Title: Machine learning for COVID-19 detection and prognostication using chest radiographs and CT scans: a systematic methodological review

Authors: Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, James H.F. Rudd*, Evis Sala*, Carola-Bibiane Schönlieb* on behalf of the AIX-COVNET collaboration[†]

[†] Members are listed in the appendix.
* Joint senior authors.

Affiliations:
Department of Applied Mathematics and Theoretical Physics (M Roberts PhD, D Driggs MS, J Gilbey PhD, AI Aviles-Rivero PhD, C Etmann PhD, Prof C-B Schönlieb PhD) **and Department of Radiology** (M Yeung BA, S Ursprung MD, C McCague MBBS, L Beer PhD, JR Weir-McCall PhD, Z Teng PhD, Prof Evis Sala PhD) **and Department of Medicine** (JHF Rudd PhD) **and Cancer Research UK Cambridge Centre** (S Ursprung MD, C McCague MBBS, Prof Evis Sala PhD) **at the University of Cambridge, Cambridge, UK**.

Oncology R&D (M Roberts PhD) **at AstraZeneca, Cambridge, UK.**

Department of Mathematics (M Thorpe PhD) **at the University of Manchester.**

Royal Papworth Hospital (JR Weir-McCall PhD)**, Cambridge, UK.**

Corresponding author: Michael Roberts at the Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK (michael.roberts@maths.cam.ac.uk) and Oncology R&D at AstraZeneca, Cambridge, UK (michael.roberts2@astrazeneca.com).

Summary

Background: Machine learning methods offer great potential for fast and accurate detection and prognostication of COVID-19 from standard-of-care chest radiographs (CXR) and computed tomography (CT) images. In this systematic review we critically evaluate the machine learning methodologies employed in the rapidly growing literature.

Methods: In this systematic review we reviewed EMBASE via OVID, MEDLINE via PubMed, bioRxiv, medRxiv and arXiv for published papers and preprints uploaded from Jan 1, 2020 to June 24, 2020. Studies which consider machine learning models for the diagnosis or prognosis of COVID-19 from CXR or CT images were included. A methodology quality review of each paper was performed against established benchmarks to ensure the review focusses only on high-quality reproducible papers. This study is registered with PROSPERO [CRD42020188887].

Findings: Our search identified 952 studies, of which 168 were included after initial screening. After quality screening, 29 studies were included in this systematic review.

- Twenty of the studies are deep learning focused, seven are focussed on other machine learning approaches and two incorporate both deep and non-deep machine learning methods.
- There is more focus in the literature towards studies for COVID-19 detection, with 23 of the papers addressing this and six studies developing models for prognostication.
- CT images were used to develop the model in 19 studies, with nine using CXR and one using both CT and CXR.
- All studies reviewed have a high or unclear risk of bias, primarily due to: (i) studies' use of public datasets where the accurate diagnosis of COVID-19 cannot be confirmed; (ii) inappropriate control groups for non-COVID-19 patients (such as paediatric patients); (iii) unknown demographics in the datasets; (iv) use of small datasets, and (v) no validation on an external dataset.
- All papers report extremely optimistic results; some diagnostic models reporting area under the receiver operating characteristic curve (AUC) values of 1.00, for example.

Interpretation: Our review finds that none of the developed models discussed are of potential clinical use due to methodological flaws and underlying biases. This is a major weakness, given the urgency with which validated COVID-19 models are needed. Typically, we find that the documentation of a model's development is not sufficient to make the results reproducible and therefore of 168 candidate papers only 29 are deemed to be reproducible and subsequently considered in this review. We therefore encourage authors to use established machine learning checklists to ensure sufficient documentation is made available, and to follow the PROBAST (prediction model risk of bias assessment tool) framework to determine the underlying biases in their model development process and to mitigate these where possible. This is key to safe clinical implementation which is urgently needed.

Research in context:

Evidence before this study

Machine learning methods applied to medical imaging for COVID-19 offer fantastic promise for improving the accuracy of diagnosis, compared to the gold-standard RT-PCR, whilst also providing valuable insight for prognostication of patient outcomes. Since the pandemic began in early 2020, researchers have answered the 'call to arms' and numerous machine learning models for diagnosis and prognosis of COVID-19 using radiological imaging have been developed and hundreds of manuscripts have been written. In this systematic review, we searched the literature for both published papers and preprints from Jan 1, 2020 to June 24, 2020 using EMBASE via OVID, MEDLINE via PubMed, arXiv, bioRxiv and medRxiv. Many studies are hampered by issues with poor quality data, poor application of machine learning methodology, poor reproducibility, and biases in study design. No systematic review addresses in detail the methodological flaws in the current literature, considering issues in both the development of the machine learning method and the clinical applicability. An objective assessment of the quality of the documentation of the methodology and reproducibility for existing models is also missing from the literature.

Added value of this study

This is the first systematic review to consider the current machine learning literature for COVID-19 diagnosis and prognosis which emphasises the quality of the methodologies applied and the reproducibility of the methods. We found that no papers in the literature currently have all of: (i) a sufficiently documented manuscript describing a reproducible method, (ii) a method which follows best practice for developing a machine learning model, and (iii) sufficient external validation to justify the wider applicability of the method. We give detailed recommendations for data curators, machine learning researchers, manuscript authors and reviewers to ensure the best quality methods are developed which are reproducible and free from biases in either the underlying data or the model development.

Implications of all available evidence

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.

Introduction

In December 2019, a novel coronavirus was first recognised in Wuhan, China [1]. On January 30, 2020, as infection rates and deaths across China soared and the first death outside China was recorded, the 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 February 12, 2020 [3], and was declared a pandemic on March 11, 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 anti-viral therapies.

Although reverse transcription polymerase chain reaction (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, CXR abnormalities are visible in patients who initially had a negative RT-PCR test [5] and several studies have shown that chest 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. These models have the potential to exploit the large amount of multi-modal 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 which can not only distinguish COVID-19 from non-COVID-19 patients but also discern alternative types of pneumonia such as those of bacterial or other viral aetiologies.

With no standardisation, 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: (i) the bias in small data sets; (ii) the variability of large internationally-sourced data sets; (iii) the poor integration of multi-stream data, particularly imaging data; (iv) the difficulty of the task of prognostication, and (v) 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.

In this paper 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 preprints works to ensure maximal coverage of the literature. While earlier reviews provided a broad analysis of predictive models for COVID-19 diagnosis and prognosis [10–12], this review highlights the unique challenges researchers face when developing classical machine learning and deep learning models using imaging data. In this review, as in [10], we assessed the risk of bias in the papers considered; we then build upon [10] by incorporating a quality screening stage to ensure only those papers with sufficiently documented methodologies are reviewed. We also improve upon prior reviews by making detailed recommendations in five domains: (i) considerations when collating COVID-19 imaging datasets that are to be made public; (ii) methodological considerations for algorithm developers; (iii) specific issues about reproducibility of the results in the literature; (iv) considerations for authors to ensure sufficient documentation of methodologies in manuscripts, and (v) considerations for reviewers performing peer review of manuscripts. This review 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, whilst 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 952 papers that satisfied our search criteria; removing duplicates we retained 927 papers, and of these, 213 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 168 papers, of which, after quality review, 29 were included for discussion in this review (see Figure 1). Of these, 20 were deep learning papers, 7 were traditional machine learning papers and 2 were hybrid papers (using both approaches). The two hybrid papers both failed the CLAIM check but passed RQS.

Deep learning papers. There are seven CLAIM criteria not satisfied in at least half of the 20 papers:

(1) 16 do not perform any robustness or sensitivity analysis of their model.
(2) 16 do not complete any external validation.
(3) 15 do not report the demographics of their data partitions.
(4) 14 do not report the statistical tests used to assess significance of results or determine confidence intervals.

(5) 12 do not report confidence intervals for the performance.
(6) 11 do not present any interpretability techniques.
(7) 10 do not sufficiently report their limitations, biases or issues around generalizability.
The full CLAIM results are in Supp. Mat. A9.

Traditional machine learning papers. Of the nine papers, including the two hybrid papers, none use longitudinal imaging, perform a prospective study for validation or standardise image acquisition by using either a phantom study or a public protocol. Only two papers describe performing external validation and only one paper reports the calibration statistics and associated statistical significance for the model predictions. The full RQS scores are in Supp. Mat. A9.

Datasets considered:

Public datasets were used extensively in the literature appearing in 14/29 papers [13–15]. Private data is used in 17/29 papers with 11 using data from China, two use data from France and the remainder use data from Iran, the USA, Belgium or undisclosed locations.

Diagnostic models for COVID-19:

Diagnosis models using CXRs. Nine papers consider diagnosis of COVID-19 from CXR images [16–24]. All of these papers use off-the-shelf networks including ResNet-18 or ResNet-50 [16,17,21], DenseNet-121 [18,20,21,23], VGG-16 [22,24] and EfficientNet [19,25]. Most papers classify images into the three classes COVID-19, non-COVID-19 pneumonia and normal [16,18,19,21–24], while [17] considers an extra class by dividing non-COVID-19 pneumonia into viral and bacterial pneumonia. ResNet and DenseNet architectures reported better performance than the others, with accuracies ranging from 0·64 to 0·99. However, we caution against direct comparison since the papers use different training and testing settings (e.g. different datasets and data partition sizes) and consider a different number of classes.

Diagnostic models using CT scans and deep learning. Thirteen 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 [24,26–32] and/or a non-COVID-19 class [26,27,30,32–36]. The full 3D volumes were only considered in three papers [26,29,35] with the remainder considering isolated 2D slices or even 2D patches [31]. In most 2D models, authors employed transfer learning, with networks pre-trained on ImageNet [37]. Almost all models used lung segmentation as a pre-processing step. One paper [33] used a Generative Adversarial Network (GAN) [38] approach to address the paucity of COVID-19 CT imaging. AUCs are reported ranging from 0·70 to 1·00.

Diagnostic models using CT scans and traditional machine learning methods. Five papers employed traditional machine learning methods for COVID-19 diagnosis, using hand-engineered features [26,39–41] or CNN-extracted features [32]. Two papers [32,41] incorporate clinical features with those obtained from the CT images. All papers using hand-engineered features employed feature reduction, using between 10 and 39 features in their final models. For final classification, three papers used logistic regression [26,40,41], one used a random forest [39] and one a multilayer perceptron [32]. Accuracies ranged from 0·76 to 0·90 [26,32,39–41]. As before, we caution against direct comparison. The traditional machine-learning model in the hybrid paper [26] had a 0·05 lower accuracy than their deep learning model.

Prognostic models for COVID-19 using CT and CXR images: Six papers developed models for the prognosis of patients with COVID-19 [36,42–46], five using CT and one using CXR. Two developed models estimating mortality risk or need for intubation (AUC=0·70, Accuracy=0·81 resp.) [42,43], two predicted the length of hospital stay (AUC=0·97, 0·88 resp.) [36,44], one estimated likelihood of conversion to severe disease (AUC=0·86) [45], and one predicted the extent of lung infection using CXR (Correlation=0·78) [46]. Predictors from radiological data were extracted using either handcrafted radiomic features [42,44,45] or deep learning [36,43,46]. Clinical data included basic observations, serology and comorbidities. Only one model integrated both radiological and clinical data [43].

Risks of bias: Following the PROBAST guidance, the risk of bias was assessed for all 29 papers in four domains: participants, predictors, outcomes and analysis; the results are shown in Table 1. We find that 26/29 papers had a high risk of bias in at least one domain with the others unclear in at least one domain.

Participants. All papers had a high (24/29) or unclear (5/29) risk of bias for their participants. This is primarily due to the following issues: (i) for public datasets it is not possible to know whether patients are truly COVID-19 positive as anybody can contribute images [16,18–21,23,24,27,30,33,34,46]; (ii) the paper uses only a subset of original datasets, applying

some exclusion criteria, without enough details to be reproducible [16,29,30,33,34,36,45,46], and/or (iii) there are significant differences in demographics between the COVID-19 cohort and the control groups [17,18,20,21,24,31,32,41,47].

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 (23/29), we cannot judge biases in the predictors. For the remaining six papers, the risk of bias is recorded as low due to the use of pre-defined hand-engineered features.

***Outcomes*.** The risk of bias in the outcome variable was found to be low for the majority (22/29) of the papers, unclear for 3/29 and high for 4/29. To evaluate the bias in the outcome, we took different approaches for papers using private datasets and public datasets (two papers use a mixture).

For the 17 papers that use private datasets, the COVID-19 diagnosis is due to either positive RT-PCR or antibody tests for 13/17 and have a low risk of bias. The other papers have a high risk of bias due to inconsistent diagnosis of COVID-19 [15,26], unclear definition of a control group [42] or using an unestablished reference to define outcome [44].

For the 14 papers that use public datasets, the outcome was assigned by the originators of the dataset and not by the papers' authors. Papers using a public dataset generally have a low risk of bias (12/14) as they have used the outcome directly obtained from the dataset and the bias due to public datasets has been considered in the participants' domain.

***Analysis*.** Only four papers have a low risk of bias for their analysis. The high risk of bias in most papers is principally due to a small sample size of COVID-19 patients leading to highly imbalanced datasets and/or using only a single internal holdout set for validating their algorithm [16,18–20,23,30,31,33,34]. One paper with a high risk of bias [21] claims external validation on dataset [13], not realising that this already includes both datasets [14] and [48] 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 is 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 generalisability of the algorithm. We find 23/29 papers consider internal validation only with 6/29 using external validation [21,27,28,36,42,43]. Five used truly external test datasets and one tested on the same data the algorithm was trained on [21].

Model evaluation. In Table 2 we give the performance metrics quoted in each paper. Five papers use cross-validation to evaluate model performance [24,34,39,44,45], one uses both cross-validation and an external test set [27], one quotes correlation metrics [46] and one has an unclear validation method [17]. The other papers all have an internal holdout or external test set with sensitivity and specificity derived from the test data using an unquoted operating point (with exception of [16] that quotes operating point 0·5). The ROC curves and AUC values are also quoted just for the test data, independent of the validation data.

Partition analysis. In Figure 2, we show the quantity of data (split by class) used in the training cohort of 19 diagnosis models. We exclude [21,24,29,31,40] as 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 13/19 papers have a reasonable balance between classes (with exceptions being [17,19,20,22,26,36]). However, the majority of datasets are quite small, with 11/19 papers using less than 2,000 datapoints for development (with exceptions [17,19,20,22,27,33,39,47]). Only four papers used both a dataset with more than 2000 datapoints that was balanced for COVID-19 positive and the other classes [27,33,39,47].

Figure 3 shows the number of images of each class used in the holdout/test cohorts. We find 2/19 papers had an imbalanced testing dataset [17,22]. Only 3/19 papers tested on more than 1,000 images [17,27,47]. Only 2/19, [27,47], had both a large and balanced testing dataset.

Public availability of the algorithms and models: Only 6/29 papers [19,23,32,44,46,47] publish the code for reproducing their results (even including their pre-trained parameters), with one to be shared in the future [42].

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 utilisation in clinical practice.

The current paper complements the work of Wynants et al. who have published a living systematic review [10] 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 symptomatic patients, and prognostication in patients with confirmed COVID-19 infection. While Wynants et al. reviewed multivariable models with any type of clinical input data, the present review focuses specifically on machine learning based diagnostic and prognostic models using medical imaging. Furthermore, this systematic review employed specialised 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 [11,12]. 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 likely lead to overly optimistic performance evaluations.

Datasets: Many papers gave little attention to establishing the original source of the images (Supp. Mat. B2). 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 such as [14,49–52]. 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). 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 (8/29) use the pneumonia dataset of Kermany et al. [48] as a control group. They commonly fail to mention that this consists of paediatric patients aged between one and five. Developing a model using adult COVID-19 patients and very young pneumonia patients is likely to overperform as it is merely detecting children vs. adults. This dataset is also erroneously referred to as the Mooney dataset in many papers (being the Kermany dataset deployed on Kaggle [53]). 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 in [54] that by excluding the lung region the authors could identify the source with and AUC between 0.9210 to 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, dataset [55] combines datasets [48–50] and dataset [53] combines [14,50,55], ignoring that dataset [50] is already included in dataset [55]. This repackaging of datasets, although pragmatic, inevitably leads to problems with algorithms being trained and tested on identical or overlapping datasets whilst believing them to be from distinct sources.

- ***Implicit biases in the source data.*** Images uploaded to a public repository and those extracted from publications [55] 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, e.g. 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 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 utilise transfer learning in developing their model, which assumes an inherent benefit to performance. However, it is unclear whether transfer learning offers significant performance benefit due to the over-parametrisation of the models [27,40]. 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: (i) the data used for model development and common pitfalls; (ii) improving the evaluation of trained models to ensure robust testing; (iii) improving reproducibility; (iv) improving the quality of documentation in manuscripts, and (v) improving the quality of peer review.

Recommendations for data. Firstly, we advise caution over the use of public repositories, which can lead to significant risks of bias: there are problems of duplication, resulting in identical copies of images appearing in different dataset partitions. Furthermore, authors should aim to match demographics across cohorts, an often neglected but significant potential source of bias; this can be impossible with public datasets that do not include demographic information, and including paediatric images [48] 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 (e.g. 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 (e.g. scanner manufacturer, slice thickness, etc.).

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, in sick, immobile patients, an anteroposterior CXR view is used for practicality rather than the standard posteroanterior CXR projection. Overrepresentation of severe disease is not only bad from the machine learning perspective, but also in terms of clinical utility, since the most useful algorithms are those that can diagnose disease at an early stage [56]. The timing between imaging and RT-PCR tests were also largely not documented, which has implications for the validity of the ground truth used. It is also important to recognise 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 [57], 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 emphasise the importance of using external validation in order 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 the sampled population. For prognostic models, calibration statistics should be calculated to inform predictive error, and decision curve analysis performed for assessing clinical utility. For models that process 3D data as 2D images, it should be stated how images from the same patient were not included in the different dataset partitions, such as describing patient-level splits. 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, and instead should aim for models to either outperform clinicians, or even better to aid clinicians by providing interpretable outputs.

Most papers derive their performance metrics from the test data alone with an unstated operating point to calculate sensitivity and specificity. However, the operating point to determine the performance of a model should be derived from the development data with the difference in sensitivity and specificity of the model 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 generalisability. Omitting it would see an FDA 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 should 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 Git6Hub, for example. The commands used to run the code may also be significant.) 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, TRIPOD, PROBAST and QUADAS [58–62]. 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 and normalisation used prior to 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 [58–62] in order 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 world-wide 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 which 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 was 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 finalising our analyses, several new preprints have been released; these are not included in this study. However, we will continue to monitor the growing body of literature, updating our review when relevant evidence comes to light at https://covid19ai.maths.cam.ac.uk/.

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 might 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, we require 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 [61]. However, events-per-variable may not be a useful metric to determine if a deep learning model will overfit. Despite their gross over-parameterisation, deep learning models generalise 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 [63]. Therefore, for deep learning models we have assessed this as low or high risk on the basis of the number of samples used to develop the model, along with how many samples were COVID-19 positive.

Conclusions

Despite the huge efforts of researchers to develop machine learning models for COVID-19 diagnosis and prognosis, we find methodological flaws and significant biases throughout the literature, leading to highly optimistic reported performance. Therefore, we conclude that none of these diagnosis and prognosis machine learning models has clinical utility. We make significant recommendations in many areas for researchers, data curators, authors and manuscript reviewers based on our observations of systematic issues in the literature.

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 have followed the PRISMA checklist [64] and include this in Supp. Mat. C. We performed our search to identify published and unpublished works using the arXiv and the "Living Evidence on COVID-19" database [65], a collation of all COVID-19 related papers from EMBASE via OVID, MEDLINE via PubMed, bioRxiv and medRxiv. The databases were searched from Jan 1, 2020 through to June 24, 2020. The full search strategy is detailed in the appendix. The initial cut-off is chosen to specifically include all early COVID-19 research, given that the World Health Organisation was only informed of the "pneumonia of unknown cause" on Dec 31, 2019 [66]. An initial search was performed on May 28, 2020 and an updated search performed on June 24, 2020 to identify any relevant new papers published in the intervening period. Since many of the papers identified are preprints, some of them were updated between these dates; in such cases, we used the preprint as it was at the later search date. Some papers were identified as duplicates by Covidence [67]; we accepted the version that Covidence determined. We used a three-stage process to determine which papers would be included in this review.

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 (see Supp. Mat. A4).

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. Explicitly, 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 Radiomic Quality Score (RQS) of Lambin et al. [58], while the deep learning papers were assessed against the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) of Mongan et al. [59]. 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 Supp. Mat. A7. 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 many 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 Prediction model Risk Of Bias Assessment Tool (PROBAST) of Wolff et al. [61] to assess the bias in the datasets, predictors and model analysis in each paper. The papers that passed the quality assessment stage were split amongst 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, with conflicts resolved by a third reviewer.

Data analysis: In the fifth and final stage, the papers were allocated amongst five teams of two reviewers. These reviewers independently extracted the following information: (i) whether the paper described a diagnosis or prognosis model; (ii) the data used to construct the model; (iii) whether there were predictive features used for the model construction; (iv) the sample sizes used for the development and holdout cohorts (along with the number of COVID-19 positive cases); (v) the type of validation performed; (vi) the best performance quoted in the paper, and (vii) whether the code for training the model and the trained model were publicly available. Any 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.

Funding

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.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.R.), the NIHR Cambridge Biomedical Research Centre (Z.T., J.R.), HEFCE (J.R.), The Gates Cambridge Trust (D.D.), the Cambridge International Trust (S.U.), the EPSRC Centre Nr. EP/N014588/1 (A.A-R., J.R., C-B.S.) and the Cantab Capital Institute for the Mathematics of Information (A.A-R, C-B.S.)

Additionally, 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.

Acknowledgments

M.R. would like to acknowledge the support of his supervisors Mishal Patel and Andrew Reynolds at AstraZeneca.

Contributors

M.R., D.D., M.T., J.G., M.Y., A.A-R., C.E., C.M., S.U., L.B. contributed to the literature search, screening, quality assessment. M.R., D.D., M.T., J.G., M.Y., A.A-R., C.E., C.M., 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.A-R., C.E., C.M., S.U., L.B., J. W-M., Z.T., J.R., E.S. and C-B.S. contributed critical revisions of the manuscript, and all authors approved the final manuscript.

Declaration of interests

M.R., D.D., M.T., J.G., M.Y., A.A-R., C.E., C.M., S.U., L.B., J.W-M., Z.T., J.R., E.S. and C-B.S. have no conflicts of interest to declare.

References

1. Wang, D. *et al.* Clinical Characteristics of 138 Hospitalized Patients with 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA - J. Am. Med. Assoc.* **323**, 1061–1069 (2020).

2. Zheng, Y. Y., Ma, Y. T., Zhang, J. Y. & Xie, X. COVID-19 and the cardiovascular system. *Nature Reviews Cardiology* vol. 17 259–260 (2020).

3. WHO. WHO Director-General's remarks at the media briefing on 2019-nCoV on 11 February 2020. *World Health Organization, Geneva* (2020).

4. WHO. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. *World Health Organization, Geneva* (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**, (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* 200642 (2020) doi:10.1148/radiol.2020200642.

9. Sperrin, M., Grant, S. W. & Peek, N. Prediction models for diagnosis and prognosis in Covid-19. *The BMJ* vol. 369 (2020).

10. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ* **369**, (2020).

11. Hamzeh, A. *et al.* Artificial intelligence techniques for Containment COVID-19 Pandemic: A Systematic Review. *Res. Sq.* (2020) doi:https://doi.org/10.21203/rs.3.rs-30432/v1.

12. 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. *Journal of Infection and Public Health* (2020) doi:10.1016/j.jiph.2020.06.028.

13. COVID-19 Radiography Database | Kaggle. https://www.kaggle.com/tawsifurrahman/covid19-radiography-database.

14. Cohen, J. P., Morrison, P. & Dao, L. COVID-19 Image Data Collection. *arXiv* http://arxiv.org/abs/2003.11597 (2020).

15. Yang, X. *et al.* COVID-CT-Dataset: A CT Scan Dataset about COVID-19. *arXiv* http://arxiv.org/abs/2003.13865 (2020).

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. *arXiv* http://arxiv.org/abs/2004.05405 (2020).

17. Ghoshal, B. & Tucker, A. Estimating Uncertainty and Interpretability in Deep Learning for Coronavirus (COVID-19) Detection. *arXiv* http://arxiv.org/abs/2003.10769 (2020).

18. Ezzat, D., Hassanien, A. ell & Ella, H. A. GSA-DenseNet121-COVID-19: a Hybrid Deep Learning Architecture for the Diagnosis of COVID-19 Disease based on Gravitational Search Optimization Algorithm. *arXiv* http://arxiv.org/abs/2004.05084 (2020).

19. Luz, E. *et al.* Towards an Effective and Efficient Deep Learning Model for COVID-19 Patterns Detection in X-ray Images. *arXiv* http://arxiv.org/abs/2004.05717 (2020).

20. Bassi, P. R. A. S. & Attux, R. A Deep Convolutional Neural Network for COVID-19 Detection Using Chest

X-Rays. *arXiv* http://arxiv.org/abs/2005.01578 (2020).

21. 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). *medRxiv* 2020.04.21.20063263 (2020) doi:10.1101/2020.04.21.20063263.

22. Heidari, M. *et al.* Improving performance of CNN to predict likelihood of COVID-19 using chest X-ray images with preprocessing algorithms. *arXiv* http://arxiv.org/abs/2006.12229 (2020).

23. Li, X., Li, C. & Zhu, D. COVID-MobileXpert: On-Device COVID-19 Screening using Snapshots of Chest X-Ray. *arXiv* http://arxiv.org/abs/2004.03042 (2020).

24. Zokaeinikoo, M., Mitra, P., Kumara, S. & Kazemian, P. AIDCOV: An Interpretable Artificial Intelligence Model for Detection of COVID-19 from Chest Radiography Images. *medRxiv* 2020.05.24.20111922 (2020) doi:10.1101/2020.05.24.20111922.

25. Tan, M. & Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. in *International Conference of Machine Learning ICML* (2019).

26. Georgescu, B. *et al.* Machine Learning Automatically Detects COVID-19 using Chest CTs in a Large Multicenter Cohort. *arXiv* http://arxiv.org/abs/2006.04998 (2020).

27. 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).

28. Wang, S. *et al.* A deep learning algorithm using CT images to screen for corona virus disease (COVID-19). *medRxiv* 1–19 (2020) doi:10.1101/2020.02.14.20023028.

29. Pu, J. *et al.* Any unique image biomarkers associated with COVID-19? *Eur. Radiol.* 1–7 (2020) doi:10.1007/s00330-020-06956-w.

30. Amyar, A., Modzelewski, R. & Ruan, S. Multi-task Deep Learning Based CT Imaging Analysis For COVID-19: Classification and Segmentation. *medRxiv* 2020.04.16.20064709 (2020) doi:10.1101/2020.04.16.20064709.

31. 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).

32. Mei, X. *et al.* Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. *Nat. Med.* 1–5 (2020) doi:10.1038/s41591-020-0931-3.

33. Acar, E., ŞAhİN, E. & Yilmaz, İ. Improving effectiveness of different deep learning-based models for detecting COVID-19 from computed tomography (CT) images. *medRxiv* 2020.06.12.20129643 (2020) doi:10.1101/2020.06.12.20129643.

34. Chen, X., Yao, L., Zhou, T., Dong, J. & Zhang, Y. Momentum Contrastive Learning for Few-Shot COVID-19 Diagnosis from Chest CT Images. *arXiv* http://arxiv.org/abs/2006.13276 (2020).

35. Jin, S. *et al.* AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system in four weeks. *medRxiv* 1–22 (2020) doi:10.1101/2020.03.19.20039354.

36. Wang, S. *et al.* A Fully Automatic Deep Learning System for COVID-19 Diagnostic and Prognostic Analysis. *medRxiv* 2020.03.24.20042317 (2020) doi:10.1101/2020.03.24.20042317.

37. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. in 248–255 (Institute of Electrical and Electronics Engineers (IEEE), 2010). doi:10.1109/cvpr.2009.5206848.

38. Goodfellow, I. J. *et al. Generative Adversarial Nets*. (2014).

39. Shi, F. *et al.* Large-Scale Screening of COVID-19 from Community Acquired Pneumonia using Infection Size-Aware Classification. *arXiv* http://arxiv.org/abs/2003.09860 (2020).

40. Guiot, J. *et al.* Development and validation of an automated radiomic CT signature for detecting

COVID-19. *medRxiv* 2020.04.28.20082966 (2020) doi:10.1101/2020.04.28.20082966.

41. 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.* 1–10 (2020) doi:10.1007/s00330-020-06829-2.

42. Chassagnon, G. *et al.* AI-Driven CT-based quantification, staging and short-term outcome prediction of COVID-19 pneumonia. *medRxiv* (2020) doi:10.1101/2020.04.17.20069187.

43. 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. *medRxiv* 2020.05.14.20101972 (2020) doi:10.1101/2020.05.14.20101972.

44. Qi, X. *et al. Machine learning-based CT radiomics model for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: A multicenter study*. *medRxiv* (2020) doi:10.1101/2020.02.29.20029603.

45. Zhu, X. *et al.* Joint Prediction and Time Estimation of COVID-19 Developing Severe Symptoms using Chest CT Scan. *arXiv* http://arxiv.org/abs/2005.03405 (2020).

46. Cohen, J. P. *et al.* Predicting COVID-19 Pneumonia Severity on Chest X-ray with Deep Learning. *arXiv* http://arxiv.org/abs/2005.11856 (2020).

47. Bai, H. X. *et al.* AI Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Etiology on Chest CT. *Radiology* 201491 (2020).

48. Kermany, D. S. *et al.* Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* **172**, 1122-1131.e9 (2018).

49. COVID-19 | Radiology Reference Article | Radiopaedia.org. https://radiopaedia.org/articles/covid-19-4?lang=gb.

50. COVID-19 DATABASE | SIRM. https://www.sirm.org/en/category/articles/covid-19-database/.

51. CORONACASES.org by RAIOSS.COM. https://coronacases.org/.

52. Eurorad. https://www.eurorad.org/.

53. Chest X-Ray Images (Pneumonia) | Kaggle. https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia.

54. Maguolo, G. & Nanni, L. A Critic Evaluation of Methods for COVID-19 Automatic Detection from X-Ray Images. *arXiv* http://arxiv.org/abs/2004.12823 (2020).

55. RSNA Pneumonia Detection Challenge | Kaggle. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge.

56. Bachtiger, P., Peters, N. & Walsh, S. L. Machine learning for COVID-19—asking the right questions. *Lancet Digit. Heal.* **2**, e391–e392 (2020).

57. Banerjee, I. *et al.* Was there COVID-19 back in 2012? Challenge for AI in Diagnosis with Similar Indications. *arXiv* http://arxiv.org/abs/2006.13262 (2020).

58. Lambin, P. *et al.* Radiomics: The bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* **14**, 749–762 (2017).

59. 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).

60. 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).

61. 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).

62. Whiting, P. F. *et al.* QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann. Intern. Med.* **155**, 529 (2011).

63. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning requires rethinking generalization. *5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc.* (2016).

64. 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**, (2009).

65. Covid-19 living Data. https://covid-nma.com/.

66. WHO | Pneumonia of unknown cause – China. https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/.

67. Veritas Health Innovation, M. A. Covidence systematic review software. https://www.covidence.org/.

FIGURE 1: FLOWCHART FOR PAPER INCLUSION

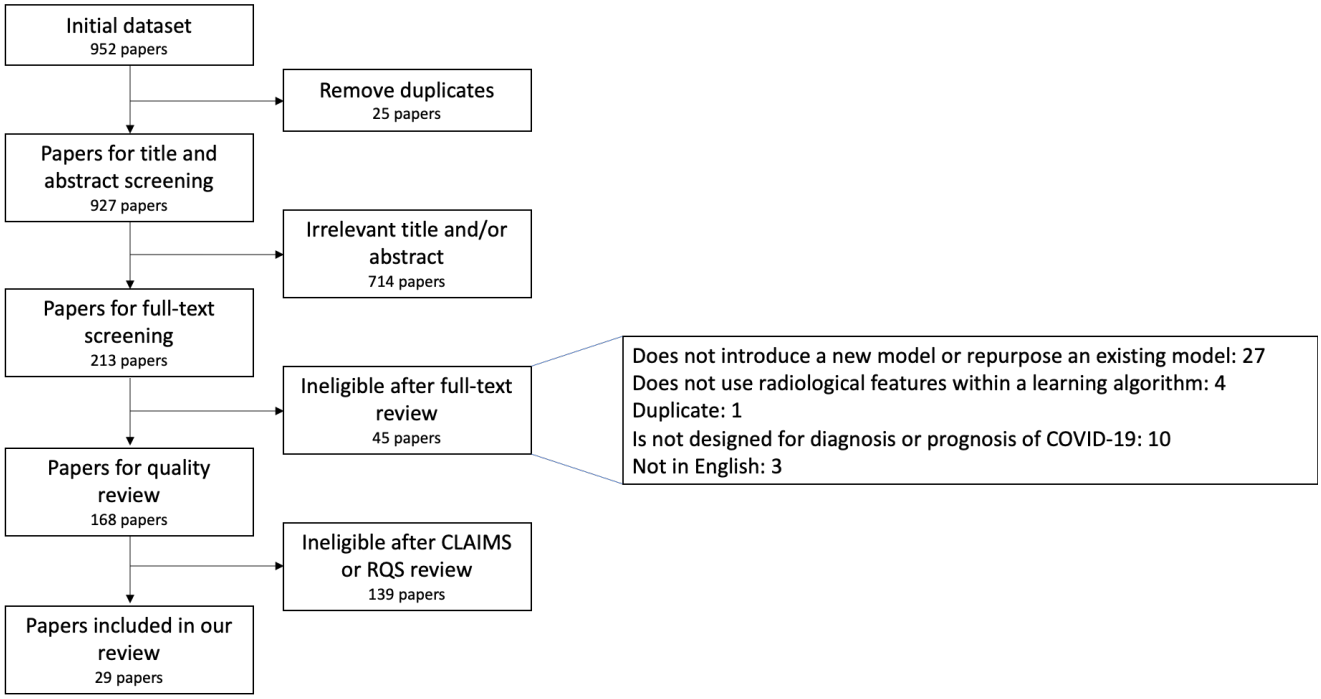

TABLE 1: PROBAST Results

REFERENCE	PARTICIPANTS	PREDICTORS	OUTCOMES	ANALYSIS
Ghoshal et al. [17]	HIGH	UNCLEAR (DL)	LOW	HIGH
Li et al. [23]	UNCLEAR	UNCLEAR (DL)	LOW	HIGH
Ezzat et al. [18]	HIGH	UNCLEAR (DL)	LOW	HIGH
Tartaglione et al. [16]	HIGH	UNCLEAR (DL)	LOW	HIGH
Luz et al. [19]	UNCLEAR	UNCLEAR (DL)	LOW	HIGH
Bassi et al. [20]	HIGH	UNCLEAR (DL)	LOW	HIGH
Kana et al. [21]	HIGH	UNCLEAR (DL)	LOW	HIGH
Heidari et al. [22]	HIGH	UNCLEAR (DL)	LOW	UNCLEAR

Zokaeinikoo et al. [24]	HIGH	UNCLEAR (DL)	LOW	UNCLEAR
Amyar et al. [30]	HIGH	UNCLEAR (DL)	LOW	HIGH
Ardakani et al. [31]	HIGH	UNCLEAR (DL)	LOW	HIGH
Bai et al. [47]	HIGH	UNCLEAR (DL)	LOW	LOW
Jin et al. [35]	HIGH	UNCLEAR (DL)	LOW	UNCLEAR
Wang et al. [28]	HIGH	UNCLEAR (DL)	LOW	UNCLEAR
Ko et al. [27]	HIGH	UNCLEAR (DL)	HIGH	UNCLEAR
Acar et al. [33]	HIGH	UNCLEAR (DL)	UNCLEAR	HIGH
Pu et al. [29]	UNCLEAR	UNCLEAR (DL)	LOW	UNCLEAR
Georgescu et al. [26]	HIGH	UNCLEAR (DL)	HIGH	UNCLEAR
Chen et al. [34]	HIGH	UNCLEAR (DL)	LOW	UNCLEAR
Guiot et al. [40]	HIGH	LOW	LOW	UNCLEAR
Shi et al. [39]	HIGH	LOW	LOW	LOW
Mei et al. [32]	HIGH	UNCLEAR (DL)	LOW	UNCLEAR
Chen et al. [41]	HIGH	LOW	LOW	HIGH
Wang et al. [36]	UNCLEAR	UNCLEAR (DL)	LOW	LOW
Cohen et al. [46]	HIGH	UNCLEAR (DL)	LOW	UNCLEAR
Qi et al. [44]	HIGH	LOW	HIGH	HIGH
Zhu et al. [45]	HIGH	LOW	LOW	LOW
Lassau et al. [43]	HIGH	UNCLEAR (DL)	HIGH	UNCLEAR
Chassagnon et al. [42]	UNCLEAR	LOW	LOW	UNCLEAR

FIGURE 2: DEVELOPMENT DATA SAMPLES AND CLASSES

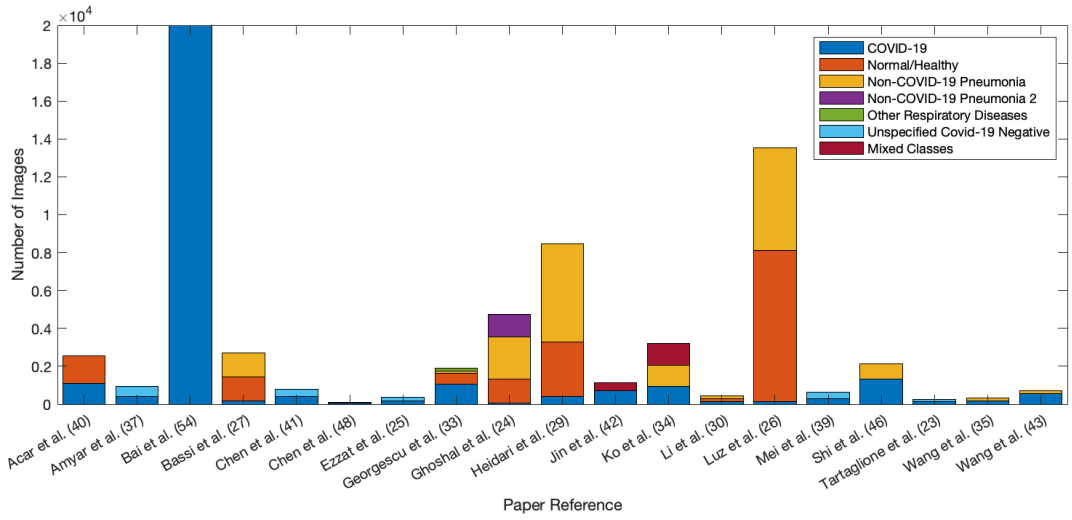

Figure 2: The number of images used in each paper for model training split by image class. Note: the figure is truncated at 20,000 images and Bai et al. [47] use significantly more (118,401 images in total).

FIGURE 3: TESTING DATA SAMPLES AND CLASSES

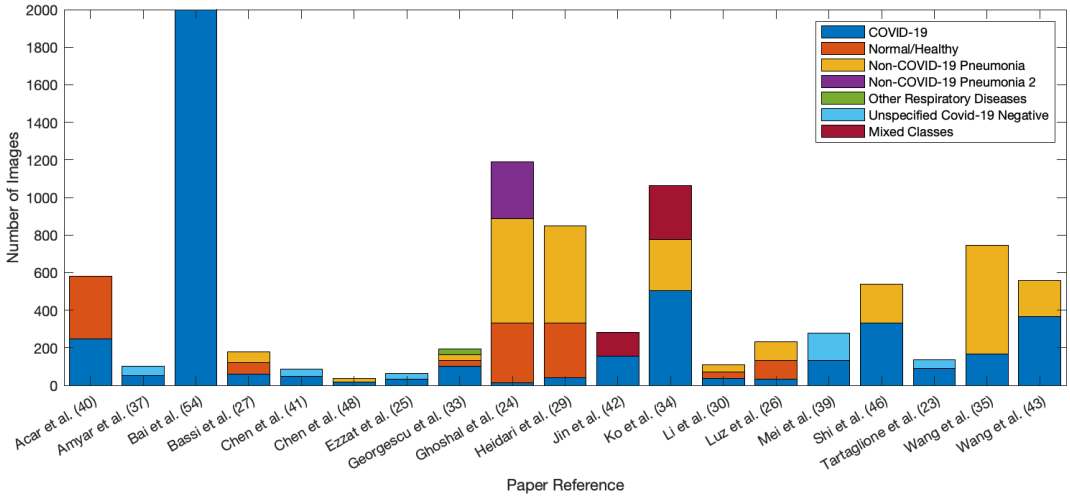

Figure 3: The number of images used for model testing split by image class. Note: the figure is truncated at 2,000 images and Bai et al. [47] use significantly more (14,182 images in total).

TABLE 2: DATA EXTRACTION

OUR 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?) does it include clinical data?	What are the predictors? In purely deep-learning models, this is DL.	Total sample size used for development (i.e. 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, etc	Performance of diagnosis model, AUC, confidence interval, sensitivity, specificity, etc. 95% CI if available.	Is there code available ? (Is the trained model available ?)
Ghoshal et al. [17]	Diagnosis	CXR	DL	4,752 images, 54 COVID-19.	1,189 images, 14 COVID-19.	Unclear validation procedure.	Accuracy: 0·8982.	No
Li et al. [23]	Diagnosis	CXR	DL	429 images, 143 COVID-19	108 images, 36 COVID-19.	Internal holdout validation.	Accuracy: 0·89 AUC: 0·992	Yes (Yes)
Ezzat et al. [18]	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. ***(w): weighted average***	No
Tartaglione et al. [16]	Diagnosis	CXR	DL	231 images, 126 COVID-19.	135 images, 90 COVID-19.	Internal holdout validation.	Sensitivity: 0·61, Specificity: 0·71 F-score: 0·70, Accuracy: 0·64, Balanced Accuracy: 0·66, AUC: 0·67, Odds-Ratio = 3·87.	No

Luz et al. [19]	Diagnosis	CXR	DL	13,539 images, 152 COVID-19.	231 images, 31 COVID-19.	Internal holdout validation.	Accuracy = 0·939, Sensitivity: 0·968, PPV: 1·000.	Yes (Yes)
Bassi et al. [20]	Diagnosis	CXR	DL	2724 images, 159 COVID-19.	180 images, 60 COVID-19.	Internal holdout validation	Recall: 0·978, Precision: 0·978.	No
Kana et al. [21]	Diagnosis	CXR	DL	Unclear in the paper.	Unclear in the paper.	External validation.	Accuracy: 0·99, Recall: 0·99, Precision: 0·99, F1: 0·99	No
Heidari et al. [22]	Diagnosis	CXR	DL	8474 images, 415 COVID-19.	848 images, 42 COVID-19.	Internal holdout validation.	Precision (w): 0·94, Recall (w): 0.94, F1-score (w): 0·94. ***(w): weighted average***	No
Zokaeinikoo et al. [24]	Diagnosis	CXR and CT	DL	Unclear in the paper.	Unclear in the paper.	Ten-fold internal cross-validation.	Accuracy: 0·992, Sensitivity: 0·993, Specificity: 0·9998, PPV: 0·996	No
Amyar et al. [30]	Diagnosis	CT	DL	944 patients, 399 COVID-19.	100 patients, 50 COVID-19.	Internal holdout validation.	Accuracy: 0·86, Sensitivity: 0·94, Specificity: 0·79, AUC: 0·93.	No

Ardakani et al.[31]	Diagnosis	CT	DL	Unclear as splits do not total correctly.	Unclear as splits do not total correctly.	Internal holdout validation	AUC: 0·994, Sensitivity, 1·000, Specificity, 0·9902, Accuracy, 0·9951; PPV, 0·9903, NPV, 1·000.	No
Bai et al. [47]	Diagnosis	CT	DL	118,401 images, 60,776 COVID-19.	14,182 images, 5,030 COVID-19.	Internal holdout validation	AUC: 0·95, Accuracy: 0·96, Sensitivity: 0·95, Specificity: 0·96.	Yes (Yes)
Jin et al. [35]	Diagnosis	CT	DL	1136 images, 723 COVID-19.	282 images, 154 COVID-19.	Internal holdout validation.	Sensitivity: 0·974, Specificity: 0·922, AUC: 0·991	No
Wang et al. [28]	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 [0·90,0·69]. **External validation:** AUC: 0·81 [0·71,0·84].	No
Ko et al. [27]	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	Five-fold internal cross-validation and external validation.	**Internal validation:** Sensitivity: 0·9958, Specificity: 1·00, Accuracy: 0·9987, AUC: 1·00. **External validation:** Accuracy: 0·9697.	No

Acar et al. [33]	Diagnosis	CT	DL	2,552 images, 1,085 COVID-19.	580 images, 246 COVID-19.	Internal holdout validation.	Accuracy: 0·9980, Error: 0·005571, Precision: 0·9980, Recall: 0·9980, F1-Score: 0·9980, AUC: 0·999979.	No
Pu et al. [29]	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
Georgescu et al. [26]	Diagnosis	CT	DL and hand-engineered 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
Chen et al. [34]	Diagnosis	CT	DL	770 (CV) images, 413 COVID-19.	Cross-validation: 86 (CV) images, 46 COVID-19	Ten-fold internal cross-validation.	Accuracy: 0·881±0·012, Precision: 0·898±0·011, Recall: 0·882±0·011, AUC: 0·942±0·012.	No
Guiot et al. [40]	Diagnosis	CT	Hand-engineered radiomic features	Unclear in the paper.	Unclear in the paper.	Internal holdout validation.	Sensitivity: 0·789, Specificity: 0·911, Accuracy: 0·897 [0·84, 0·939], AUC = 0·939 [0·875,1·00]	No
Shi et al. [39]	Diagnosis	CT	Hand-engineered radiomic features	2,148 (CV) images, 1,326 COVID-19.	Cross-validation: 537 (CV) images, 332 COVID-19	Five-fold internal cross-validation.	AUC: 0·942, Sensitivity: 0·907, Specificity: 0·833, Accuracy 0·879.	No

Mei et al. [32]	Diagnosis	CT	DL and CNN extracted features and clinical data	626 images, 285 COVID-19.	279 images, 134 COVID-19.	Internal holdout validation.	Sensitivity: 0·843 [0·771, 0·900], Specificity: 0·828 [0·756, 0·885], AUC = 0·92 [0·887,0·948].	Yes (Yes)
Chen et al. [41]	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: 0·936 [0·866,1·000], Accuracy: 0·763, Specificity: 0·789, Sensitivity: 0·737.	No
Wang et al. [36]	Diagnosis and prognosis for length of hospital stay.	CT	Detection 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 COVID-19. Validation 4: 117 images, All COVID-19.	External validation.	**Validation 1:** AUC: 0·87 **Validation 2** AUC: 0·88 **Validation 3:** KM separation p = 0·013 **Validation 4:** KM separation p = 0·014	Yes (Yes)
Cohen et al. [46]	Prognosis of lung opacity and extent of lung involvement with GGOs for COVID-19 patients.	CXR	Features from a trained CNN extracted at various layers.	47 patients, All COVID-19.	47 patients, All COVID-19.	Internal holdout validation.	Opacity correlation: 0·80 Extent correlation: 0·78	Yes (Yes)

Qi et al. [44]	Prognosing hospital stay in for COVID-19 patients	CT	Hand-engineered radiomic features.	25 (CV) patients, All COVID-19.	Cross-validation: 6 (CV) patients All COVID-19.	Five-fold internal cross-validation.	AUC: 0·97 [0·83,1·00], Sensitivity: 1·00, Specificity: 0·89, NPV: 1·00, PPV: 0·80.	Yes**
Zhu et al. [45]	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	Five-fold internal cross-validation run 20 times, average reported.	AUC: 0·8591±0·0227 Accuracy: 0·8569±0·0220 Sensitivity: 0·7697±0·0336 Specificity: 0·8802±0·0145	No
Lassau et al. [43]	The prognostic model used for predicting the risk of death, need for ventilation and ICU.	CT, clinical data	CNN extracted features and clinical data	646 patients, All COVID-19.	Internal validation: 150 images, All COVID-19. External validation: 137 patients, All COVID-19.	Internal holdout validation and external validation.	**Internal validation:** AUC: 0·76 **External validation:** AUC: 0·70	No
Chassagnon et al. [42]	Short term outcome prediction: intubation and death within 4 days	CT	Hand-engineered radiomic features	383 COVID-19 patients, 84 severe outcomes.	95 COVID-19 patients, 26 severe outcomes.	External validation.	Precision (w): 0·9, Sensitivity (w): 0·96, Specificity (w): 0·74, Balanced Accuracy: 0·81. ***(w): weighted***	No*

*** the authors state that "the developed tool will be made publicly available."**
**** the authors state that "imaging or algorithm data used in this study are available upon request."**

Appendix

1. Search strategy

Initial extraction. For the ArXiv papers, we initially extract papers for the relevant date ranges that include “ncov”, “corona”, “covid” or “SARS-CoV-2" in their title or abstract which are submitted in one of the following categories:

1. cs (computer science),
2. eess (electrical engineering and systems science),
3. math (mathematics),
4. q-bio (quantitative biology)
5. stat (statistics)

For the Living Systematic Review, we download all papers in the appropriate date range.

Refined search. We then filter 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”, “imaging”, “image*”, “radiograph*”

2. Supplementary author list

Name	Affiliation
Alessandro Ruggiero	Royal Papworth Hospital, Cambridge, UK; Qureight Ltd, Cambridge, UK
Anna Korhonen	Language Technology Laboratory, University of Cambridge, Cambridge, UK
Effrossyni Gkrania-Klotsas	Department of Infectious Diseases, Cambridge University Hospitals NHS Trust, Cambridge, UK
Emily Jefferson	Population Health and Genomics, School of Medicine, University of Dundee, Dundee, UK
Emmanuel Ako	Chelsea and Westminster NHS Trust and Royal Brompton NHS Hospital, London, UK
Georg Langs	Department of Biomedical Imaging and Image-guided Therapy, Computational Imaging Research Lab Medical University of Vienna, Vienna, Austria
Ghassem Gozaliasl	Department of Physics, University of Helsinki, Helsinki, Finland
Guang Yang	National Heart & Lung Institute, Imperial College London, London, UK
Helmut Prosch	Department of Biomedical Imaging and Image-Guided Therapy, Medical University of Vienna, Austria; Boehringer Ingelheim
Jacobus Preller	Addenbrooke’s Hospital, Cambridge University Hospitals NHS Trust, Cambridge, UK
Jan Stanczuk	Department of Mathematics and Theoretical Physics, Cambridge University, Cambridge, UK
Jing Tang	Research Program in System Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland
Johannes Hofmanninger	Department of Biomedical Imaging and Image-guided Therapy, Computational Imaging Research Lab Medical University of Vienna, Vienna, Austria
Judith Babar	Addenbrooke’s Hospital, Cambridge University Hospitals NHS Trust, Cambridge, UK
Lorena Escudero Sánchez	Department of Radiology, University of Cambridge, UK; Cancer Research UK Cambridge Institute, Cambridge, UK
Muhunthan Thillai	Interstitial Lung Disease Unit, Royal Papworth Hospital, Cambridge, UK; Department of Medicine, University of Cambridge, Cambridge, UK

Paula Martin Gonzalez	Cancer Research UK Cambridge Centre, University of Cambridge, Cambridge, UK
Philip Teare	Biopharmaceuticals R&D, AstraZeneca, Cambridge, UK
Xiaoxiang Zhu	Signal Processing in Earth Observation, Technical University of Munich, Munich, Germany
Mishal Patel	Biopharmaceuticals R&D, AstraZeneca, Cambridge, UK
Conor Cafolla	Department of Chemistry, University of Cambridge, Cambridge, UK
Hojjat Azadbakht	Ainostics Ltd, Manchester, UK
Joseph Jacob	Centre for Medical Image Computing, University College London, London, UK
Josh Lowe	SparkBeyond UK Ltd, London, UK
Kang Zhang	Center for Biomedicine and Innovations at Faculty of Medicine, Macau University of Science and Technology, Macau, China
Kyle Bradley	SparkBeyond UK Ltd, London, UK
Marcel Wassin	contextflow GmbH, Vienna, Austria
Markus Holzer	contextflow GmbH, Vienna, Austria
Kangyu Ji	Cavendish Laboratory, University of Cambridge, Cambridge, UK
Maria Delgado Ortet	Department of Radiology, Cambridge University, Cambridge, UK
Tao Ai	Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
Nicholas Walton	Institute of Astronomy, University of Cambridge, Cambridge, UK
Pietro Lio	Department of Computer Science and Technology, University of Cambridge, Cambridge, UK
Samuel Stranks	Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, UK
Tolou Shadbahr	Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland
Weizhe Lin	Department of Engineering, University of Cambridge, Cambridge, UK
Yunfei Zha	Department of Radiology, Renmin Hospital of Wuhan University, Wuhan, China
Zhangming Niu	Aladdin Healthcare Technologies Ltd, London, UK