[4,5,36] and Carola-Bibiane Schönlieb[1,36]

Machine learning methods offer great promise for fast and accurate detection and prognostication of coronavirus disease 2019 (COVID-19) from standard-of-care chest radiographs (CXR) and chest computed tomography (CT) images. Many articles have been published in 2020 describing new machine learning-based models for both of these tasks, but it is unclear which are of potential clinical utility. In this systematic review, we consider all published papers and preprints, for the period from 1 January 2020 to 3 October 2020, which describe new machine learning models for the diagnosis or prognosis of COVID-19 from CXR or CT images. All manuscripts uploaded to bioRxiv, medRxiv and arXiv along with all entries in EMBASE and MEDLINE in this timeframe are considered. Our search identified 2,212 studies, of which 415 were included after initial screening and, after quality screening, 62 studies were included in this systematic review. Our review finds that none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases. This is a major weakness, given the urgency with which validated COVID-19 models are needed. To address this, we give many recommendations which, if followed, will solve these issues and lead to higher-quality model development and well-documented manuscripts.

In December 2019, a novel coronavirus was first recognized in Wuhan, China[1]. On 30 January 2020, as infection rates and deaths across China soared and the first death outside China was recorded, the World Health Organization (WHO) described the then-unnamed disease as a Public Health Emergency of International Concern[2]. The disease was officially named coronavirus disease 2019 (COVID-19) by 11 February 2020[3], and was declared a pandemic on 11 March 2020[4]. Since its first description in late 2019, the COVID-19 infection has spread across the globe, causing massive societal disruption and stretching our ability to deliver effective healthcare. This was caused by a lack of knowledge about the virus's behaviour along with a lack of an effective vaccine and antiviral therapies.

Although PCR with reverse transcription (RT–PCR) is the test of choice for diagnosing COVID-19, imaging can complement its use to achieve greater diagnostic certainty or even be a surrogate in some countries where RT–PCR is not readily available. In some cases, chest radiograph (CXR) abnormalities are visible in patients who initially had a negative RT–PCR test[5] and several studies have shown that chest computed tomography (CT) has a higher sensitivity for COVID-19 than RT–PCR, and could be considered as a primary tool for diagnosis[6–9]. In response to the pandemic, researchers have rushed to develop models using artificial intelligence (AI), in particular machine learning, to support clinicians.

Given recent developments in the application of machine learning models to medical imaging problems[10,11], there is fantastic promise for applying machine learning methods to COVID-19 radiological imaging for improving the accuracy of diagnosis, compared with the gold-standard RT–PCR, while also providing valuable insight for prognostication of patient outcomes. These models have the potential to exploit the large amount of multimodal data collected from patients and could, if successful, transform detection, diagnosis and triage of patients with suspected COVID-19. Of greatest potential utility is a model that can not only distinguish patients with COVID-19 from patients without COVID-19 but also discern alternative types of pneumonia such as those of bacterial or other viral aetiologies. With no standardization, AI algorithms for COVID-19 have been developed with a very broad range of applications, data collection procedures and performance assessment metrics. Perhaps as a result, none are currently ready to be deployed clinically. Reasons for this include: (1) the bias in small datasets; (2) the variability of large internationally sourced datasets; (3) the poor integration of multistream data, particularly imaging data; (4) the difficulty of the task of prognostication; and (5) the necessity for clinicians and data analysts to work side-by-side to ensure the developed AI algorithms are clinically relevant and implementable into routine clinical care. Since the pandemic began in early 2020, researchers have answered the 'call to arms' and numerous machine

[1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK. [2]Oncology R&D, AstraZeneca, Cambridge, UK. [3]Department of Mathematics, University of Manchester, Manchester, UK. [4]Department of Radiology, University of Cambridge, Cambridge, UK. [5]Cancer Research UK Cambridge Centre, University of Cambridge, Cambridge, UK. [6]Royal Papworth Hospital, Cambridge, Royal Papworth Hospital NHS Foundation Trust, Cambridge, UK. [7]Department of Infectious Diseases, Cambridge University Hospitals NHS Trust, Cambridge, UK. [8]Department of Medicine, University of Cambridge, Cambridge, UK. [36]These authors contributed equally: James H. F. Rudd, Evis Sala, Carola-Bibiane Schönlieb. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: michael.roberts@maths.cam.ac.uk

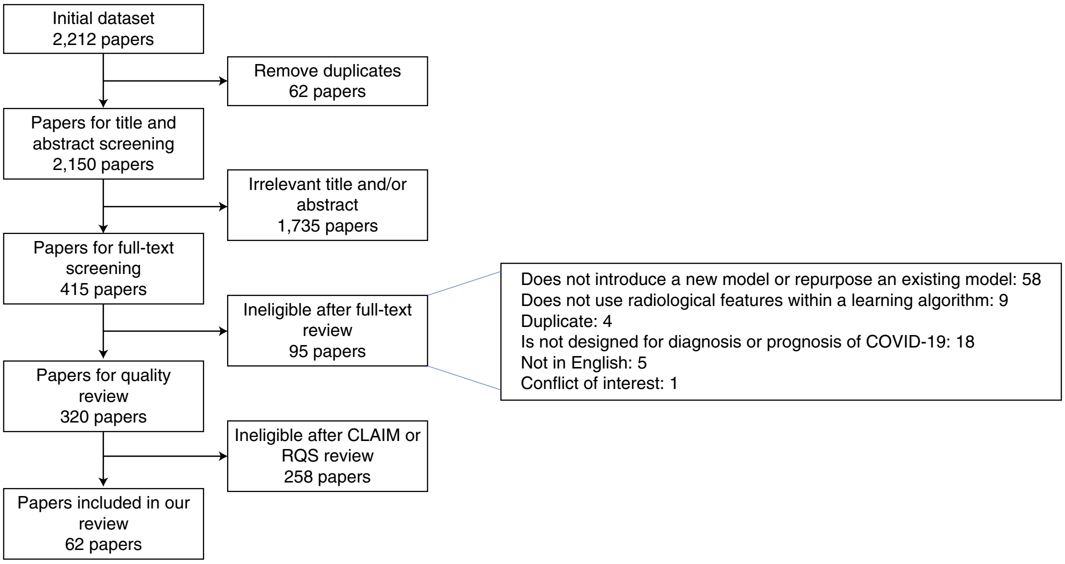
Initial dataset
2,212 papers
Remove duplicates
62 papers
Papers for title and abstract screening
2,150 papers
Irrelevant title and/or abstract
1,735 papers
Papers for full-text screening
415 papers
Ineligible after full-text review
95 papers
Does not introduce a new model or repurpose an existing model: 58
Does not use radiological features within a learning algorithm: 9
Duplicate: 4
Is not designed for diagnosis or prognosis of COVID-19: 18
Not in English: 5
Conflict of interest: 1
Papers for quality review
320 papers
Ineligible after CLAIM or RQS review
258 papers
Papers included in our review
62 papers

Fig. 1 | PRISMA flowchart for our systematic review. The flowchart highlights the inclusion and exclusion of papers at each stage.

learning models for diagnosis and prognosis of COVID-19 using radiological imaging have been developed and hundreds of manuscripts have been written. In this Analysis, we reviewed the entire literature of machine learning methods as applied to chest CT and CXR for the diagnosis and prognosis of COVID-19. As this is a rapidly developing field, we reviewed both published and preprint studies to ensure maximal coverage of the literature.

While earlier reviews provided a broad analysis of predictive models for COVID-19 diagnosis and prognosis[12–15], this Analysis highlights the unique challenges researchers face when developing classical machine learning and deep learning models using imaging data. This Analysis builds on the approach of Wynants et al.[12]: we assess the risk of bias in the papers considered, going further by incorporating a quality screening stage to ensure only those papers with sufficiently documented methodologies are reviewed in most detail. We also focus our review on the systematic methodological flaws in the current machine learning literature for COVID-19 diagnosis and prognosis models using imaging data. We also give detailed recommendations in five domains: (1) considerations when collating COVID-19 imaging datasets that are to be made public; (2) methodological considerations for algorithm developers; (3) specific issues about reproducibility of the results in the literature; (4) considerations for authors to ensure sufficient documentation of methodologies in manuscripts; and (5) considerations for reviewers performing peer review of manuscripts.

This Analysis has been performed, and informed, by both clinicians and algorithm developers, with our recommendations aimed at ensuring the most clinically relevant questions are addressed appropriately, while maintaining standards of practice to help researchers develop useful models and report reliable results even in the midst of a pandemic.

Results

Study selection. Our initial search highlighted 2,212 papers that satisfied our search criteria; removing duplicates we retained 2,150 papers, and of these, 415 papers had abstracts or titles deemed relevant to the review question, introducing machine learning methods for COVID-19 diagnosis or prognosis using radiological imaging. Full-text screening retained 320 papers, of which, after quality review, 62 were included for discussion in this Analysis (Fig. 1). Of these, 37 were deep learning papers, 23 were traditional machine learning papers and 2 were hybrid papers (using both approaches).

Quality screening failures. *Deep learning papers.* There were 254/320 papers that described deep learning-based models and 215 of these were excluded from the detailed review (including one hybrid paper). We found that 110 papers (51%) fail at least three of our identified mandatory criteria from the CLAIM checklist (Supplementary Discussion 1), with 23% failing two and 26% failing just one. In the rejected papers, the three most common reasons for a paper failing the quality check was due to insufficient documentation of the following.

(1) How the final model was selected in 61% (132)
(2) The method of pre-processing of the images in 58% (125)
(3) The details of the training approach (for example, the optimizer, the loss function, the learning rate) in 49% (105)

Traditional machine learning papers. There were 69 papers that described traditional machine learning methods and 44 of these were excluded from the review, that is, the RQS was less than 6 or the datasets used were not specified in the paper. There were only two papers that had an RQS ≥6, but which failed to disclose the datasets used in the analysis. Of the remaining papers, the two factors that lead to the lowest RQS results were omission of the following.

(1) Feature reduction techniques in 52% of papers (23)
(2) Model validation in 61% of papers (27)

The two hybrid papers both failed the checklist for artificial intelligence in medical imaging (CLAIM) check but passed the radiomic quality score (RQS) criteria. Full details can be found in Supplementary Data 2.

Remaining papers for detailed analysis. *Deep learning papers.* There were six non-mandatory CLAIM criteria not satisfied in at least half of the 37 papers.

(1) 29 did not complete any external validation
(2) 30 did not perform any robustness or sensitivity analysis of their model
(3) 26 did not report the demographics of their data partitions
(4) 25 did not report the statistical tests used to assess significance of results or determine confidence intervals
(5) 23 did not report confidence intervals for the performance
(6) 22 did not sufficiently report their limitations, biases or issues around generalizability

The full CLAIM results are in Supplementary Data 2.

Traditional machine learning papers. Of the 25 papers, including the two hybrid papers, none used longitudinal imaging, performed a prospective study for validation, or standardized image acquisition by using either a phantom study or a public protocol. Only six papers described performing external validation and only four papers reported the calibration statistics (the level of agreement between predicted risks and those observed) and associated statistical significance for the model predictions. The full RQS scores are in Supplementary Data 2.

Datasets considered. Public datasets were used extensively in the literature appearing in 32/62 papers (see Supplementary Discussion 2 for list of public datasets, three papers use both public and private data). Private data were used in 33/62 papers with 21 using data from mainland China, three using data from France and the remainder using data from Iran, the United States, Belgium, Brazil, Hong Kong and the Netherlands.

Diagnostic models for COVID-19. *Diagnosis models using CXRs.* Twenty-two papers considered diagnosis of COVID-19 from CXR images[16–36]. Most of these papers used off-the-shelf networks, including ResNet-18 or ResNet-50[16,17,20,26,29,32,37], DenseNet-121[27,28,31,32,34], VGG-16 or VGG-19[19,33,35], Inception[21,38] and EfficientNet[30,39], with three considering custom architectures[18,25,36] and three using hand-engineered features[22–24]. Most papers classified images into the three classes, that is, COVID-19, non-COVID-19 pneumonia and normal[16,19,21,23,25,26,28,30,32–37], while two considered an extra class by dividing non-COVID-19 pneumonia into viral and bacterial pneumonia[17,29]. ResNet and DenseNet architectures showed better performance than the others, with accuracies ranging from 0.88 to 0.99. However, we caution against direct comparison as the papers use different training and testing settings (for example, different datasets and data partition sizes) and consider a different number of classes.

Diagnostic models using CT scans and deep learning. Eighteen papers applied deep learning techniques to CT imaging, all of which were framed as a classification task to distinguish COVID-19 from other lung pathologies such as (viral or bacterial) pneumonia, interstitial lung disease[35,40–47] and/or a non-COVID-19 class[40,41,44,46,48–52]. The full three-dimensional (3D) volumes were only considered in seven papers[40,43,47,50,52–54] with the remainder considering isolated 2D slices or even 2D patches[45]. In most 2D models, authors employed transfer learning, with networks pre-trained on ImageNet[55]. Almost all models used lung segmentation as a pre-processing step. One paper[48] used a generative adversarial network[56] approach to address the paucity of COVID-19 CT imaging. Values reported for the area under the receiver operating characteristic curve (AUC) ranged from 0.70 to 1.00.

Diagnostic models using CT scans and traditional machine learning methods. Eight papers employed traditional machine learning methods for COVID-19 diagnosis using hand-engineered features[40,57–62] or convolutional neural network (CNN)-extracted features[46]. Four papers[46,59,60,62] incorporated clinical features with those obtained from the CT images. All papers using hand-engineered features employed feature reduction, using between 4 and 39 features in their final models. For final classification, five papers used logistic regression[40,58–61], one used a random forest[57], one a multilayer perceptron[46] and one compared many different machine learning classifiers to determine the best[62]. Accuracies ranged from 0.76 to 0.98 (refs. [40,46,57–59]). As before, we caution against direct comparison. The traditional machine-learning model in the hybrid paper[40] had a 0.05 lower accuracy than their deep learning model.

Prognostic models for COVID-19 using CT and CXR images. Nineteen papers developed models for the prognosis of patients with COVID-19[51,63–80], fifteen using CT and four using CXR. These models were developed for predicting severity of outcomes including: death or need for ventilation[72,78,79], a need for intensive care unit (ICU) admission[63,73,77–79], progression to acute respiratory distress syndrome[80], the length of hospital stay[51,74], likelihood of conversion to severe disease[64,65,75] and the extent of lung infection[76]. Most papers used models based on a multivariable Cox proportional hazards model[51,72,78,79], logistic regression[65,73–75,80], linear regression[75,76], random forest[74,77] or compare a huge variety of machine learning models such as tree-based methods, support vector machines, neural networks and nearest-neighbour clustering[63,64].

Predictors from radiological data were extracted using either handcrafted radiomic features[63,64,68–70,72,74,75,77–80] or deep learning[51,66,70,71,73,76]. Clinical data included basic observations, serology and comorbidities. Only eight models integrated both radiological and clinical data[62,63,69,72,73,77–79].

Risks of bias. Following the prediction model risk of bias assessment tool (PROBAST) guidance, the risk of bias was assessed for all 62 papers in four domains: participants, predictors, outcomes and analysis. The results are shown in Table 1. We found that 55/62 papers had a high risk of bias in at least one domain with the others unclear in at least one domain.

Participants. Almost all papers had a high (45/62) or unclear (11/62) risk of bias for their participants, with only six assessed as having a low risk of bias. This was primarily due to the following issues: (1) for public datasets it is not possible to know whether patients are truly COVID-19 positive, or if they have underlying selection biases, as anybody can contribute images[16,24,26,28–32,34,35,37,41,44,48,49,76]; (2) the paper uses only a subset of original datasets, applying some exclusion criteria, without enough details to be reproducible[16,43,44,48,49,51,61,70,71,75,76]; and/or (3) there are large differences in demographics between the COVID-19 cohort and the control groups, with, for example, paediatric patients as controls[17,24,28,29,31,32,35,37,45,46,59,81].

Predictors. For models where the features have been extracted using deep learning models, the predictors are unknown and abstract imaging features. Therefore, for these papers (38/62), we could not judge biases in the predictors. For 20 papers, the risk of bias was recorded as low due to the use of pre-defined hand-engineered features. For the remaining 4 papers, a high risk of bias was recorded due to the predictors being assessed with knowledge of the associated outcome.

Outcomes. The risk of bias in the outcome variable was found to be low for the majority (25/62) of the papers, unclear for 26/62 and high for 11/62. To evaluate the bias in the outcome, we took different approaches for papers using private datasets and public datasets (three papers use a mixture).

For the 35 papers that used public datasets, the outcome was assigned by the originators of the dataset and not by the papers' authors. Papers that used a public dataset generally have an unclear risk of bias (30/35) as they used the outcome directly sourced from the dataset originator.

For the 33 papers that used private datasets, the COVID-19 diagnosis was due to either positive RT–PCR or antibody tests for 24/33 and have a low risk of bias. The other papers have a high (7/33) or unclear (2/33) risk of bias due to inconsistent diagnosis of COVID-19[40,82], unclear definition of a control group[63,65], ground truths being assigned using the images themselves[26,54,60,71], using an unestablished reference to define outcome[74] or by combining public and private datasets[41,66,83].

Table 1 | PROBAST results for each domain considered for each paper included in our systematic review

Reference	Domain			
	Participants	Predictors	Outcomes	Analysis
Ghoshal and Tucker[17]	High	Unclear (DL)	Unclear	High
Li et al.[34]	Unclear	Unclear (DL)	Unclear	High
Ezzat et al.[28]	High	Unclear (DL)	Unclear	High
Tartaglione et al.[16]	High	Unclear (DL)	Unclear	High
Luz et al.[30]	Unclear	Unclear (DL)	Unclear	High
Bassi and Attux[31]	High	Unclear (DL)	Unclear	High
Gueguim Kana et al.[32]	High	Unclear (DL)	Unclear	High
Heidari et al.[33]	High	Unclear (DL)	Unclear	Unclear
Farooq and Hafeez[29]	High	Unclear (DL)	Unclear	High
Zhang et al.[27]	Low	Unclear (DL)	Low	Unclear
Zhang et al.[37]	High	Unclear (DL)	Unclear	High
Wang et al.[26]	High	Unclear (DL)	High	High
Bararia et al.[25]	High	Unclear (DL)	Unclear	High
Tsiknakis et al.[21]	High	Unclear (DL)	Unclear	High
Malhotra et al.[18]	High	High	Unclear	High
Sayyed et al.[36]	High	Low	Unclear	High
Rahaman et al.[19]	High	Unclear (DL)	Unclear	High
Amer et al.[20]	High	Unclear (DL)	Unclear	High
Elaziz et al.[22]	High	Low	Unclear	High
Tamal et al.[24]	High	High	Unclear	High
Gil et al.[23]	High	Low	Unclear	Unclear
Zokaeinikoo et al.[35]	High	Unclear (DL)	Unclear	Unclear
Amyar et al.[44]	High	Unclear (DL)	Unclear	High
Ardakani et al.[45]	High	Unclear (DL)	Low	High
Bai et al.[81]	High	Unclear (DL)	Low	Low
Jin et al.[50]	High	Unclear (DL)	Low	Unclear
Wang et al.[42]	High	Unclear (DL)	Low	Unclear
Ko et al.[41]	High	Unclear (DL)	High	Low
Acar et al.[48]	High	Unclear (DL)	High	Unclear
Pu et al.[43]	Unclear	Unclear (DL)	Low	Unclear
Chen et al.[49]	High	Unclear (DL)	Unclear	Unclear
Shah et al.[52]	High	Unclear (DL)	Unclear	High
Han et al.[47]	High	Unclear (DL)	Unclear	High
Wang et al.[53]	Unclear	Unclear (DL)	Low	Low
Wang et al.[54]	High	Unclear (DL)	High	Unclear
Goncharov et al.[71]	High	Unclear (DL)	High	Low
Xie et al.[61]	High	Low	Low	High
Xu et al.[62]	High	Low	Low	Unclear
Qin et al.[60]	Low	High	High	High
Georgescu et al.[40]	High	Unclear (DL)	High	Unclear
Guiot et al.[58]	High	Low	Low	Unclear
Shi et al.[57]	High	Low	Low	Low
Mei et al.[46]	High	Unclear (DL)	Low	High
Chen et al.[59]	High	Low	Low	High
Wang et al.[51]	Unclear	Unclear (DL)	Low	Low
Li et al.[66]	Low	Unclear (DL)	Unclear	High
Li et al.[67]	LOW	Unclear (DL)	Low	High

Continued

Table 1 | PROBAST results for each domain considered for each paper included in our systematic review

Reference	Domain			
	Participants	Predictors	Outcomes	Analysis
Schalekamp et al.[68]	Unclear	Low	Low	Low
Cohen et al.[76]	High	Unclear (DL)	Unclear	Unclear
Yue et al.[74]	High	Low	High	High
Zhu et al.[75]	High	Low	Low	Low
Lassau et al.[73]	High	Unclear (DL)	High	Unclear
Chassagnon et al.[63]	Unclear	Low	Low	Unclear
Chao et al.[77]	Low	Low	Low	High
Wu et al.[78]	Unclear	Low	Low	High
Zheng et al.[79]	Unclear	Low	Low	High
Chen et al.[80]	High	Low	Low	High
Ramtohul et al.[72]	Unclear	Low	Low	High
Ghosh et al.[64]	High	High	High	Low
Wei et al.[65]	Low	Low	High	High
Wang et al.[69]	Unclear	Low	Low	Low
Yip et al.[70]	High	Low	Unclear	Unclear

DL, deep learning.

Analysis. Only ten papers have a low risk of bias for their analysis. The high risk of bias in most papers was principally due to a small sample size of patients with COVID-19 (leading to highly imbalanced datasets), use of only a single internal holdout set for validating their algorithm (rather than cross-validation or bootstrapping) and a lack of appropriate evaluation of the performance metrics (for example, no discussion of calibration/discrimination)[18–20,22,23,44,48,52,64,72,80]. One paper with a high risk of bias[32] claimed external validation on the dataset from ref. [84], not realizing that this already includes datasets from both ref. [85] and ref. [86] that were used to train the algorithm.

Data analysis. There are two approaches for validating the performance of an algorithm, namely internal and external validation. For internal validation, the test data are from the same source as the development data and for external validation they are from different sources. Including both internal and external validation allows more insight to generalizability of the algorithm. We found that 48/62 papers consider internal validation only, with 13/62 using external validation[22,32,41,42,51,54,63,66,67,69,73,78,79]. Twelve used truly external test datasets and one tested on the same data the algorithm was trained on[32].

Model evaluation. In Table 2, we give the performance metrics quoted in each paper. Ten papers used cross-validation to evaluate model performance[21,35,36,47,49,57,65,72,74,75,77], one used both cross-validation and an external test set[41], one quoted correlation metrics[76] and one had an unclear validation method[17]. The other papers all had an internal holdout or external test set with sensitivity and specificity derived from the test data using an unquoted operating point (with the exception of ref. [16], which quotes operating point 0.5). It would be expected that an operating point be chosen based on the algorithm performance for the validation data used to tune and select the final algorithm. However, the receiver operating characteristic (ROC) curves and AUC values are given for the internal holdout or external test data independent of the validation data.

Partition analysis. In Fig. 2, we show the quantity of data (split by class) used in the training cohort of 32 diagnosis models. We excluded many studies[18,20,22,23,25,28,29,32,35,43,45,58,71] because it was unclear how many images were used. If a paper only stated the number of patients (and not the number of images), we assumed that there was only one image per patient. We see that 20/32 papers have a reasonable balance between classes (with exceptions being refs. [17,24,26,30,31,33,36,37,40,51,61,62]. However, the majority of datasets were quite small, with 19/32 papers using fewer than 2,000 datapoints for development (with exceptions being refs. [17,18,26,27,30,31,33,36,41,48,53,57,81]). Only seven papers used both a dataset with more than 2,000 datapoints that was balanced for COVID-19 positive and the other classes[27,41,48,53,54,57,81].

Figure 3 shows the number of images of each class used in the holdout/test cohorts. We found that 6/32 papers had an imbalanced testing dataset[17,24,33,36,37,61]. Only 6/32 papers tested on more than 1,000 images[17,27,36,41,54,81]. Only 4/32 had both a large and balanced testing dataset[27,41,54,81].

Public availability of the algorithms and models. Only 13/62 papers[21,23,27,30,34,36,46,51,66,67,74,76,81] published the code for reproducing their results (seven including their pre-trained parameters) and one stated that it is available on request[74].

Discussion

Our systematic review highlights the extensive efforts of the international community to tackle the COVID-19 pandemic using machine learning. These early studies show promise for diagnosis and prognostication of pneumonia secondary to COVID-19. However, we have also found that current reports suffer from a high prevalence of deficiencies in methodology and reporting, with none of the reviewed literature reaching the threshold of robustness and reproducibility essential to support utilization in clinical practice. Many studies are hampered by issues with poor-quality data, poor application of machine learning methodology, poor reproducibility and biases in study design. The current paper complements the work of Wynants et al. who have published a living systematic review[12] on publications and preprints of studies describing multivariable models for screening of COVID-19 infections in the general population, differential diagnosis of COVID-19 infection in patients that are symptomatic and prognostication in patients with

Table 2 | Summary of the data extracted for each paper included in our systematic review

Reference	Diagnosis/ prognosis	Data used in model	Predictors	Sample size development	Sample size test	Type of validation	Evaluation	Public code
	Is this paper describing a COVID-19 diagnosis or prognosis model (or both)?	**Does this use CXR or CT (or both)?**	**What are the predictors? In purely deep learning models, this is DL.**	**Total sample size used for development (that is, training and validation and NOT test set), along with number of positive outcomes.**	**Total sample size used for testing of the algorithm, along with the number of positive outcomes.**	***k*-fold CV, external validation in *k* centres, no validation and so on**	**Performance of the model, AUC, confidence interval, sensitivity, specificity and so on. 95% CI if available.**	**Is there code available? (Is the trained model available?)**
Ghoshal and Tucker[17]	Diagnosis	CXR	DL	4,752 images, 54 COVID-19	1,189 images, 14 COVID-19	Unclear validation procedure	Unclear in the paper	No
Li et al.[34]	Diagnosis	CXR	DL	429 images, 143 COVID-19	108 images, 36 COVID-19	Internal holdout validation	Accuracy, 0.880; AUC, 0.970	Yes (Yes)
Ezzat et al.[28]	Diagnosis	CXR	DL	Unclear in the paper	Unclear in the paper	Internal holdout validation	Precision (*w*), 0.98; recall (*w*), 0.98; F1 score (*w*), 0.98	No
Tartaglione et al.[16]	Diagnosis	CXR	DL	231 images, 126 COVID-19	135 images, 90 COVID-19	Internal holdout validation	Unclear in the paper	No
Luz et al.[30]	Diagnosis	CXR	DL	13,569 images, 152 COVID-19	231 images, 31 COVID-19	Internal holdout validation	Accuracy, 0.94; sensitivity, 0.97; PPV, 1.00	Yes (Yes)
Bassi and Attux[31]	Diagnosis	CXR	DL	2,724 images, 159 COVID-19	180 images, 60 COVID-19	Internal holdout validation	Recall, 0.98; precision, 1.00	No
Gueguim Kana et al.[32]	Diagnosis	CXR	DL	Unclear in the paper	Unclear in the paper	External validation	Accuracy, 0.99; recall, 1.00; precision, 0.99; F1 score, 1.00	No
Heidari et al.[33]	Diagnosis	CXR	DL	8,474 images, 415 COVID-19	848 images,42 COVID-19	Internal holdout validation	Precision (*w*), 0.95; recall (*w*), 0.94; F1 score (*w*), 0.94	No
Farooq and Hafeez[29]	Diagnosis	CXR	DL	Unclear in the paper	637 images, 8 COVID-19	Internal holdout validation	Accuracy, 0.96; sensitivity, 0.97; PPV, 0.99; F1 score, 0.98	No
Zhang et al.[27]	Diagnosis	CXR	DL	5,236 images, 2,582 COVID-19	5,869 images, 3,223 COVID-19	Internal holdout validation	AUC, 0.92; sensitivity, 0.88; specificity, 0.79	Yes (No)
Zhang et al.[37]	Diagnosis	CXR	DL	386 images, 150 COVID-19	101 images, 39 COVID-19	Internal holdout validation	Accuracy, 0.91	No
Wang et al.[26]	Diagnosis	CXR	DL	3,522 images, 204 COVID-19	61 images, 20 COVID-19	Internal holdout validation	AUC, 1.00; accuracy, 0.99	No
Bararia et al.[25]	Diagnosis	CXR	DL	Unclear in the paper	1,000 images, 341 COVID-19	Internal holdout validation	Accuracy, 0.81; sensitivity, 0.81; specificity, 0.90; precision, 0.74; recall, 0.77; F1 score, 0.75	No
Tsiknakis et al.[21]	Diagnosis	CXR	DL	458 (CV) images, 98 COVID-19	114 (CV) images, 24 COVID-19	Fivefold internal cross-validation	AUC, 1.00; accuracy, 1.00; sensitivity, 0.99; specificity, 1.00	Yes (No)

Continued

Table 2 | Summary of the data extracted for each paper included in our systematic review (continued)

Reference	Diagnosis/ prognosis	Data used in model	Predictors	Sample size development	Sample size test	Type of validation	Evaluation	Public code
	Is this paper describing a COVID-19 diagnosis or prognosis model (or both)?	**Does this use CXR or CT (or both)?**	**What are the predictors? In purely deep learning models, this is DL.**	**Total sample size used for development (that is, training and validation and NOT test set), along with number of positive outcomes.**	**Total sample size used for testing of the algorithm, along with the number of positive outcomes.**	***k*-fold CV, external validation in *k* centres, no validation and so on**	**Performance of the model, AUC, confidence interval, sensitivity, specificity and so on. 95% CI if available.**	**Is there code available? (Is the trained model available?)**
Malhotra et al.[18]	Diagnosis	CXR	DL	26,464 images, 1,740 COVID-19[a]	6,299 images, 125 COVID-19[a]	Internal holdout validation	Sensitivity, 0.87; specificity, 0.97	No
Sayyed et al.[36]	Diagnosis	CXR	DL	5,018 (CV) images, 334 COVID-19	1,255 (CV) images, 83 COVID-19	Fivefold internal cross-validation	Accuracy, 0.99 ± 0.05	Yes (No)
Rahaman et al.[19]	Diagnosis	CXR	DL	720 images, 220 COVID-19	140 images, 40 COVID-19	Internal holdout validation	Accuracy, 0.89; precision, 0.90; recall, 0.89; F1 score, 0.90	No
Amer et al.[20]	Diagnosis	CXR	DL	Unclear in the paper	Unclear in the paper	Internal holdout validation	AUC, 0.98; accuracy, 0.94; sensitivity, 0.92; specificity, 0.97; PPV, 0.98	No
Elaziz et al.[22]	Diagnosis	CXR	Hand-engineered radiomic features	Unclear in the paper	Unclear in the paper	Internal holdout validation and external validation	Internal validation: accuracy, 0.96; recall, 0.99; precision, 0.96 External validation: accuracy, 0.98; recall, 0.99; precision, 0.99	No
Tamal et al.[24]	Diagnosis	CXR	Hand-engineered radiomic features.	378 images, 226 COVID-19	165 images, 115 COVID-19	Internal holdout validation	Sensitivity, 1.00; specificity, 0.85	No[b]
Gil et al.[23]	Diagnosis	CXR	Hand-engineered radiomic features	Unclear in the paper	Unclear in the paper	Internal holdout validation	Accuracy, 0.96; sensitivity, 0.98; specificity, 0.93; precision, 0.96	Yes (Yes)
Zokaeinikoo et al.[35]	Diagnosis	CXR and CT	DL	Unclear in the paper	Unclear in the paper	Tenfold internal cross-validation	Accuracy, 0.99; sensitivity, 0.99; specificity, 1.00; PPV, 1.00	No
Amyar et al.[44]	Diagnosis	CT	DL	944 patients, 399 COVID-19	100 patients, 50 COVID-19	Internal holdout validation	Accuracy, 0.95; sensitivity, 0.96; specificity, 0.92; AUC, 0.97	No
Ardakani et al.[45]	Diagnosis	CT	DL	Unclear as splits do not total correctly	Unclear as splits do not total correctly	Internal holdout validation	AUC, 0.99; sensitivity, 1.00; specificity, 0.99; accuracy, 1.00; PPV, 0.99; NPV, 1.00	No

Continued

Table 2 | Summary of the data extracted for each paper included in our systematic review (continued)

Reference	Diagnosis/ prognosis	Data used in model	Predictors	Sample size development	Sample size test	Type of validation	Evaluation	Public code
	Is this paper describing a COVID-19 diagnosis or prognosis model (or both)?	Does this use CXR or CT (or both)?	What are the predictors? In purely deep learning models, this is DL.	Total sample size used for development (that is, training and validation and NOT test set), along with number of positive outcomes.	Total sample size used for testing of the algorithm, along with the number of positive outcomes.	*k*-fold CV, external validation in *k* centres, no validation and so on	Performance of the model, AUC, confidence interval, sensitivity, specificity and so on. 95% CI if available.	Is there code available? (Is the trained model available?)
Bai et al.[81]	Diagnosis	CT	DL	118,401 images, 60,776 COVID-19	14,182 images, 5,040 COVID-19	Internal holdout validation	AUC, 0.95; accuracy, 0.96; sensitivity, 0.95; specificity, 0.96	Yes (Yes)
Jin et al.[50]	Diagnosis	CT	DL	1,136 images, 723 COVID-19	282 images, 154 COVID-19	Internal holdout validation	Sensitivity, 0.97; specificity, 0.92; AUC, 0.99	No
Wang et al.[42]	Diagnosis	CT	DL	320 images, 160 COVID-19	Internal validation: 455 images, 95 COVID-19 External validation: 290 images, 70 COVID-19	Internal holdout validation and external validation	Internal validation: AUC, 0.93 [0.90, 0.96] External validation: AUC, 0.81 [0.71, 0.84]	No
Ko et al.[41]	Diagnosis	CT	DL	3,194 (CV) images, 955 COVID-19	Internal cross-validation: 799 (CV) images, 239 COVID-19 External validation: 264 images, all COVID-19	Fivefold internal cross-validation and external validation	Internal validation: AUC, 1.00; accuracy, 1.00; sensitivity, 1.00; specificity, 1.00 External validation: accuracy, 0.97	No
Acar et al.[48]	Diagnosis	CT	DL	2,552 images, 1,085 COVID-19	580 images, 246 COVID-19	Internal holdout validation	AUC, 1.00; accuracy, 1.00; error, 0.01; precision, 1.00; recall, 1.00; F1 score, 1.00	No
Pu et al.[43]	Diagnosis	CT	DL	Unclear in the paper	Unclear in the paper	Internal holdout validation	AUC, 0.70 [0.56, 0.85]; sensitivity, 0.98; specificity, 0.28	No
Chen et al.[49]	Diagnosis	CT	DL	770 (CV) images, 413 COVID-19	Internal cross-validation: 86 (CV) images, 46 COVID-19	Tenfold internal cross-validation	AUC, 0.94 ± 0.01; accuracy, 0.88 ± 0.01; precision, 0.90 ± 0.01; recall, 0.88 ± 0.01	No
Shah et al.[52]	Diagnosis	CT	DL	664 images, 314 COVID-19	74 images, 35 COVID-19	Internal holdout validation	Accuracy, 0.95	No
Han et al.[47]	Diagnosis	CT	DL	368 (CV) images, 184 COVID-19	92 (CV) images, 46 COVID-19	Fivefold internal cross-validation	AUC, 0.99; accuracy, 0.98	No[b]
Wang et al.[53]	Diagnosis	CT	DL	3,997 images, 1,095 COVID-19	600 images, 200 COVID-19	Internal holdout validation	AUC, 0.97; accuracy, 0.93; specificity, 0.96; precision, 0.88; recall, 0.88	No

Continued

Table 2 | Summary of the data extracted for each paper included in our systematic review (continued)

Reference	Diagnosis/ prognosis	Data used in model	Predictors	Sample size development	Sample size test	Type of validation	Evaluation	Public code
	Is this paper describing a COVID-19 diagnosis or prognosis model (or both)?	**Does this use CXR or CT (or both)?**	**What are the predictors? In purely deep learning models, this is DL.**	**Total sample size used for development (that is, training and validation and NOT test set), along with number of positive outcomes.**	**Total sample size used for testing of the algorithm, along with the number of positive outcomes.**	***k*-fold CV, external validation in *k* centres, no validation and so on**	**Performance of the model, AUC, confidence interval, sensitivity, specificity and so on. 95% CI if available.**	**Is there code available? (Is the trained model available?)**
Wang et al.[54]	Diagnosis	CT	DL	2,447 images, 1,647 COVID-19	Internal validation: 639 images, 439 COVID-19 External validation: 2,120 images, 217 COVID-19	Internal holdout and external validation	Internal validation: AUC, 0.99; sensitivity, 0.97; specificity, 0.85 External validation: AUC, 0.95; sensitivity: 0.92; specificity, 0.85	No
Goncharov et al.[71]	Diagnosis and severity prognosis	CT	DL	Unclear in the paper	Diagnosis: 101 images, 33 COVID-19 Severity: 38 images of differing severity	Internal holdout validation	Diagnosis model: AUC, 0.95 Severity model: correlation, 0.98	No[c]
Xie et al.[61]	Diagnosis	CT	Hand-engineered radiomic features	225 images, 27 COVID-19	76 images, 6 COVID-19	Internal holdout validation	AUC, 0.91; accuracy, 0.90; sensitivity, 0.83; specificity, 0.90	No
Xu et al.[62]	Diagnosis	CT	DL and hand-engineered radiomic features	551 images, 289 COVID-19	138 images, 73 COVID-19	Internal holdout validation	Accuracy, 0.98; F1 score, 0.99	No[d]
Qin et al.[60]	Diagnosis	CT	Hand-engineered radiomic features	118 patients, 62 COVID-19	50 patients, 26 COVID-19	Internal holdout validation	AUC, 0.85 [0.74, 0.96]; sensitivity, 0.89; specificity, 0.92	No
Georgescu et al.[40]	Diagnosis	CT	DL and hand-engineered radiomic features	1,902 patients, 1,050 COVID-19	194 patients, 100 COVID-19	Internal holdout validation	AUC, 0.90; sensitivity, 0.86; specificity, 0.81	No
Guiot et al.[58]	Diagnosis	CT	Hand-engineered radiomic features	Unclear in the paper	Unclear in the paper	Internal holdout validation	AUC, 0.94 [0.88, 1.00]; accuracy, 0.90 [0.84, 0.94]; sensitivity, 0.79; specificity, 0.91	No
Shi et al.[57]	Diagnosis	CT	Hand-engineered radiomic features	2,148 (CV) images, 1,326 COVID-19	Internal cross-validation: 537 (CV) images, 332 COVID-19	Fivefold internal cross-validation	AUC, 0.94; accuracy, 0.88; sensitivity, 0.91; specificity, 0.83	No
Mei et al.[46]	Diagnosis	CT	DL and CNN extracted features and clinical data	626 images, 285 COVID-19	279 images, 134 COVID-19	Internal holdout validation	AUC, 0.92 [0.89, 0.95]; sensitivity, 0.843 [0.77, 0.90]; specificity, 0.83 [0.76, 0.89]	Yes (Yes)

Continued

Table 2 | Summary of the data extracted for each paper included in our systematic review (continued)

Reference	Diagnosis/ prognosis	Data used in model	Predictors	Sample size development	Sample size test	Type of validation	Evaluation	Public code
	Is this paper describing a COVID-19 diagnosis or prognosis model (or both)?	**Does this use CXR or CT (or both)?**	**What are the predictors? In purely deep learning models, this is DL.**	**Total sample size used for development (that is, training and validation and NOT test set), along with number of positive outcomes.**	**Total sample size used for testing of the algorithm, along with the number of positive outcomes.**	***k*-fold CV, external validation in *k* centres, no validation and so on**	**Performance of the model, AUC, confidence interval, sensitivity, specificity and so on. 95% CI if available.**	**Is there code available? (Is the trained model available?)**
Chen et al.[59]	Diagnosis	CT	Clinical features, qualitative imaging features and hand-engineered radiomic imaging features	98 patients, 51 COVID-19	38 images, 19 COVID-19	Internal holdout validation	AUC, 0.94 [0.87, 1.00]; accuracy, 0.76; sensitivity, 0.74; specificity, 0.79	No
Wang et al.[51]	Diagnosis and prognosis for length of hospital stay	CT	Diagnosis model: DL Prognosis model: 64 CNN features and clinical factors	709 images, 560 COVID-19	Validation 1: 226 images, 102 COVID-19 Validation 2: 161 images, 92 COVID-19 Validation 3: 53 images, all with length of hospital stay Validation 4: 117 images, all with length of hospital stay	External validation	Validation 1 (diagnosis): AUC, 0.87 Validation 2 (diagnosis): AUC, 0.88 Validation 3 (prognosis): KM separation, $P=0.01$ Validation 4 (prognosis): KM separation, $P=0.01$	Yes (Yes)
Li et al.[66]	Prognosis for severity	CXR	DL	354 images of differing severities	Internal validation: 108 images External validation: 111 images	Internal holdout validation and external validation	Internal validation: correlation, 0.88 External validation: correlation, 0.90	Yes (No)
Li et al.[67]	Prognosis for severity	CXR	DL	314 images of differing severities	Internal validation: 154 images External validation: 113 images	Internal holdout validation and external validation	Internal validation: correlation, 0.86 External validation: correlation, 0.86	Yes (No)
Schalekamp et al.[68]	Prognosis for severity	CXR	Hand-engineered radiomic features and clinical factors	Unclear in the paper	Unclear in the paper	Internal holdout validation	AUC, 0.77	No
Cohen et al.[76]	Prognosis of lung opacity and extent of lung involvement with GGOs for patients with COVID-19	CXR	Features from a trained CNN extracted at various layers	47 patients of varying severity	47 patients of varying severity	Internal holdout validation	Opacity correlation, 0.80; extent correlation, 0.78	Yes (Yes)

Continued

Table 2 | Summary of the data extracted for each paper included in our systematic review (continued)

Reference	Diagnosis/ prognosis	Data used in model	Predictors	Sample size development	Sample size test	Type of validation	Evaluation	Public code
	Is this paper describing a COVID-19 diagnosis or prognosis model (or both)?	Does this use CXR or CT (or both)?	What are the predictors? In purely deep learning models, this is DL.	Total sample size used for development (that is, training and validation and NOT test set), along with number of positive outcomes.	Total sample size used for testing of the algorithm, along with the number of positive outcomes.	*k*-fold CV, external validation in *k* centres, no validation and so on	Performance of the model, AUC, confidence interval, sensitivity, specificity and so on. 95% CI if available.	Is there code available? (Is the trained model available?)
Yue et al.[74]	Prognosing short- and long-term (>10 days) hospital stay for patients with COVID-19	CT	Hand-engineered radiomic features	26 patients, 16 long term	Internal validation: 5 patients, 3 long term Temporal-split internal validation: 6 patients, all long term	Internal holdout and temporal-split validation	AUC, 0.97 [0.83,1.00]; sensitivity, 1.00; specificity, 0.89; NPV, 1.00; PPV, 0.80	Yes[d]
Zhu et al.[75]	The prognosis for whether patients will convert to a severe stage of COVID-19 and regression to predict the time to that conversion	CT	Hand-engineered radiomic features	Unclear in the paper	Unclear in the paper	Fivefold internal cross-validation run 20 times, average reported	AUC, 0.86 ± 0.02; accuracy, 0.86 ± 0.02; sensitivity, 0.77 ± 0.03; specificity, 0.88 ± 0.015	No
Lassau et al.[73]	The prognostic model used for predicting the risk of death, need for ventilation or requirement for over 15 l min^{-1} oxygen	CT	CNN extracted features and clinical data	646 patients, all with COVID-19; 243 with severe outcomes	Internal validation: 150 images, all COVID-19, 48 with severe outcome External validation: 135 patients, all with COVID-19, unclear number of severe patients	Internal holdout validation and external validation	Internal validation: AUC, 0.76 External validation: AUC, 0.75	No[c]
Chassagnon et al.[63]	Short-term prognosis intubation and death within four days Long-term prognosis: death within one month after CT	CT	Hand-engineered radiomic features and clinical data	536 patients with COVID-19, 108 severe short-term outcomes, unclear for long term	157 patients with COVID-19, 31 severe short-term outcomes, unclear for long term	External validation	Short-term prognosis: precision (*w*), 0.94; sensitivity (*w*), 0.94; specificity (*w*), 0.81; balanced accuracy, 0.88 Long-term prognosis: precision (*w*), 0.77; sensitivity (*w*), 0.94; specificity (*w*), 0.82; balanced accuracy, 0.71	No[b]

Continued

Table 2 | Summary of the data extracted for each paper included in our systematic review (continued)

Reference	Diagnosis/ prognosis	Data used in model	Predictors	Sample size development	Sample size test	Type of validation	Evaluation	Public code
	Is this paper describing a COVID-19 diagnosis or prognosis model (or both)?	Does this use CXR or CT (or both)?	What are the predictors? In purely deep learning models, this is DL.	Total sample size used for development (that is, training and validation and NOT test set), along with number of positive outcomes.	Total sample size used for testing of the algorithm, along with the number of positive outcomes.	*k*-fold CV, external validation in *k* centres, no validation and so on	Performance of the model, AUC, confidence interval, sensitivity, specificity and so on. 95% CI if available.	Is there code available? (Is the trained model available?)
Chao et al.[77]	Prognosing for ICU admission	CT	Hand-engineered radiomic features and clinical data	236 (CV) images, 125 admitted to ICU	59 (CV) images, 31 admitted to ICU	Fivefold internal cross-validation	Unclear in the paper	No
Wu et al.[78]	Prognosing for death, ventilation and ICU admission in early- and late-stage COVID-19	CT	Hand-engineered radiomic features	351 images, 25 severe outcomes	141 images, 26 severe outcomes	External validation	Early-stage COVID-19: AUC, 0.86; sensitivity, 0.80; specificity, 0.86 Late-stage COVID-19: AUC, 0.98; sensitivity, 1.00; specificity, 0.94	No
Zheng et al.[79]	Prognosing for admission to an ICU, use of mechanical ventilation or death	CT	Hand-engineered radiomic features and clinical data	166 images, 35 severe outcomes	72 images, 10 severe outcomes	External validation	*C* index, 0.89	No
Chen et al.[80]	Prognosis for acute respiratory distress syndrome	CT	Hand-engineered radiomic features and clinical data	247 images, 36 severe cases	105 images, 15 severe cases	Internal holdout validation	Accuracy, 0.88; sensitivity, 0.55; specificity, 0.95	No
Ghosh et al.[64]	Prognosing COVID-19 severity	CT	Hand-engineered radiomic features	36 images, unclear number of severe cases	24 images, unclear number of severe cases	Internal holdout validation	Accuracy, 0.88	No
Ramtohul et al.[72]	Prognosing mortality for patients with COVID-19 in a cancer population	CT	Hand-engineered radiomic features and clinical data	35 (CV) patients, unclear number of deaths	70 patients, unclear number of deaths	Twofold internal cross-validation	*C* index, 0.83 [0.73, 0.93]	No
Wei et al.[65]	Prognosing COVID-19 severity	CT	Hand-engineered radiomic features	Unclear in the paper	Unclear in the paper	One-hundred-fold leave-group-out cross-validation	AUC, 0.93 accuracy, 0.91; sensitivity, 0.81; specificity, 0.95	No
Wang et al.[69]	Prognosis for survival	CT	Hand-engineered radiomic features	161 patients, 15 non-survivors	135 patients, unclear number of non-survivors	External validation	*C* index, [0.92, 0.95]; accuracy, [0.85, 0.87]; sensitivity, [0.71, 0.76]; specificity, [0.91, 0.92]	No
Yip et al.[70]	Prognosing COVID-19 severity	CT	Hand-engineered radiomic features	657 images of various severities	441 images of various severities	Internal holdout validation	AUC, 0.85	No

[a]Number of samples after augmentation, the original number of COVID-19 images is unclear. [b]The authors state that the algorithm will be made publicly available. [c]The paper states that code 'is available on a public GitHub repository' but no link is provided and the authors could not locate it. [d]The authors state that 'imaging or algorithm data used in this study are available upon request'. *w*, weighted average; CV, cross-validation; CI, 95% confidence interval; PPV, positive predictive value; NPV, negative predictive value; KM, Kaplan–Meier; GGOs, ground-glass opacities.

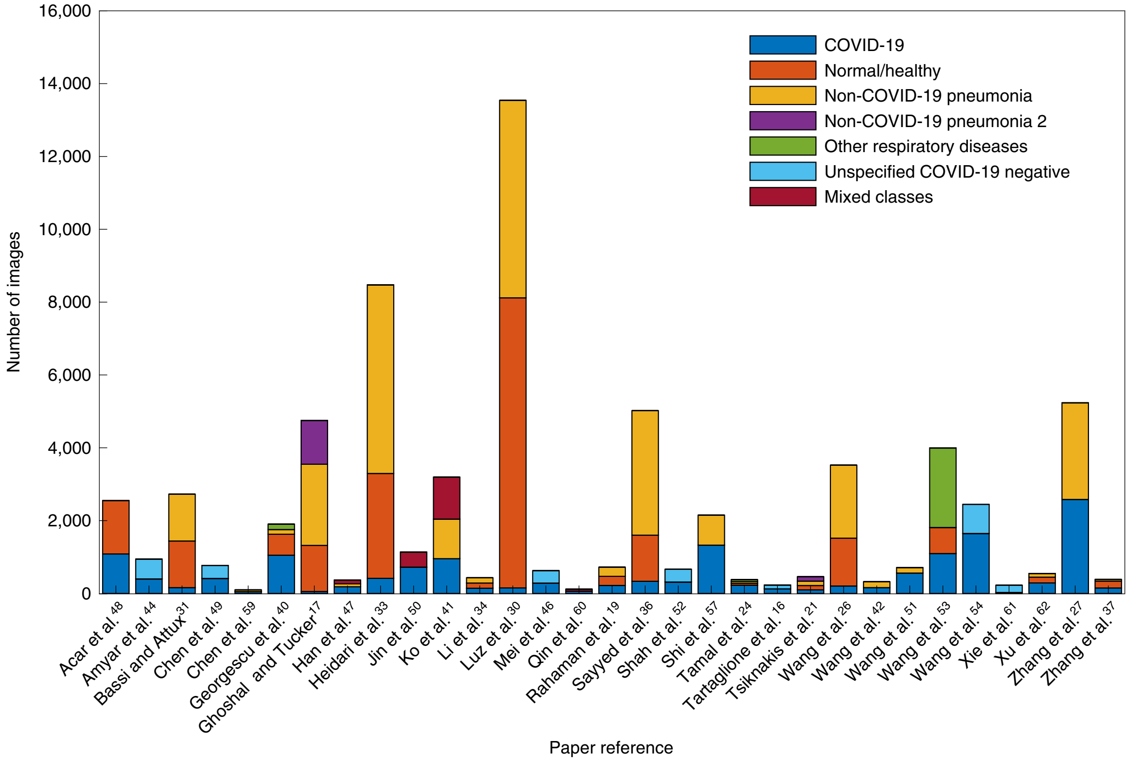

Fig. 2 | The number of images used in each paper for model training split by image class. It is noted that we exclude Bai et al.[81] from the figure as they used much more training data (118,401 images) than other papers. For Xu et al.[62], two COVID-19 classes are shown in the graph as one combined class.

this systematic review employed specialized quality metrics for the assessment of radiomics and deep learning-based diagnostic models in radiology. This is also in contrast to previous studies that have assessed AI algorithms in COVID-19[13,14]. Limitations of the current literature most frequently reflect either a limitation of the dataset used in the model or methodological mistakes repeated in many studies that probably lead to overly optimistic performance evaluations.

Datasets. Many papers gave little attention to establishing the original source of the images (Supplementary Discussion 2). When considering papers that use public data, readers should be aware of the following.

Duplication and quality issues. There is no restriction for a contributor to upload COVID-19 images to many of the public repositories[85,87–90]. There is high likelihood of duplication of images across these sources and no assurance that the cases included in these datasets are confirmed COVID-19 cases (authors take a great leap to assume this is true) so great care must be taken when combining datasets from different public repositories. Also, most of the images have been pre-processed and compressed into non-DICOM formats leading to a loss in quality and a lack of consistency/comparability.

- Source issues. Many papers (16/62) used the pneumonia dataset of Kermany et al.[86] as a control group. They commonly failed to mention that this consists of paediatric patients aged between one and five. Developing a model using adult patients with COVID-19 and very young patients with pneumonia is likely to overperform as it is merely detecting children versus adults. This dataset is also erroneously referred to as the Mooney dataset in many papers (being the Kermany dataset deployed on Kaggle[91]). It is also important to consider the sources of each image class, for example, if images for different diagnoses are from different sources. It is demonstrated by Maguolo et al.[92] that by excluding the lung region entirely, the authors could identify the source of the images in the Cohen et al.[85] and Kermany et al.[86] datasets with an AUC between 0.9210 and 0.9997, and 'diagnose' COVID-19 with an AUC = 0.68.
- Frankenstein datasets. The issues of duplication and source become compounded when public 'Frankenstein' datasets are used, that is, datasets assembled from other datasets and redistributed under a new name. For instance, one dataset[91] combined several other datasets[85,88,93] without realizing that one of the component datasets[93] already contains another component[88]. This repackaging of datasets, although pragmatic, inevitably leads to problems with algorithms being trained and tested on identical or overlapping datasets while believing them to be from distinct sources.
- Implicit biases in the source data. Images uploaded to a public repository and those extracted from publications[93] are likely to have implicit biases due to the contribution source. For example, it is likely that more interesting, unusual or severe cases of COVID-19 appear in publications.

Methodology. All proposed models suffer from a high or unclear risk of bias in at least one domain. There are several methodological issues driven by the urgency in responding to the COVID-19 crisis and subtler sources of bias due to poor application of machine learning.

The urgency of the pandemic led to many studies using datasets that contain obvious biases or are not representative of the target population, for example, paediatric patients. Before evaluating a model, it is crucial that authors report the demographic statistics for their datasets, including age and sex distributions. Diagnostic studies commonly compare their models' performance to that of RT–PCR. However, as the ground-truth labels are often determined

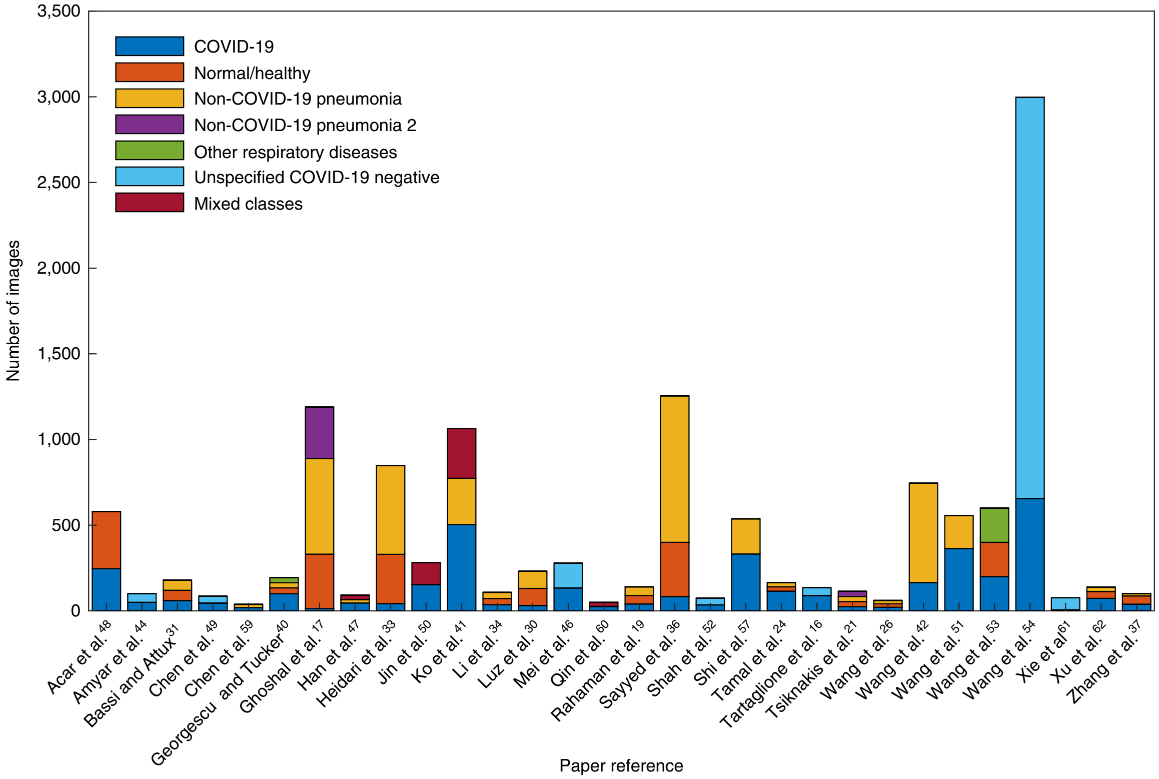

Fig. 3 | The number of images used for model testing split by image class. It is noted that we exclude Bai et al.[81] and Zhang et al.[27] from the figure as they used far more testing data (14,182 and 5,869 images respectively) than other papers. There were a large number of images (1,237) in the testing dataset in Wang et al.[54] that were unidentified in the paper (we include these in the unspecified COVID-19 negative).

by RT–PCR, there is no way to measure whether a model outperforms RT–PCR from accuracy, sensitivity or specificity metrics alone. Ideally, models should aim to match clinicians using all available clinical and radiomic data, or to aid them in decision-making.

Many papers utilized transfer learning in developing their model, which assumes an inherent benefit to performance. However, it is unclear whether transfer learning offers a large performance benefit due to the over-parameterization of the models[41,58]. Many publications used the same resolutions such as 224-by-224 or 256-by-256 for training, which are often used for ImageNet classification, indicating that the pre-trained model dictated the image rescaling used rather than clinical judgement.

Recommendations. Based on the systematic issues we encountered in the literature, we offer recommendations in five distinct areas: (1) the data used for model development and common pitfalls; (2) the evaluation of trained models; (3) reproducibility; (4) documentation in manuscripts; and (5) the peer-review process. Our recommendations in areas (3) and (4) are largely informed by the 258 papers that did not pass our initial quality check, while areas (1), (2) and (5) follow from our analysis of the 62 papers receiving our full review.

Recommendations for data. First, we advise caution over the use of public repositories, which can lead to high risks of bias due to source issues and Frankenstein datasets as discussed above. Furthermore, authors should aim to match demographics across cohorts, an often neglected but important potential source of bias; this can be impossible with public datasets that do not include demographic information, and including paediatric images[86] in the COVID-19 context introduces a strong bias.

Using a public dataset alone without additional new data can lead to community-wide overfitting on this dataset. Even if each individual study observes sufficient precautions to avoid overfitting, the fact that the community is focused on outperforming benchmarks on a single public dataset encourages overfitting. Many public datasets containing images taken from preprints receive these images in low-resolution or compressed formats (for example, JPEG and PNG), rather than their original DICOM format. This loss of resolution is a serious concern for traditional machine learning models if the loss of resolution is not uniform across classes, and the lack of DICOM metadata does not allow exploration of model dependence on image acquisition parameters (for example, scanner manufacturer, slice thickness and so on).

Regarding CXRs, researchers should be aware that algorithms might associate more severe disease not with CXR imaging features, but the view that has been used to acquire that CXR. For example, for patients that are sick and immobile, an anteroposterior CXR view is used for practicality rather than the standard posteroanterior CXR projection. Also, overrepresentation of severe disease is bad not only from the machine learning perspective but also in terms of clinical utility, as the most useful algorithms are those that can diagnose disease at an early stage[94]. The timing between imaging and RT–PCR tests was also largely undocumented, which has implications for the validity of the ground truth used. It is also important to recognize that a negative RT–PCR test does not necessarily mean that a patient does not have COVID-19. We encourage authors to evaluate their algorithms on datasets from the pre-COVID-19 era, such as performed by ref. [95], to validate any claims that the algorithm is isolating COVID-19-specific imaging features. It is common for non-COVID-19 diagnoses (for example, non-COVID-19 pneumonia) to be determined from imaging alone. However, in many cases, these images are the only predictors of the developed model, and using predictors to inform outcomes leads to optimistic performance.

Recommendations for evaluation. We emphasize the importance of using a well-curated external validation dataset of appropriate size to assess generalizability to other cohorts. Any useful model for diagnosis or prognostication must be robust enough to give reliable results for any sample from the target population rather than just on the sampled population. Calibration statistics should be calculated for the developed models to inform predictive error and decision curve analysis[96] performed for assessing clinical utility. It is important for authors to state how they ensured that images from the same patient were not included in the different dataset partitions, such as describing patient-level splits. This is an issue for approaches that consider 2D and 3D images as a single sample and also for those that process 3D volumes as independent 2D samples. It is also important when using datasets containing multiple images from each patient. When reporting results, it is important to include confidence intervals to reflect the uncertainty in the estimate, especially when training models on the small sample sizes commonly seen with COVID-19 data. Moreover, we stress the importance of not only reporting results but also demonstrating model interpretability with methods such as saliency maps, which is a necessary consideration for adoption into clinical practice. We remind authors that it is inappropriate to compare model performance to RT–PCR or any other ground truths. Instead, authors should aim for models to either improve the performance and efficiency of clinicians, or, even better, to aid clinicians by providing interpretable predictions. Examples of interpretability techniques include: (1) informing the clinician of which features in the data most influenced the prediction of the model, (2) linking the prognostic features to the underlying biology and (3) overlaying an activation/saliency map on the image to indicate the region of the image that influenced the model's prediction, and (4) identifying patients that had a similar clinical pathway.

Most papers derive their performance metrics from the test data alone with an unstated operating point to calculate sensitivity and specificity. Clinical judgement should be used to identify the desired sensitivity or specificity of the model and the operating point should be derived from the development data. The differences in the sensitivity and specificity of the model should be recorded separately for the validation and test data. Using an operating point of 0.5 and only reporting the test sensitivity and specificity fails to convey the reliability of the threshold. This is a key aspect of generalizability. Omitting it, in the process of device regulation, would see a US Food and Drug Administration 510K submission rejected.

Recommendations for replicability. A possible ambiguity arises due to updating of publicly available datasets or code. Therefore, we recommend that a cached version of the public dataset be saved, or the date/version quoted, and specific versions of data or code be appropriately referenced. (Git commit ids or tags can be helpful for this purpose to reference a specific version on GitHub, for example.) We acknowledge that although perfect replication is potentially not possible, details such as the seeds used for randomness and the actual partitions of the dataset for training, validation and testing would form very useful supplementary materials.

Recommendations for authors. For authors, we recommend assessing their paper against appropriate established frameworks, such as RQS, CLAIM, transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD), PROBAST and Quality Assessment of Diagnostic Accuracy Studies (QUADAS)[97–101]. By far the most common point leading to exclusion was failure to state the data pre-processing techniques in sufficient detail. As a minimum, we expected papers to state any image resizing, cropping and normalization used before model input, and with this small addition many more papers would have passed through the quality review stage. Other commonly missed points include details of the training (such as number of epochs and stopping criteria), robustness or sensitivity analysis, and the demographic or clinical characteristics of patients in each partition.

Recommendations for reviewers. For reviewers, we also recommend the use of the checklists[97–101] to better identify common weaknesses in reporting the methodology. The most common issues in the papers we reviewed was the use of biased datasets and/or methodologies. For non-public datasets, it may be difficult for reviewers to assess possible biases if an insufficiently detailed description is given by the authors. We strongly encourage reviewers to ask for clarification from the authors if there is any doubt about bias in the model being considered. Finally, we suggest using reviewers from a combination of both medical and machine learning backgrounds, as they can judge the clinical and technical aspects in different ways.

Challenges and opportunities. Models developed for diagnosis and prognostication from radiological imaging data are limited by the quality of their training data. While many public datasets exist for researchers to train deep learning models for these purposes, we have determined that these datasets are not large enough, or of suitable quality, to train reliable models, and all studies using publicly available datasets exhibit a high or unclear risk of bias. However, the size and quality of these datasets can be continuously improved if researchers worldwide submit their data for public review. Because of the uncertain quality of many COVID-19 datasets, it is likely more beneficial to the research community to establish a database that has a systematic review of submitted data than it is to immediately release data of questionable quality as a public database.

The intricate link of any AI algorithm for detection, diagnosis or prognosis of COVID-19 infections to a clear clinical need is essential for successful translation. As such, complementary computational and clinical expertise, in conjunction with high-quality healthcare data, are required for the development of AI algorithms. Meaningful evaluation of an algorithm's performance is most likely to occur in a prospective clinical setting. Like the need for collaborative development of AI algorithms, the complementary perspectives of experts in machine learning and academic medicine were critical in conducting this systematic review.

Limitations. Due to the fast development of diagnostic and prognostic AI algorithms for COVID-19, at the time of finalizing our analyses, several new preprints have been released; these are not included in this study.

Our study has limitations in terms of methodologic quality and exclusion. Several high-quality papers published in high-impact journals—including *Radiology*, *Cell* and *IEEE Transactions on Medical Imaging*—were excluded due to the lack of documentation on the proposed algorithmic approaches. As the AI algorithms are the core for the diagnosis and prognosis of COVID-19, we only included works that are reproducible. Furthermore, we acknowledge that the CLAIM requirements are harder to fulfil than the RQS ones, and the paper quality check is therefore not be fully comparable between the two. We underline that several excluded papers were preprint versions and may possibly pass the systematic evaluation in a future revision.

In our PROBAST assessment, for the 'Were there a reasonable number of participants?' question of the analysis domain, we required a model to be trained on at least 20 events per variable for the size of the dataset to score a low risk of bias[100]. However, events per variable may not be a useful metric to determine whether a deep learning model will overfit. Despite their gross over-parameterization, deep learning models generalize well in a variety of tasks, and it is difficult to determine a priori whether a model will overfit given the number of training examples[102]. A model that was trained using fewer than 500 COVID-19 positive images was deemed to have a high risk of bias in answer to this and more than 2,000 COVID-19 positive images qualified as low risk. However, in determining the overall risk of bias for the analysis domain, we factor in nine PROBAST questions, so it is possible for a paper using fewer than 500 images to achieve at best an unclear overall risk of bias for its analysis. Similarly, it is possible for papers that have over 2,000 images to have an overall high risk of bias for their analysis if it does not account for other sources of bias.

Conclusions

This systematic review specifically considers the current machine learning literature using CT and CXR imaging for COVID-19 diagnosis and prognosis, which emphasizes the quality of the methodologies applied and the reproducibility of the methods. We found that no papers in the literature currently have all of: (1) a sufficiently documented manuscript describing a reproducible method; (2) a method that follows best practice for developing a machine learning model; and (3) sufficient external validation to justify the wider applicability of the method. We give detailed specific recommendations for data curators, machine learning researchers, manuscript authors and reviewers to ensure the best-quality methods are developed that are reproducible and free from biases in either the underlying data or the model development.

Despite the huge efforts of researchers to develop machine learning models for COVID-19 diagnosis and prognosis, we found methodological flaws and many biases throughout the literature, leading to highly optimistic reported performance. In their current reported form, none of the machine learning models included in this review are likely candidates for clinical translation for the diagnosis/prognosis of COVID-19. Higher-quality datasets, manuscripts with sufficient documentation to be reproducible and external validation are required to increase the likelihood of models being taken forward and integrated into future clinical trials to establish independent technical and clinical validation as well as cost-effectiveness.

Methods

The methods for performing this systematic review are registered with PROSPERO (CRD42020188887) and were agreed by all authors before the start of the review process, to avoid bias.

Search strategy and selection criteria. We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist[103] and include this in Supplementary Table 1. We performed our search to identify published and unpublished works using the arXiv and the 'Living Evidence on COVID-19' database[104], a collation of all COVID-19 related papers from EMBASE via OVID, MEDLINE via PubMed, bioRxiv and medRxiv. The databases were searched from 1 January 2020 to 3 October 2020. The full search strategy is detailed in 'Search strategy'. The initial cut-off is chosen to specifically include all early COVID-19 research, given that the WHO was only informed of the 'pneumonia of unknown cause' on 31 December 2019[105]. An initial search was performed on 28 May 2020, with updated searches performed on 24 June 2020, 14 August 2020, 15 August 2020 and 3 October 2020 to identify any relevant new papers published in the intervening period. As many of the papers identified are preprints, some of them were updated or published between these dates; in such cases, we used the preprint as it was at the later search date or the published version. Some papers were identified as duplicates ourselves or by Covidence[106]; in these instances, we ensured that the latest version of the paper was reviewed. We used a three-stage process to determine which papers would be included in this review. During the course of the review, one author (A.I.A.-R.) submitted a paper[107] that was in scope for this review; however, we excluded it due to the potential for conflict of interest.

Title and abstract screening. In the first stage, a team of ten reviewers assessed papers for eligibility, screening the titles and abstracts to ensure relevance. Each paper was assessed by two reviewers independently and conflicts were resolved by consensus of the ten reviewers (Supplementary Data 1).

Full-text screening. In the second stage, the full text of each paper was screened by two reviewers independently to ensure that the paper was eligible for inclusion with conflicts resolved by consensus of the ten reviewers.

Quality review. In the third stage, we considered the quality of the documentation of methodologies in the papers. Note that exclusion at this stage is not a judgement on the quality or impact of a paper or algorithm, merely that the methodology is not documented with enough detail to allow the results to be reliably reproduced.

At this point, we separated machine learning methods into deep learning methods and non-deep learning methods (we refer to these as traditional machine learning methods). The traditional machine learning papers were scored using the RQS of Lambin et al.[97], while the deep learning papers were assessed against the CLAIM of Mongan et al.[98]. The ten reviewers were assigned to five teams of two: four of the ten reviewers have a clinical background and were paired with non-clinicians in four of the five teams to ensure a breadth of experience when reviewing these papers. Within each team, the two reviewers independently assessed each paper against the appropriate quality measure. Where papers contained both deep learning and traditional machine learning methodologies, these were assessed using both CLAIM and RQS. Conflicts were resolved by a third reviewer.

To restrict consideration to only those papers with the highest-quality documentation of methodology, we excluded papers that did not fulfil particular CLAIM or RQS requirements. For the deep learning papers evaluated using the CLAIM checklist, we selected eight checkpoint items deemed mandatory to allow reproduction of the paper's method and results. For the traditional machine learning papers, evaluated using the RQS, we used a threshold of 6 points out of 36 for inclusion in the review along with some basic restrictions, such as detail of the data source and how subsets were selected. The rationale for these CLAIM and RQS restrictions is given in Supplementary Discussion 1. If a paper was assessed using both CLAIM and RQS then it only needed to pass one of the quality checks to be included.

In a number of cases, various details of pre-processing, model configuration or training setup were not discussed in the paper, even though they could be inferred from a referenced online code repository (typically GitHub). In these cases, we have assessed the papers purely on the content in the paper, as it is important to be able to reproduce the method and results independently of the authors' code.

Risk of bias in individual studies. We use the PROBAST of Wolff et al.[100] to assess bias in the datasets, predictors and model analysis in each paper. The papers that passed the quality assessment stage were split among three teams of two reviewers to complete the PROBAST review. Within each team, the two reviewers independently scored the risk of bias for each paper and then resolved by conflicts any remaining conflicts were resolved by a third reviewer.

Data analysis. The papers were allocated among five teams of two reviewers. These reviewers independently extracted the following information: (1) whether the paper described a diagnosis or prognosis model; (2) the data used to construct the model; (3) whether there were predictive features used for the model construction; (4) the sample sizes used for the development and holdout cohorts (along with the number of COVID-19 positive cases); (5) the type of validation performed; (6) the best performance quoted in the paper for the validation cohort (whether internal, external or both); and (7) whether the code for training the model and the trained model were publicly available. Any conflicts were initially resolved by team discussions and remaining conflicts were resolved by a third reviewer.

Role of the funding source. The funders of the study had no role in the study design, data collection, data analysis, data interpretation or writing of the manuscript. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Search strategy. *Initial extraction.* For the arXiv papers, we initially extracted papers for the relevant date ranges that included 'ncov' (as a complete word), 'coronavirus', 'covid', 'sars-cov-2' or 'sars-cov2' in their title or abstract. For the 'Living Evidence on COVID-19' database[104], we downloaded all papers in the appropriate date range.

Refined search. We then filtered the identified papers using the following criteria: title or abstract contain one of: 'ai', 'deep', 'learning', 'machine', 'neural', 'intelligence', 'prognos', 'diagnos', 'classification', ' segmentation' and also contain one of 'ct', 'cxr', 'x-ray', 'xray', 'imaging', 'image', 'radiograph'. Only 'ai', 'ct' and 'cxr' are required to be complete words. This differs slightly from the original PROSPERO description; the search was widened to identify some additional papers. The full history of the searching and filtering source code can be explored at https://gitlab.developers.cam.ac.uk/maths/cia/covid-19-systematic-review

Data availability

All data generated or analysed during this study are included in this published article (and its supplementary information files).

Received: 20 August 2020; Accepted: 27 January 2021;
Published online: 15 March 2021

References

1. Wang, D. et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA* **323**, 1061–1069 (2020).
2. Zheng, Y. Y., Ma, Y. T., Zhang, J. Y. & Xie, X. COVID-19 and the cardiovascular system. *Nat. Rev. Cardiol.* **17**, 259–260 (2020).
3. *WHO Director-General's Remarks at the Media Briefing on 2019-nCoV on 11 February 2020* (World Health Organization, 2020); https://www.who.int/director-general/speeches/detail/who-director-general-s-remarks-at-the-media-briefing-on-2019-ncov-on-11-february-2020
4. *WHO Director-General's Opening Remarks at the Media Briefing on COVID-19—11 March 2020* (World Health Organization, 2020); https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020
5. Wong, H. Y. F. et al. Frequency and distribution of chest radiographic findings in COVID-19 positive patients. *Radiology* **296**, 201160 (2019).
6. Long, C. et al. Diagnosis of the coronavirus disease (COVID-19): rRT–PCR or CT? *Eur. J. Radiol.* **126**, 108961 (2020).
7. Fang, Y. et al. Sensitivity of chest CT for COVID-19: comparison to RT–PCR. *Radiology* **296**, 200432 (2020).
8. Ai, T. et al. Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. *Radiology* https://doi.org/10.1148/radiol.2020200642 (2020).
9. Sperrin, M., Grant, S. W. & Peek, N. Prediction models for diagnosis and prognosis in COVID-19. *BMJ* **369**, m1464 (2020).
10. Irvin, J. et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. *Proc. AAAI Conf. Artif. Intell.* **33**, 590–597 (2019).
11. Huang, P. et al. Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method. *Lancet Digit. Health* **1**, e353–e362 (2019).
12. Wynants, L. et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *BMJ* **369**, m1328 (2020).
13. Hamzeh, A. et al. Artificial intelligence techniques for containment COVID-19 pandemic: a systematic review. *Res. Sq.* https://doi.org/10.21203/rs.3.rs-30432/v1 (2020).
14. Albahri, O. S. et al. Systematic review of artificial intelligence techniques in the detection and classification of COVID-19 medical images in terms of evaluation and benchmarking: taxonomy analysis, challenges, future solutions and methodological aspects. *J. Infect. Public Health* https://doi.org/10.1016/j.jiph.2020.06.028 (2020).
15. Feng, S. et al. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19. *IEEE Rev. Biomed. Eng.* **14**, 4–15 (2021).
16. Tartaglione, E., Barbano, C. A., 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* **17**, 6933 (2020).
17. Ghoshal, B. & Tucker, A. Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. Preprint at http://arxiv.org/abs/2003.10769 (2020).
18. Malhotra, A. et al. Multi-task driven explainable diagnosis of COVID-19 using chest X-ray images. Preprint at https://arxiv.org/abs/2008.03205 (2020).
19. Rahaman, M. M. 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.* **28**, 821–839 (2020).
20. Amer, R., Frid-Adar, M., Gozes, O., Nassar, J. & Greenspan, H. COVID-19 in CXR: from detection and severity scoring to patient disease monitoring. Preprint at https://arxiv.org/abs/2008.02150 (2020).
21. Tsiknakis, N. et al. Interpretable artificial intelligence framework for COVID-19 screening on chest X-rays. *Exp. Ther. Med.* **20**, 727–735 (2020).
22. Elaziz, M. A. et al. New machine learning method for imagebased diagnosis of COVID-19. *PLoS ONE* **15**, e0235187 (2020).
23. Gil, D., Díaz-Chito, K., Sánchez, C. & Hernández-Sabaté, A. Early screening of SARS-CoV-2 by intelligent analysis of X-ray images. Preprint at https://arxiv.org/abs/2005.13928 (2020).
24. Tamal, M. et al. An integrated framework with machine learning and radiomics for accurate and rapid early diagnosis of COVID-19 from chest X-ray. Preprint at *medRxiv* https://doi.org/10.1101/2020.10.01.20205146 (2020).
25. Bararia, A., Ghosh, A., Bose, C. & Bhar, D. Network for subclinical prognostication of COVID 19 patients from data of thoracic roentgenogram: a feasible alternative screening technology. Preprint at *medRxiv* https://doi.org/10.1101/2020.09.07.20189852 (2020).
26. Wang, Z. et al. Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest X-rays. *Pattern Recognit.* **110**, 107613 (2021).
27. Zhang, R. et al. Diagnosis of COVID-19 pneumonia using chest radiography: value of artificial intelligence. *Radiology* https://doi.org/10.1148/radiol.2020202944 (2020).
28. Ezzat, D., Hassanien, A. E. & Ella, H. A. An optimized deep learning architecture for the diagnosis of COVID-19 disease based on gravitational search optimization. *Appl. Soft Comput. J.* https://doi.org/10.1016/j.asoc.2020.106742 (2020).
29. Farooq, M. & Hafeez, A. COVID-ResNet: a deep learning framework for screening of COVID19 from radiographs. Preprint at https://arxiv.org/abs/2003.14395 (2020).
30. Luz, E. et al. Towards an effective and efficient deep learning model for COVID-19 patterns detection in X-ray images. Preprint at http://arxiv.org/abs/2004.05717 (2020).
31. Bassi, P. R. A. S. & Attux, R. A deep convolutional neural network for COVID-19 detection using chest X-rays. Preprint at http://arxiv.org/abs/2005.01578 (2020).
32. Gueguim Kana, E. B., Zebaze Kana, M. G., Donfack Kana, A. F. & Azanfack Kenfack, R. H. A web-based diagnostic tool for COVID-19 using machine learning on chest radiographs (CXR). Preprint at *medRxiv* https://doi.org/10.1101/2020.04.21.20063263 (2020).
33. Heidari, M. et al. Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms. *Int. J. Med. Inform.* **144**, 104284 (2020).
34. Li, X., Li, C. & Zhu, D. COVID-MobileXpert: on-device COVID-19 screening using snapshots of chest X-ray. Preprint at http://arxiv.org/abs/2004.03042 (2020).
35. Zokaeinikoo, M., Mitra, P., Kumara, S. & Kazemian, P. AIDCOV: an interpretable artificial intelligence model for detection of COVID-19 from chest radiography images. Preprint at *medRxiv* https://doi.org/10.1101/2020.05.24.20111922 (2020).
36. Sayyed, A. Q. M. S., Saha, D. & Hossain, A. R. CovMUNET: a multiple loss approach towards detection of COVID-19 from chest X-ray. Preprint at https://arxiv.org/abs/2007.14318 (2020).
37. Zhang, R. 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.* **12**, 555–565 (2020).
38. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proce. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2818–2826 (IEEE Computer Society, 2016).
39. Tan, M. & Le, Q. EfficientNet: rethinking model scaling for convolutional neural networks. In *International Conference of Machine Learning ICML* 6105–6114 (PMLR, 2019).
40. Georgescu, B. et al. Machine learning automatically detects COVID-19 using chest CTs in a large multicenter cohort. Preprint at http://arxiv.org/abs/2006.04998 (2020).
41. Ko, H. 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.* **22**, e19569 (2020).
42. Wang, S. et al. A deep learning algorithm using CT images to screen for corona virus disease (COVID-19). Preprint at *medRxiv* https://doi.org/10.1101/2020.02.14.20023028 (2020).
43. Pu, J. et al. Any unique image biomarkers associated with COVID-19? *Eur. Radiol.* https://doi.org/10.1007/s00330-020-06956-w (2020).
44. Amyar, A., Modzelewski, R. & Ruan, S. Multi-task deep learning based CT imaging analysis for COVID-19: classification and segmentation. Preprint at *medRxiv* https://doi.org/10.1101/2020.04.16.2006470 (2020).
45. Ardakani, A. A., Kanafi, A. R., Acharya, U. R., 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.* **121**, 103795 (2020).
46. Mei, X. et al. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. *Nat. Med.* https://doi.org/10.1038/s41591-020-0931-3 (2020).
47. Han, Z. et al. Accurate screening of COVID-19 using attention-based deep 3D multiple instance learning. *IEEE Trans. Med. Imaging* **39**, 2584–2594 (2020).
48. Acar, E., Şahin, E. & Yilmaz, İ. Improving effectiveness of different deep learning-based models for detecting COVID-19 from computed tomography (CT) images. Preprint at *medRxiv* https://doi.org/10.1101/2020.06.12.20129643 (2020).
49. Chen, X., Yao, L., Zhou, T., Dong, J. & Zhang, Y. Momentum contrastive learning for few-shot COVID-19 diagnosis from chest CT images. Preprint at http://arxiv.org/abs/2006.13276 (2020).
50. Jin, S. et al. AI-assisted CT imaging analysis for COVID-19 screening: building and deploying a medical AI system in four weeks. Preprint at *medRxiv* https://doi.org/10.1101/2020.03.19.20039354 (2020).
51. Wang, S. et al. A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. *Eur. Respir. J.* https://doi.org/10.1183/13993003.00775-2020 (2020).

52. Shah, V. et al. Diagnosis of COVID-19 using CT scan images and deep learning techniques. Preprint at *medRxiv* https://doi.org/10.1101/2020.07.11.20151332 (2020).
53. Wang, J. et al. Prior-attention residual learning for more discriminative COVID-19 screening in CT images. *IEEE Trans. Med. Imaging* **39**, 2572–2583 (2020).
54. Wang, M. et al. Deep learning-based triage and analysis of lesion burden for COVID-19: a retrospective study with external validation. *Lancet Digit. Health* **2**, e506–e515 (2020).
55. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2010); https://doi.org/10.1109/cvpr.2009.5206848
56. Goodfellow, I. J. et al. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems* Vol. 2 2672–2680 (MIT Press, 2014).
57. Shi, F. et al. Large-scale screening of COVID-19 from community acquired pneumonia using infection size-aware classification. Preprint at http://arxiv.org/abs/2003.09860 (2020).
58. Guiot, J. et al. Development and validation of an automated radiomic CT signature for detecting COVID-19. Preprint at *medRxiv* https://doi.org/10.1101/2020.04.28.20082966 (2020).
59. Chen, X. X. et al. A diagnostic model for coronavirus disease 2019 (COVID-19) based on radiological semantic and clinical features: a multi-center study. *Eur. Radiol.* https://doi.org/10.1007/s00330-020-06829-2 (2020).
60. Qin, L. et al. A predictive model and scoring system combining clinical and CT characteristics for the diagnosis of COVID-19. *Eur. Radiol.* https://doi.org/10.1007/s00330-020-07022-1 (2020).
61. Xie, C. 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* **7**, 100271 (2020).
62. Xu, M. et al. Accurately differentiating COVID-19, other viral infection, and healthy individuals using multimodal features via late fusion learning. Preprint at *medRxiv* https://doi.org/10.1101/2020.08.18.20176776 (2020).
63. Chassagnon, G. et al. AI-driven CT-based quantification, staging and short-term outcome prediction of COVID-19 pneumonia. Preprint at *medRxiv* https://doi.org/10.1101/2020.04.17.20069187 (2020).
64. Ghosh, B. et al. A quantitative lung computed tomography image feature for multi-center severity assessment of COVID-19. Preprint at *medRxiv* https://doi.org/10.1101/2020.07.13.20152231 (2020).
65. Wei, W., Hu, X. W., Cheng, Q., Zhao, Y. M. & Ge, Y. Q. Identification of common and severe COVID-19: the value of CT texture analysis and correlation with clinical characteristics. *Eur. Radiol.* https://doi.org/10.1007/s00330-020-07012-3 (2020).
66. Li, M. D. et al. Improvement and multi-population generalizability of a deep learning-based chest radiograph severity score for COVID-19. Preprint at *medRxiv* https://doi.org/10.1101/2020.09.15.20195453 (2020).
67. Li, M. D. et al. Automated assessment and tracking of COVID-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks. *Radiol. Artif. Intell.* **2**, e200079 (2020).
68. Schalekamp, S. et al. Model-based prediction of critical illness in hospitalized patients with COVID-19. *Radiology* https://doi.org/10.1148/radiol.2020202723 (2020).
69. Wang, X. 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.* https://doi.org/10.2214/ajr.20.24044 (2020).
70. Yip, S. S. F. et al. Performance and robustness of machine learning-based radiomic COVID-19 severity prediction. Preprint at *medRxiv* https://doi.org/10.1101/2020.09.07.20189977 (2020).
71. Goncharov, M. et al. CT-based COVID-19 triage: deep multitask learning improves joint identification and severity quantification. Preprint at https://arxiv.org/abs/2006.01441 (2020).
72. Ramtohul, T. 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.* **10**, 1560 (2020).
73. Lassau, N. et al. AI-based multi-modal integration of clinical characteristics, lab tests and chest CTs improves COVID-19 outcome prediction of hospitalized patients. Preprint at *medRxiv* https://doi.org/10.1101/2020.05.14.20101972 (2020).
74. Yue, H. 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.* **8**, 859–859 (2020).
75. Zhu, X. et al. Joint prediction and time estimation of COVID-19 developing severe symptoms using chest CT scan. Preprint at http://arxiv.org/abs/2005.03405 (2020).
76. Cohen, J. P. et al. Predicting COVID-19 pneumonia severity on chest X-ray with deep learning. Preprint at https://arxiv.org/abs/2005.11856 (2020).
77. Chao, H. et al. Integrative analysis for COVID-19 patient outcome prediction. Preprint at https://arxiv.org/abs/2007.10416 (2020).
78. Wu, Q. et al. Radiomics analysis of computed tomography helps predict poor prognostic outcome in COVID-19. *Theranostics* **10**, 7231–7244 (2020).
79. Zheng, Y. 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.* **21**, 1007–1017 (2020).
80. Chen, Y. 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.* **17**, 1773–1782 (2020).
81. Bai, H. X. et al. Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other etiology on chest CT. *Radiology* **296**, E156–E165 (2020).
82. Yang, X. et al. COVID-CT-Dataset: a CT scan dataset about COVID-19. Preprint at http://arxiv.org/abs/2003.13865 (2020).
83. 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.* **126**, 104037 (2020).
84. *COVID-19 Radiography Database* (Kaggle, accessed 29 July 2020); https://www.kaggle.com/tawsifurrahman/covid19-radiography-database
85. Cohen, J. P., Morrison, P. & Dao, L. COVID-19 image data collection. Preprint at http://arxiv.org/abs/2003.11597 (2020).
86. Kermany, D. S. et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131.e9 (2018).
87. COVID-19: radiology reference article. *Radiopaedia* https://radiopaedia.org/articles/covid-19-4?lang=gb (accessed 29 July 2020).
88. *COVID-19 Database* (SIRM, accessed 29 July 2020); https://www.sirm.org/en/category/articles/covid-19-database/
89. *CORONACASES.org* (RAIOSS.com, accessed 30 July 2020); https://coronacases.org/
90. *Eurorad* (ESR, accessed 29 July 2020); https://www.eurorad.org/
91. *Chest X-Ray Images (Pneumonia)* (Kaggle, accessed 29 July 2020); https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia
92. Maguolo, G. & Nanni, L. A critic evaluation of methods for COVID-19 automatic detection from X-ray images. Preprint at http://arxiv.org/abs/2004.12823 (2020).
93. *RSNA Pneumonia Detection Challenge* (Kaggle, accessed 29 July 2020); https://www.kaggle.com/c/rsna-pneumonia-detection-challenge
94. Bachtiger, P., Peters, N. & Walsh, S. L. Machine learning for COVID-19—asking the right questions. *Lancet Digit. Health* **2**, e391–e392 (2020).
95. Banerjee, I. et al. Was there COVID-19 back in 2012? Challenge for AI in diagnosis with similar indications. Preprint at http://arxiv.org/abs/2006.13262 (2020).
96. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Making* **26**, 565–574 (2006).
97. Lambin, P. et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* **14**, 749–762 (2017).
98. Mongan, J., Moy, L. & Kahn, C. E. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol. Artif. Intell.* **2**, e200029 (2020).
99. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann. Intern. Med.* **162**, 55–63 (2015).
100. Wolff, R. F. et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann. Intern. Med.* **170**, 51 (2019).
101. Whiting, P. F. et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann. Intern. Med.* **155**, 529 (2011).
102. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations ICLR 2017* (ICLR, 2017).
103. Liberati, A. et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* **339**, b2700 (2009).
104. *COVID-19 Open Access Project: Living Evidence on COVID-19* (accessed 21 January 2021, COVID-19 Open Access Project); https://ispmbern.github.io/covid-19/living-review/
105. Pneumonia of unknown cause—China. *WHO* https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/ (2020).
106. Covidence Systematic Review Software v2486 6284093b (Veritas Health Innovation, 2020); https://www.covidence.org/
107. Aviles-Rivero, A. I., Sellars, P., Schönlieb, C.-B. & Papadakis, N. GraphXCOVID: explainable deep graph diffusion pseudo-labelling for identifying COVID-19 on chest X-rays. Preprint at https://arxiv.org/abs/2010.00378 (2020).

Acknowledgements

M.R. acknowledges the support of his supervisors M. Patel and A. Reynolds at AstraZeneca. The AIX-COVNET collaboration also acknowledges the support of Intel in supporting the collaboration. There is no direct funding source for this study;

however, the following authors do report indirect funding as follows: AstraZeneca (M.R.), European Research Council under the European Union's Horizon 2020 research and innovation programme grant agreement no. 777826 (M.T.), Wellcome Trust (C.E, C.M., J.H.F.R.), The Mark Foundation for Cancer Research (L.B., E.S.), Cancer Research UK Cambridge Centre (C9685/A25177) (L.B., E.S., C.-B.S.), British Heart Foundation (Z.T., J.H.F.R.), the NIHR Cambridge Biomedical Research Centre (Z.T., J.H.F.R.), HEFCE (J.H.F.R.), Lyzeum Ltd. (J.G.), The Gates Cambridge Trust (D.D.), the Cambridge International Trust (S.U.), the EPSRC Cambridge Mathematics of Information in Healthcare Hub EP/T017961/1 (A.I.A.-R., J.H.F.R., C.-B.S.) and the Cantab Capital Institute for the Mathematics of Information (A.I.A.-R, C.-B.S.). In addition, C.-B.S. acknowledges support from the Leverhulme Trust project on 'Breaking the non-convexity barrier', the Philip Leverhulme Prize, the EPSRC grants EP/S026045/1 and EP/T003553/1, the Wellcome Innovator Award RG98755, European Union Horizon 2020 research and innovation programmes under the Marie Skłodowska-Curie grant agreement no. 777826 NoMADS and no. 691070 CHiPS and the Alan Turing Institute. All authors had full access to all the data in the study and the corresponding author had final responsibility for the decision to submit for publication.

Author contributions

M.R., D.D., M.T., J.G., M.Y., A.I.A.-R., C.E., C.M., S.U. and L.B. contributed to the literature search, screening and quality assessment. M.R., D.D., M.T., J.G., M.Y., A.I.A.-R., C.E., C.M. and S.U. contributed to the data extraction, analysis and interpretation along with writing the original draft manuscript. M.R., D.D., M.T., J.G., M.Y., A.I.A.-R., C.E., C.M., S.U., L.B., J.R.W.-M., Z.T., E.G.-K., J.H.F.R., E.S. and C.-B.S. contributed critical revisions of the manuscript, and all authors approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-021-00307-0.

Correspondence and requests for materials should be addressed to M.R.

Peer review information *Nature Machine Intelligence* thanks Pengtao Xie and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

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

AIX-COVNET

Michael Roberts[1,2], Derek Driggs[1], Matthew Thorpe[3], Julian Gilbey[1], Michael Yeung[4], Stephan Ursprung[4,5], Angelica I. Aviles-Rivero[1], Christian Etmann[1], Cathal McCague[4,5], Lucian Beer[4], Jonathan R. Weir-McCall[4,6], Zhongzhao Teng[4], Effrossyni Gkrania-Klotsas[7], Alessandro Ruggiero[6,9], Anna Korhonen[10], Emily Jefferson[11], Emmanuel Ako[12], Georg Langs[13], Ghassem Gozaliasl[14], Guang Yang[15], Helmut Prosch[13,16], Jacobus Preller[17], Jan Stanczuk[1], Jing Tang[18], Johannes Hofmanninger[13], Judith Babar[17], Lorena Escudero Sánchez[4,5], Muhunthan Thillai[8,9,19], Paula Martin Gonzalez[5], Philip Teare[20], Xiaoxiang Zhu[21], Mishal Patel[20], Conor Cafolla[22], Hojjat Azadbakht[23], Joseph Jacob[24], Josh Lowe[25], Kang Zhang[26], Kyle Bradley[25], Marcel Wassin[27], Markus Holzer[27], Kangyu Ji[28], Maria Delgado Ortet[4], Tao Ai[29], Nicholas Walton[30], Pietro Lio[31], Samuel Stranks[32], Tolou Shadbahr[18], Weizhe Lin[33], Yunfei Zha[34], Zhangming Niu[35], James H. F. Rudd[8,36], Evis Sala[4,5,36] and Carola-Bibiane Schönlieb[1,36]

[9]Qureight Ltd, Cambridge, UK. [10]Language Technology Laboratory, University of Cambridge, Cambridge, UK. [11]Population Health and Genomics, School of Medicine, University of Dundee, Dundee, UK. [12]Chelsea and Westminster NHS Trust and Royal Brompton NHS Hospital, London, UK. [13]Department of Biomedical Imaging and Image-guided Therapy, Computational Imaging Research Lab Medical University of Vienna, Vienna, Austria. [14]Department of Physics, University of Helsinki, Helsinki, Finland. [15]National Heart and Lung Institute, Imperial College London, London, UK. [16]Boehringer, Ingelheim, Germany. [17]Addenbrooke's Hospital, Cambridge University Hospitals NHS Trust, Cambridge, UK. [18]Research Program in System Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland. [19]Interstitial Lung Disease Unit, Royal Papworth Hospital, Cambridge, Royal Papworth Hospital NHS Foundation Trust, Cambridge, UK. [20]Biopharmaceuticals R&D, AstraZeneca, Cambridge, UK. [21]Signal Processing in Earth Observation, Technical University of Munich, Munich, Germany. [22]Department of Chemistry, University of Cambridge, Cambridge, UK. [23]AINOSTICS Ltd, Manchester, UK. [24]Centre for Medical Image Computing, University College London, London, UK. [25]SparkBeyond UK Ltd, London, UK. [26]Center for Biomedicine and Innovations at Faculty of Medicine, Macau University of Science and Technology, Macau, China. [27]contextflow GmbH, Vienna, Austria. [28]Cavendish Laboratory, University of Cambridge, Cambridge, UK. [29]Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China. [30]Institute of Astronomy, University of Cambridge, Cambridge, UK. [31]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK. [32]Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, UK. [33]Department of Engineering, University of Cambridge, Cambridge, UK. [34]Department of Radiology, Renmin Hospital of Wuhan University, Wuhan, China. [35]Aladdin Healthcare Technologies Ltd, London, UK.