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

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Artificial intelligence for imaging-based COVID-19 detection: systematic review comparing added value of AI versus human readers

Christine Kriza, Valeria Amenta, Alexandre Zenié, Dimitris Panidis, Hubert Chassaigne, Patricia Urbán, Uwe Holzwarth, Aisha Vanessa Sauer, Vittorio Reina, Claudius Benedict Griesinger

PII: S0720-048X(21)00509-X
DOI: https://doi.org/10.1016/j.ejrad.2021.110028
Reference: EURR 110028

To appear in: European Journal of Radiology

Received Date: 31 August 2021
Revised Date: 4 November 2021
Accepted Date: 9 November 2021

Please cite this article as: C. Kriza, V. Amenta, A. Zenié, D. Panidis, H. Chassaigne, P. Urbán, U. Holzwarth, A. Vanessa Sauer, V. Reina, C. Benedict Griesinger, Artificial intelligence for imaging-based COVID-19 detection: systematic review comparing added value of AI versus human readers, European Journal of Radiology (2021), doi: https://doi.org/10.1016/j.ejrad.2021.110028

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 Published by Elsevier B.V.
Artificial intelligence for imaging-based COVID-19 detection: systematic review comparing added value of AI versus human readers

Authors:
Dr Christine Kriza1*, Dr Valeria Amenta2, Dr Alexandre Zenié3, Dr Dimitris Panidis4,
Dr Hubert Chassaigne5, Dr Patrícia Urbán6, Dr Uwe Holzwarth7, Dr Aisha Vanessa Sauer8, Vittorio Reina9, Dr Claudius Benedict Griesinger10
1-10 European Commission, Joint Research Centre (JRC), Via E. Fermi 2749 (TP 281) Ispra, Lombardy, Italy

Email addresses of all authors:
Christine.KRIZA@ec.europa.eu; Valeria.AMENTA@ec.europa.eu; Alexandre.ZENIE@ec.europa.eu;
Dimitrios.PANIDIS@ec.europa.eu; Hubert.CHASSAIGNE@ec.europa.eu;
Patricia.URBAN-LOPEZ@ec.europa.eu; Uwe.HOLZWARTH@ec.europa.eu;
Aisha.SAUER@ec.europa.eu; Vittorio.REINA@ec.europa.eu; Claudius.GRIESINGER@ec.europa.eu

Contact details corresponding author Dr Christine Kriza:
European Commission, Joint Research Centre (JRC)
Directorate F - Health, Consumers and Reference Materials, Unit F.2 - Consumer Products Safety
Via E. Fermi 2749 (TP 281), 21027 Ispra (VA), Lombardy, Italy
+39-0332-78-5899, Christine.Kriza@ec.europa.eu

Abstract

Purpose: A growing number of studies have examined whether Artificial Intelligence (AI) systems can support imaging-based diagnosis of COVID-19-caused pneumonia, including both gains in diagnostic performance and speed. However, what is currently missing is a combined appreciation of studies comparing human readers and AI.

Methods: We followed PRISMA-DTA guidelines for our systematic review, searching EMBASE, PUBMED and Scopus databases. To gain insights into the potential value of AI methods, we focused on studies comparing the performance of human readers versus AI models or versus AI-supported human readings.
**Results:** Our search identified 1270 studies, of which 12 fulfilled specific selection criteria. Concerning diagnostic performance, in testing datasets reported sensitivity was 42-100% (human readers, n=9 studies), 60-95% (AI systems, n=10) and 81-98% (AI-supported readers, n=3), whilst reported specificity was 26-100% (human readers, n=8), 61-96% (AI systems, n=10) and 78-99% (AI-supported readings, n=2). One study highlighted the potential of AI-supported readings for the assessment of lung lesion burden changes, whilst two studies indicated potential time savings for detection with AI.

**Conclusions:** Our review indicates that AI systems or AI-supported human readings show less performance variability (interquartile range) in general, and may support the differentiation of COVID-19 pneumonia from other forms of pneumonia when used in high-prevalence and symptomatic populations. However, inconsistencies related to study design, reporting of data, areas of risk of bias, as well as limitations of statistical analyses complicate clear conclusions. We therefore support efforts for developing critical elements of study design when assessing the value of AI for diagnostic imaging.
Introduction

The field of medical imaging has seen rapid progress in the last years for the application of artificial intelligence (AI) methodologies, especially machine learning (ML) and deep learning (DL). AI systems have progressed in image-recognition tasks relevant to disease diagnosis and detection, thus mimicking expert data interpretation capacities. Recent studies have shown that AI-supported imaging technologies for specific diseases have a diagnostic performance comparable to medical experts.

According to the WHO, there have been over 247 million confirmed cases of COVID-19 globally as of November 3, 2021 and the pandemic has claimed over 5 million deaths worldwide. COVID-19 diagnosis is largely based on reverse transcriptase-polymerase chain reaction (RT-PCR) testing, but problems related to test kit availability, test reliability, and test turnaround time have persisted in many countries. Additional fast, low-cost, and easily scalable tools for triaging and detecting COVID-19 suspected patients are crucial. The COVID-19 pandemic has provided considerable momentum to this research area, with high expectations from the clinical community, which warrants an overview assessing the current evidence regarding the potential of AI methods to support the accurate, and fast detection of COVID-19 pneumonia. However, the accurate differentiation between COVID-19 and pneumonia of other origins remains challenging, due to subtle radiologic differences, especially in asymptomatic patients and those with early onset of symptoms.

We performed a systematic review of peer-reviewed publications that used AI for the evaluation of lung imaging to support the detection of COVID-19, and compared findings between a selected AI system and human readers, or AI-supported readings versus human readers alone, to obtain a comprehensive view of the current status of published evidence.
To our knowledge, this is the first published systematic review to date on this focused topic. Due to the large heterogeneity in the reporting of relevant results, the information was not amenable to quantitative data grouping or meta-analysis. We therefore focused our study on the following aims: (a) firstly, to compare the evidence from published studies that used AI methodologies for supporting the detection of COVID-19 in lung imaging and also included an element of comparison between the performance values of human readers versus AI models or human readers versus AI-supported readings, (b) secondly, to look into the consistency of the reporting of outcomes across different studies, and (c) finally, to assess the risk of bias in the published studies according to a standard appraisal tool.
Material and Methods

Search strategy and selection criteria

Our systematic review was performed in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy (PRISMA-DTA), including the related checklist.6,7 We used the systematic review methods outlined in the York Centre for Reviews and Dissemination’s guidance for undertaking reviews in healthcare, specifically focusing on the guidance for systematic reviews of clinical tests.8

The relevant population included COVID-19 pneumonia positive patients versus those with other pneumonia symptoms that tested negative for COVID-19. The intervention or index test was imaging, which was evaluated by both AI and human readers. As a comparator or reference standard, a positive PCR test was most commonly used in current practice, but in some cases they were supported by radiologists’ findings and a clinical consensus. We were interested in the outcome measures related to the support of COVID-19 detection, comparing AI and human readers, i.e. sensitivity, specificity per patient, and related values, such as AUC (area under the Receiver Operator Characteristic -ROC- curve). We also explored additional outcomes, such as time required for diagnosis.

We only included peer-reviewed studies in our review. The included studies had to focus on supporting the detection of COVID-19 with a lung-imaging modality, and applying an AI methodology for the analysis of the imaging outcomes, including ML and DL. Furthermore, the studies had to include a comparison between human readers versus AI models or human readers versus AI-supported readings and report outcomes related to sensitivity and specificity at least, but ideally also further outcome measures. We excluded studies not
focusing on lung imaging of COVID-19, for example those using ex-vivo imaging and pathology studies, and those focused on aspects related to segmentation, features extraction, treatment, survival and disease risk prediction (see appendix, table 1 for full list of selection criteria).

We conducted a systematic search of the literature including EMBASE, PubMed and Scopus databases for papers published from January 1, 2019 in the English language (see appendix, table 2 for search strings). The searches were performed on November 30, 2020. Additional papers were identified through automated database notifications of new publications according to our search terms until January 31, 2021. The reference lists from all included papers were checked to identify and include any other potential studies.

Two reviewers independently performed a screening of the citations by title and abstract, with discrepancies resolved by consensus. The citations identified in the systematic search were uploaded to EndNote reference manager and duplicates were automatically deleted.

Quality assessment

Two reviewers acquired the full-text versions of the included papers and independently assessed their methodological quality using the QUADAS-2 tool, which has been adapted to the systematic review objectives (see overview of signalling questions in appendix, table 3). Any discrepancies were resolved through discussion between the reviewers. QUADAS-2 is an appraisal tool recommended for the use in systematic reviews to evaluate the risk of bias in diagnostic accuracy studies.\(^9\) QUADAS-2 consists of the following four domains: patient selection, index test, reference standard and flow and timing. We used an additional domain of data management and assessed all five domains for risk of bias using relevant signalling questions. While QUADAS-2 is not intended to generate an overall score, the tool highlights
a high, medium or unclear risk of bias according to the domains assessed. We excluded studies, with a high risk of bias in at least two domains, in the final analysis of our systematic review. In addition, one reviewer evaluated the methodologies applied for statistical analysis on all available datasets, where appropriate performance measurements were reported in the included papers.

**Data analysis**

Two reviewers extracted the following information on study characteristics for each paper, including: study category, imaging modality, country where the imaging took place, experience of human readers, selection criteria in the original research studies, reference standards used and information on blinding, as well as sample sizes for different datasets and validation type (Table 1a). In addition, information on the AI method and the data source, together with the data acquisition time period were extracted (Table 1b). Along with the original study authors’ original conclusions, the following outcome data were extracted per-patient where applicable: values for sensitivity, specificity, accuracy, positive predictive value, negative predictive value and AUC. Where indicated, other values were extracted for speeds of reading and potential time savings (Table 2).
Results

The search strategy identified a total of 1261 articles on November 30, 2020, with a further six studies identified via database notifications of the same search until January 31, 2021. Overall, 771 articles were screened for meeting the selection criteria. A total of 20 studies that met these criteria were assessed for the risk of bias with the QUADAS-2 tool. Eight studies with a high risk of bias rating in at least two domains were excluded (figure 1, appendix, table 3). A total of 12 studies were included in our systematic review. (See figure 2 for the study selection and appendix, table 4 for the PRISMA-DTA checklist and appendix, table 5 for information on statistical analysis).

Study characteristics

We identified that seven studies used computed tomography (CT) as the main imaging modality, with the remaining five studies focusing on chest X-ray (CXR) imaging. (See table 1a for details on study characteristics). All studies applied a DL model for their AI methodology that included neural networks (table 1b). For the majority of four studies, China was indicated as the country where the imaging took place, with two studies using imaging data from both China and the United States. In terms of data source origins, imaging data was considered from various geographical locations, with a focus on patients from the United States for two studies, two studies from the Netherlands, one study from Italy, and one study from Hong Kong.5

There was considerable heterogeneity between the study designs, data collection and patient selection criteria applied throughout the studies, i.e. with different inclusion criteria for the selection of COVID-19 positive patients and the related characteristics for imaging required, as well as the use of automatic assessment scoring and subgroup analysis.
For the reference standard indicated in the included papers, three studies used RT-PCR, two studies applied consensus findings between readers, two studies provided outcome values for both these options, and in four studies the use of the reference standard was not sufficiently detailed, but included the above (see table 1a). All studies were based on patient selection with the diagnosis of an RT-PCR or nucleic acid amplification (NAAT) test. The total number of human readers ranged from two readers to a panel of ten, with practice levels ranging from less than five years to over 30 years of experience in thoracic imaging.

In ten studies, various hospitals, academic centres and clinics were indicated as the data source. Two studies used a mix of data from publicly available databases as well as from hospitals. There was a large variety of patients in the total numbers used in the overall datasets, ranging from 216 to 25,146 patients. Studies that used CXR as the imaging modality had generally larger sample sizes. All studies used differentiated datasets for training, testing and validation. Ten studies provided performance measurements for testing datasets, reporting 15 different sample sizes between 18 and 2,193 patients. Four studies reported values for validation, with sample sizes ranging between 18 and 910 patients. While three studies provided performance measurements for external validation, one study did this for independent internal validation.

In eight out of 12 studies, there was a larger representation of men in the patient characteristics, ranging from 54-59% of males of the total number of patients, and up to 65.8% for specific datasets used for training. Three studies reported equal numbers of males and females, and with 53%, one study indicated a larger representation of women in the patient characteristics.
Age values were reported in a heterogeneous way across the studies. Three studies reported mean ages for the total patient population,\(^{17,20,28}\) ranging from 40.7\(^{28}\) to 58 years.\(^{17}\) Three studies reported mean ages for COVID and non-COVID groups,\(^{14,16,19}\) with one study indicating a younger average age of 48 years in the COVID group vs. 62 years in the non-COVID group,\(^{16}\) and two studies reporting more balanced ages for the two groups.\(^{18,21}\)

**Performance outcomes**

We focused on the performance data from 126 scenarios, reporting sensitivity (118 for specificity) in 12 studies, and 65 scenarios for AUC values in 11 studies, applying our grouping of the categories into human readers, AI models and AI-supported human readings, visualised in figure 4. (See figure 5 for a comparison across different datasets and table 4 for an overview of the quantitative results).

**General findings**

For overall conclusions, as claimed by the authors in the 12 original studies, six reported a diagnostic performance of the AI model which was comparable to human readers, with three focused on CT,\(^{26-28}\) and three with CXR as an imaging modality (figure 3 and table 2).\(^{17,20,21}\) Two other studies using CXR indicated that their AI model outperformed human readers.\(^{5,18}\) Compared to a reader-only approach, three studies reported that AI augmentation improved human readers’ performance using CT as an imaging modality.\(^{16,23,24}\) One study reported positive results of the AI system to aid radiologists in the assessment of changes of the lung lesion burden on pairs of CT scans, with comparable performance of the AI model to human readers.\(^{22}\)

**Sensitivity**
Testing datasets

Ten of the studies reported sensitivity with 42-100% for human readers, and 60-95% of AI models.\textsuperscript{16-18,20-24,26,27} Three studies examined AI-supported human readings, with sensitivity values from 81-98%.\textsuperscript{16,23-24} Median sensitivity was the highest for AI-supported human readings (91.3%), with lower values observed for AI models (85%) and human readers (71.1%). The interquartile range (IQR) of the above studies was lowest for AI-supported human readings (8.8%) and AI models (11.3%), with a higher value for human readers (22.3%).

Validation datasets

Three of the studies reported sensitivity at 49-100% for human readers,\textsuperscript{5,22,27} with four studies indicating ranges of 71.2-96.3% for AI models.\textsuperscript{5,22-23,27} No values were provided for AI supported human readings. Median sensitivity was higher in AI models (79.4%) than for human readers (64.2%). The IQR of the above studies was lower in AI models (14.8%) than for human readers (46.4%).

Specificity

Testing datasets

Eight studies reported specificity with 26-100% for human readers,\textsuperscript{16-18,20-21,23,26,28} while ten reported values of 61-96% for AI models.\textsuperscript{16-18,20-21,23-24,26-28} Two studies examined AI-supported human readings, with specificity values from 78-99%.\textsuperscript{16,23} Median specificity was highest for AI-supported human readings (91.1%), with lower values observed for AI models (86.5%) and human readers (85%). The IQR of the above studies was lowest for AI-supported human readings (6.1%) and AI models (10.9%), with a higher value for 21.8% for human readers.
*Validation datasets*

Three of the studies reported specificity with 72-97.9% for human readers, and 57.9-89.2% for AI models. No values were provided for AI supported human readings. Median specificity was 81% for AI models and 82.2% for human readers. The IQR of the above studies was lower in AI models (9.6%) than for human readers (18.7%).

*AUC values*

*Testing datasets*

Five of the studies reported AUC ranges of 0.73-0.963 for human readers, with ten studies indicating ranges of 0.80-0.97 for AI models. One study examined AI-supported human readings, reporting an AUC of 0.966. AI models had median values of 0.92 and human readers of 0.85. The IQR of the above studies was 0.07 for human readers and 0.09 for AI models.

*Validation datasets*

Two studies reported AUC ranges of 0.61-0.802 for human readers, with three studies indicating values of 0.71-0.98 for AI models. No values were provided for AI supported human readings. Median AUC values were higher in AI models (0.87) compared to human readers (0.64). The IQR of the above studies was lower for AI models (0.05) than for human readers (0.121).

*Comparison of performance between testing and validation datasets*

Only scenarios for AI models and human readers were comparable between testing and validation datasets. AI-supported readings showed the highest median values, and in general, the median for sensitivity, specificity and AUC values, was higher in the testing datasets. The
IQR for sensitivity in general was lower in the testing datasets, while being comparable for specificity. Only for comparable datasets in two studies, related to testing and external validation, the medians in the testing datasets were higher, showing an overall reduced performance in external validation (figure 5 j-).

**Time savings**

AI-supported triage improved the efficiency of scan-to-fever-clinician triage at each hospital in the study by Wang and colleagues, with a median reduction in triage time ranging between 18.77 and 198.28 minutes in different hospitals\textsuperscript{22}. Wehbe and colleagues reported that the time to analyse a data subset with AI took approximately 18 minutes, compared to approximately 2.5–3.5 hours for each radiologist.\textsuperscript{17}
Discussion

Our study shows promising performance results of AI-supported detection of COVID-19 imaging, specifically comparing the performance of AI and human readers, however these need to be interpreted in the context of risks of bias. The medians of all performance values in the testing datasets were in general the highest for AI-supported human readings, followed by AI models, and then human readers alone, as shown in Table 4. However, human readers alone reached the highest maximum, but also the lowest minimum values in testing datasets, with the latter especially notable for specificity. Variability of diagnostic performance focused on the IQR was in general lower in AI systems or AI-supported readings. In addition, some studies reported time savings with AI models, as detailed in Table 2. Reporting related to speedier analysis with AI also included AI-supported triage, whereby time-to-triage was faster compared to a standard clinical workflow across different clinics. Notably, this implied an ideal scenario where clinicians would respond instantly to AI notifications and could thereby potentially shorten the time to diagnosis, with multiple benefits for the isolation and the treatment of affected patients.

In the validation datasets, there was an overall reduction in performance in comparison to the testing datasets related to median values, as shown in Figure 5. However, a detailed analysis of the same performance measurements across all studies was difficult, as most studies did not include a complete reporting on all performance measurements disaggregated by datasets and differentiating all scenarios for human readers, AI models and AI-supported readings. The fragmented reporting related to performance measurements is detailed in the supplementary appendix, table 5. Notably, there was large heterogeneity in the methodology and selection criteria applied, including patient characteristics and sample sizes, as well as levels of patient detail disaggregation (see table 1a and 1b). Most of the studies in our review
focused on differentiating COVID-19 pneumonia from other types of pneumonia, showing a potential added value of imaging as a supplementary diagnostic measure, in the context of fast and accurate reporting and earlier identification of potentially infected COVID-19 patients. However, there were differences in how studies attempted to evaluate these merits, i.e. through a comparison of AI and human readers as well as AI-supported readings versus human readings alone. The studies also applied different strategies regarding subgroup analysis, and the differentiation of training, (external) testing and (external) validation. In order to ensure the robustness of final results, future studies should clearly describe the splitting of all datasets for training, testing and validation, not only subsets of these (i.e. training and validation only). In addition, the documentation of relevant patient characteristics should be consistently applied for all datasets. A variety of different approaches were used for the scenarios of human readers, including different levels of experience, as well as thresholds and cut-off points (see figure 4). The heterogeneity in scenarios also applies to AI models, which reported different model types, design features, classifiers and reference standards. Due to such heterogeneity in the study methodology and analysis applied in the different studies and the resulting lack of comparability, it is difficult to make solid conclusions regarding improved performance measurements with AI. There were also major differences in baseline characteristics between patient groups with COVID-19, and those with non–COVID-19 pneumonia, introducing possible selection bias with imbalances regarding gender and age, with a related overrepresentation of younger males, including also different datasets within a study as shown in table 1a. In addition, reporting on the details of patient characteristics for different cohorts varied, for example there was limited information on patients who may have been immunosuppressed.
Notably, some CT-focused studies featured smaller sample sizes for certain datasets – mostly for validation\textsuperscript{23,27}, and patients with early-stage COVID-19. The limitations related to patient populations were further complicated by variations in image quality, and heterogeneity in imaging acquisition and post-processing parameters as shown in our QUADAS-2 assessment. Future studies should focus on testing AI models with data from more populations and geographical areas. Imaging should also include more data and information on disaggregation related to all ethnicities, and not limited to Caucasian and Asian races\textsuperscript{21}, as well as older patients, with more equal gender representation to allow for further generalisations. To assess and validate the robustness of AI models, training with larger multi-centre datasets and consistent external validation is required, as well as more prospective study data and evaluations\textsuperscript{26,29}.

Overly positive interpretations of promising performance results attributed to AI models or AI-supported readings should be treated with caution, as there is a risk of overestimating results due to several potential confounding factors. As such, Chiu and colleagues have discussed how the use of RT-PCR as the ground truth for training models may not reflect the real performance of AI systems, since false-negative rates of RT-PCR have been reported to be as high as 30\%\textsuperscript{5}. In addition, there is a potential susceptibility of AI systems to ‘learn’ dataset characteristics instead of disease classification, for example by introducing a bias described as ‘shortcuts learning’.\textsuperscript{30} In these instances, models may rely on features that are not related to correct object classification for a disease pathology, by relying on differences in the background instead, such as textual markers in obtained images, for example related to patient positioning\textsuperscript{31}. Another area of attention is related to an intrinsic feature of AI, specifically its “black box” character: it is difficult to rule out that an algorithm is not using findings outside the lungs when discriminating for COVID-19.\textsuperscript{5} The black box problem
emphasizes another value of human readers, namely that results are difficult to explain.\textsuperscript{27} Moreover, human analysis is required to rule out motion artefacts which may cause errors when detecting COVID-19.\textsuperscript{26}

Our QUADAS-2 analysis highlighted the need to address possible risk of bias in future studies, especially related to a clear need for adequate and consistent descriptions of patient selection and study population description. Further areas of attention are clear information on the use of index tests, reference standards and flow and timing between these. In the context of imaging, detailed descriptions of imaging acquisition and processing methods are required. In addition, clear descriptions of data sources are important. An analysis of the statistical methods applied in the studies revealed further methodological concerns, such as the inconsistent use of 95\% confidence intervals for all performance values, or of \textit{p}-values for interpreting lack of significance, as shown in Table 5 of the supplementary appendix. We found considerable limitations concerning the methods applied to compare performance between AI systems and human readers. For example, the selection of sample sizes can strongly influence performance measurements and only one study analysed the threshold above which no performance gain could be obtained.\textsuperscript{18} Several studies did not consistently apply the agreement index, using instead, average values of reader performance. To enable appropriate inter-rater agreement, studies should focus on \textit{agreement} rather than \textit{correlation} indices.\textsuperscript{32,33} In addition, future statistical analysis should consistently analyse the disease prevalence, for example, by using Negative Predictive Value and Positive Predictive Value parameters, as well as positive and negative likelihood ratios to reduce uncertainties regarding the validity of diagnostic tests,\textsuperscript{34} with the latter ratios only reported in one of the 12 studies.\textsuperscript{21} In this way, meaningful analysis of the overall merit of AI-supported imaging-based COVID-19 detection beyond high-prevalence populations could be ensured.
Future studies need to address concerns of AI methodology regarding transparency, reproducibility, ethics and effectiveness, as well as specific reporting standards for AI studies. This includes attention to critical elements that are required for coherent use of methodology, including study design, clear description of data sets and patient characteristics as well as consistent application of assessment methods that analyse potential risks of bias.

We acknowledge certain limitations in our study. We only included peer-reviewed studies in English from the published literature, potentially increasing the likelihood of publication bias influencing our findings. We did not evaluate the methodology of the AI and DL approaches applied in the reported studies, which may further affect the risk of bias. Importantly, we were unable to provide a meta-analysis of the studies included in our review, due to a large heterogeneity in the methodologies and results of the studies, precluding conclusions on pooled diagnostic accuracy of AI versus human readers.
Conclusions

While the studies included in our systematic review reported promising results of AI, these need to be seen in the context of COVID-19 severity, overall disease prevalence in a population, as well as several risks of bias regarding study methodologies applied. For early-onset patients and those who are asymptomatic or have mild disease, the performance of imaging may not be satisfactory. AI-supported imaging would therefore be most promising in high prevalence areas, and as a support tool for rapid diagnosis to identify suspected patients as a priority within a triage setting.\textsuperscript{17,21,28} At the same time, the involvement of human readers remains crucial.

Our study identified and highlighted several inconsistencies of data reporting and presentation, heterogeneity for study methodologies and evaluation methods applied, certain areas of risk of bias (e.g. selection bias, “shortcut learnings” bias) and limitations of statistical analyses.

While our study presented an overall promising potential of AI models for COVID-19 imaging diagnostic decision-making, this can only be confirmed by future studies with an improved and harmonised overall methodology that allows for reliable generalisability of results. In this context, we support efforts for the development and future implementation of a set of methodological ‘critical elements’ regarding study design, and assessment methodology to facilitate data aggregation, comparability and conclusions of the possible added value of AI-based imaging systems for diagnostic decision-making.
Acknowledgements

We thank Marta Da Pian for her support regarding the development of the database search strings in her capacity as an Elsevier search professional. We also thank Úna Cullinan and Roisin McMorrow for a linguistic review of our manuscript.

Funding

This work was supported by the European Commission Joint Research Centre (JRC) within the Consumer Products Safety Unit of Directorate for Health, Consumers and Reference Materials through the JRC’s Work Programme for 2020-2021, running under Horizon 2020, the current EU Framework Programme for research and innovation funding.

None of the authors have any conflicts of interests to declare. The views expressed are those of the authors and not necessarily those of the European Commission.
References

1. Choy G, Khalilzadeh O, Michalski M, Do S, Samir AE, Pianykh OS, et al. Current Applications and Future Impact of Machine Learning in Radiology. Radiology. 2018 Aug;288(2):318–28.

2. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nature reviews Cancer. 2018 Aug;18(8):500–10.

3. Shen J, Zhang CJP, Jiang B, Chen J, Song J, Liu Z, et al. Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review. JMIR medical informatics. 2019 Aug;7(3):e10010.

4. World Health Organization. WHO Coronavirus Dashboard [Internet]. [cited 2021 November 3]. Available from: https://covid19.who.int/

5. Chiu WHK, Vardhanabhuti V, Poplavskiy D, Yu PLH, Du R, Yap AYH, et al. Detection of COVID-19 Using Deep Learning Algorithms on Chest Radiographs. Journal of thoracic imaging. 2020 Sep.

6. Frank RA, Bossuyt PM, McInnes MDF. Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy: The PRISMA-DTA Statement. Vol. 289, Radiology. United States; 2018. p. 313–4.

7. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Medicine [Internet]. 2009 Jul 21;6(7):e1000097. Available from: https://dx.plos.org/10.1371/journal.pmed.1000097

8. Centre for Reviews and Dissemination. Systematic Reviews [Internet]. 2009. Available from: https://www.york.ac.uk/media/crd/Systematic_Reviews.pdf

9. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of internal medicine. 2011 Oct;155(8):529–36.

10. Ni Q, Sun ZY, Qi L, Chen W, Yang Y, Wang L, et al. A deep learning approach to characterize 2019 coronavirus disease (COVID-19) pneumonia in chest CT images. European radiology. 2020 Dec;30(12):6517–27.

11. Javor D, Kaplan H, Kaplan A, Puchner SB, Krestan C, Baltzer P. Deep learning analysis provides accurate COVID-19 diagnosis on chest computed tomography. European journal of radiology. 2020 Dec;133:109402.

12. Xie Q, Lu Y, Xie X, Mei N, Xiong Y, Li X, et al. The usage of deep neural network improves distinguishing COVID-19 from other suspected viral pneumonia by clinicians on chest CT: a real-world study. European radiology. 2020 Dec;1–10.

13. Yang S, Jiang L, Cao Z, Wang L, Cao J, Feng R, et al. Deep learning for detecting corona virus disease 2019 (COVID-19) on high-resolution computed tomography: a pilot study. Annals of translational medicine. 2020 Apr;8(7):450.

14. Sukhiija A, Mahajan M, Joshi PC, Dsouza J, Seth NDN, Patil KH. Radiographic findings in COVID-19: Comparison between AI and radiologist. The Indian journal of radiology & imaging. 2021 Jan;31(Suppl 1):S87–93.
15. Krishnamoorthy S, Ramakrishnan S, Colaco LB, Dias A, Gopi IK, Gowda GAG, et al. Comparing a deep learning model’s diagnostic performance to that of radiologists to detect Covid-19 features on chest radiographs. The Indian journal of radiology & imaging. 2021 Jan;31(Suppl 1):S53–60.

16. Bai HX, Wang R, Xiong Z, Hsieh B, Chang K, Halsey K, et al. Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT. Radiology [Internet]. 2020/04/27. 2020 Sep;296(3):E156–65. Available from: https://pubmed.ncbi.nlm.nih.gov/32339081

17. Wehbe RM, Sheng J, Dutta S, Chai S, Dravid A, Barutcu S, et al. DeepCOVID-XR: An Artificial Intelligence Algorithm to Detect COVID-19 on Chest Radiographs Trained and Tested on a Large U.S. Clinical Data Set. Radiology. 2021 Apr;299(1):E167–76.

18. Zhang R, Tie X, Qi Z, Bevins NB, Zhang C, Griner D, et al. Diagnosis of Coronavirus Disease 2019 Pneumonia by Using Chest Radiography: Value of Artificial Intelligence. Radiology. 2021 Feb;298(2):E88–97.

19. Wang Z, Xiao Y, Li Y, Zhang J, Lu F, Hou M, et al. Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest X-rays. Pattern recognition. 2021 Feb;110:107613.

20. Murphy K, Smits H, Knoops AJG, Korst MBJM, Samson T, Scholten ET, et al. COVID-19 on Chest Radiographs: A Multireader Evaluation of an Artificial Intelligence System. Radiology. 2020 Sep;296(3):E166–72.

21. Castiglioni I, Ippolito D, Interlenghi M, Monti CB, Salvatore C, Schiaffino S, et al. Machine learning applied on chest x-ray can aid in the diagnosis of COVID-19: a first experience from Lombardy, Italy. European radiology experimental. 2021 Feb;5(1):7.

22. Wang M, Xia C, Huang L, Xu S, Qin C, Liu J, et al. Deep learning-based triage and analysis of lesion burden for COVID-19: a retrospective study with external validation. The Lancet Digital health. 2020 Oct;2(10):e506–15.

23. Wang H, Wang L, Lee EH, Zheng J, Zhang W, Halabi S, et al. Decoding COVID-19 pneumonia: comparison of deep learning and radiomics CT image signatures. European journal of nuclear medicine and molecular imaging. 2020 Oct;1–9.

24. Yang Y, Lure FYM, Miao H, Zhang Z, Jaeger S, Liu J, et al. Using artificial intelligence to assist radiologists in distinguishing COVID-19 from other pulmonary infections. Journal of X-ray science and technology. 2021;29(1):1–17.

25. Dorr F, Chaves H, Serra MM, Ramirez A, Costa ME, Seia J, et al. COVID-19 pneumonia accurately detected on chest radiographs with artificial intelligence. Intelligence-Based Medicine [Internet]. 2020;3–4:100014. Available from: https://www.sciencedirect.com/science/article/pii/S2666521220300144

26. Lessmann N, Sánchez CI, Beenen L, Boulougne LH, Brink M, Calli E, et al. Automated Assessment of COVID-19 Reporting and Data System and Chest CT Severity Scores in Patients Suspected of Having COVID-19 Using Artificial Intelligence. Radiology. 2021 Jan;298(1):E18–28.
27. Zhou M, Yang D, Chen Y, Xu Y, Xu J-F, Jie Z, et al. Deep learning for differentiating novel coronavirus pneumonia and influenza pneumonia. Annals of translational medicine. 2021 Jan;9(2):111.

28. Mei X, Lee H-C, Diao K, Huang M, Lin B, Liu C, et al. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nature Medicine [Internet]. 2020;26(8):1224–8. Available from: https://doi.org/10.1038/s41591-020-0931-3

29. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Heal. 2019 Oct;1(6):e271–97.

30. Lópe-Z-Cabrera JD, Orozco-Morales R, Portal-Diaz JA, Lovelle-Enriquez O, Pérez-Díaz M. Current limitations to identify COVID-19 using artificial intelligence with chest X-ray imaging. Health and technology. 2021 Feb;1–14.

31. DeGrave AJ, Janizek JD, Lee S-I. AI for radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell. 2021; 3(7):610-9.

32. Gallagher EJ. Correlation versus agreement: methods of measurement in medicine. Vol. 27, Annals of emergency medicine. United States; 1996. p. 236–8.

33. Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: Key concepts, approaches, and applications. Research in Social and Administrative Pharmacy [Internet]. 2013;9(3):330–8. Available from: https://www.sciencedirect.com/science/article/pii/S1551741112000642

34. Asai T. COVID-19: accurate interpretation of diagnostic tests—a statistical point of view. Journal of Anesthesia [Internet]. 2020; Available from: https://doi.org/10.1007/s00540-020-02875-8

35. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. npj Digital Medicine [Internet]. 2021;4(1):65. Available from: https://doi.org/10.1038/s41746-021-00438-z
| Study Category | Imaging modality and country where imaging took place | Experience of human readers | Reference standard and information on blinding | Total dataset patient characteristics and sample size (where indicated or calculated) | Training dataset patient characteristics and sample size (where indicated or calculated) | Validation dataset patient characteristics and sample size (where indicated or calculated) | Validation type | Testing dataset patient characteristics and sample size (where indicated or calculated) |
|---------------|---------------------------------------------------|-----------------------------|------------------------------------------------|----------------------------------------------------------------------------------|----------------------------------------------------------------------------------|----------------------------------------------------------------------------------|---------------|----------------------------------------------------------------------------------|
| Lessmann (26) | Comparison of AI model with 8 human readers | CT, Netherlands 8 readers; 4 readers with less than 5 y, 4 readers up to 27 y | Reference standard not sufficiently detailed, AI model was compared with human readers and to clinical assessment of the patients, including RT-PCR testing Information on reader blinding provided | 843 476, mean age 60 (+/-16), male 56% | 476, mean age 60 (+/-16), male 56% | Not stated | Claimed external validation, but without providing performance measurements | Internal test set: 104/105, mean age 62 (+/-16), male 58%; External test set: 262, mean age 64 (+/-16), male 54% |
| Zhang Ran (18) | Comparison of AI model with 3 human readers | CXR, USA 3 readers, all experienced thoracic radiologists, with over 9, 14, and 34 y of experience | Reader consensus between three radiologists was used as reference standard Information on reader blinding provided | 5 208, out of which 2 060 with COVID-pneumonia (mean age 62 years, +/-16, 51% male) and 3 148 with non-COVID-pneumonia (mean age 64 years, +/-18, 50 1% male) | Training and validation combined: 3 015 | Training and validation combined: 3 015 | Claimed validation, but without providing performance measurements | 2193 |
| Study | Comparison | Country, Region | Readers | Data Description | Study Type | Validation Type | Details |
|-------|------------|----------------|---------|------------------|-------------|----------------|---------|
| Wang (22) | Comparison of AI model with a radiologist panel of 2 human readers; focus on the identification of lesion burden increase on pairs of CT scans obtained from 100 patients with COVID-19 | CT, China | n/a | Data for both RT-PCR and radiological findings as reference standards provided | 5,206 | Independent internal validation: 639, median 55 years (38-66), male 48%; external validation: 2120, median 43 years (31-56), male 51% | Information on reader blinding provided |
| Harrison (16) | Comparison of 6 human readers with and without assistance of AI model; performance measurements are also reported for the AI model only | CT, USA and China | 6 readers, 3 radiologists with 10 y experience, 3 radiologists with 20 y of chest CT experience | Reader consensus was used as reference standard | 1,958 | Independent internal validation | Information on reader blinding provided |
| Mei (28) | Comparison of AI model with 2 human readers | CT, China | 2 readers, one fellowship-trained thoracic radiologist with 10 y experience and one thoracic radiology fellow | RT-PCR was used as reference standard | Information on reader blinding provided | 905 (mean age 40.7 years, +/- 16.5 years), 54% males | Training: 534, tuning: 92 | Not stated | Not stated | 279 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Murphy (20) | Comparison of AI model with 6 human readers | CXR, Netherlands | 6 readers, all chest radiologists with experience ranging from 5 y up to more than 30 y | RT-PCR was used as reference standard | Information on reader blinding provided | 25 146 | 23 138, mean age 47.8 +/- 17.0 years, 57.0% male, 43.2% female, 0.2% unknown | 1 540, mean age 47.4 +/- 17.1 years, 56.7% male | Claimed validation, but without providing performance measurements | 468, mean age 67.3 +/- 14.4 years, 55.8% male |
| Study | Comparison of AI model with human readers | Country | Radiologists | Participants | Validation | Readers | N | Notes |
|-------|------------------------------------------|---------|--------------|--------------|------------|---------|---|-------|
| Wehbe (17) | Comparison of AI model with 5 human readers | CXR, USA | 5 readers, with 4 board-certified thoracic radiologists with experience ranging from 1-8 y and one board-certified diagnostic radiologist with 38 y of post-training experience | Results for both options with RT-PCR and using consensus interpretation as reference standard | 5 853 (mean age 58 +/- 19 years, 47% male) | Internal validation | 866 | Information on reader blinding provided |
| Zhou Min (27) | Comparison of AI models with 10 human readers in two groups (specialist-level and resident-level) | CT, China | 10 radiologists from two groups (group 1, 5 specialist-level radiologists with more than 15 y of experience; group 2, 5 resident radiologists with 3-5 y of experience) in thoracic imaging | Reference standard not sufficiently detailed | 449 (gender provided separately for NCP and IP groups, calculated 57% male, 43% female on the overall number) Mean age provided separately for influenza pneumonia group (IP) and novel coronavirus pneumonia group (NCP) | 33 independent internal + 107 external (calculated 60% male, 40% female on the external set) | Independent internal validation and external validation | 34 |
| Castiglioni (21) | Comparison of AI model with 2 human readers | CXR, Italy | 2 readers, 1 radiologist with 15 y of experience in chest imaging at Centre 1 and a general radiologist with 6 y of experience at Centre 2 | RT-PCR was used as reference standard | Information on reader blinding provided | 610 | Training & validation: 500; Mean age provided separately for COVID-19 and non COVID-19 group and for Centre 1 and Centre 2, gender provided separately for centre 1 (55% male) and 2 (65% male), calculated as 60% male and 40% female on the tot of 500 | Training & validation: 500 Mean age provided separately for COVID-19 and non COVID-19 group and for Centre 1 and Centre 2, gender provided separately for centre 1 (55% male) and 2 (65% male), calculated as 60% male and 40% female on the tot of 500 | Claimed validation, but without providing performance measurements | 110 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Wang Hongmei (23) | Comparison of 3 human readers with and without assistance of AI model, performance measurements are also reported for the AI model only | CT, China and USA | 3 readers with over 5 y of experience | Reference standard not sufficiently detailed | Information on reader blinding provided | 216 (59.3% male, age medians available for patients selected from three different hospitals, total age median 44 years) | 148 | 18 internal validation, 32 external validation | Independent internal and external validation datasets | 18 |
| Study | Comparision of AI model with 3 human readers | CXR, Hong Kong | 3 readers, all board-certified radiologists, each with >10 y experience | Reference standard not sufficiently detailed, RT-PCR as well as radiologic reference standard mentioned | 762 | 51-4, median age 65 (44-82), 52.3% male, 47.7% female | 248, median age 61 (39-79), 51% male, 49% female | Prospectively collected independent validation set representing a real-world mix of patients screened for COVID-19 | not stated |
|-------|-----------------------------------------------|----------------|---------------------------------------------------------------------|-----------------------------------------------------------------|------|---------------------------------------------------|-----------------------------------------------|------------------------------------------------------------------------|---------------------|
| Chiu 5 | Comparison of AI model with 3 human readers  | CT, China      | 3 radiologists with over 20 y, nearly 20 y and 5~10 y of chest CT experience | Final diagnosis decision from the hospital, also based on radiologists' readings, was used as reference standard | 694  | 509 (ages available for subgroups, average age 49 years, 65.8% male) | Not stated | Claimed validation, but without providing performance measurements | 185 (ages available for subgroups, average age 47, 61.6% male) |

**Legend**: AI: Artificial Intelligence, CT: Computed Tomography, CXR: Chest X-Ray, RT-PCR: Reverse transcription polymerase chain reaction, SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus-2, y: years

Table 1b: AI methods and data sources overview
| AI method                                                                 | Data source and data acquisition time period for model input                                                                 | Inclusion/Exclusion criteria in original research study                                                                 |
|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|
| Lessmann (26)                                                            | Patients were included from an academic center and a large teaching hospital in the Netherlands in March and April of 2020 who met inclusion criteria. | **Inclusion criteria:** symptoms of lower respiratory tract infection including cough, clinically relevant dyspnoea requiring hospital admission and fever with anosmia; **exclusion criteria:** no CT severity scores in radiology report, COVID-19 prior to imaging, or missing RT-PCR test results. RT-PCR was used to confirm COVID-19 patients. |
| Zhang Ran (18)                                                           | Data from hospital chain in US, Henry Ford Health System, including 5 hospitals and more than 30 clinics (February 1 to May 30, 2020). | **Inclusion criteria for non-COVID-19 pneumonia group:** patients that underwent frontal view CXR, had pneumonia diagnosis, and imaging was performed in specified time period; **inclusion criteria for COVID-19 group:** patients that underwent frontal view CXR, with a RT-PCR positive test for SARS-CoV-2 with a diagnosis of pneumonia; **exclusion criteria:** patients under the age of 18 and those where CXR was performed more than 5 days prior or 14 days after RT-PCR confirmation. RT-PCR was used to confirm COVID-19 patients. |
| Wang (22)                                                                | CT images were obtained from Tongji Hospital in Wuhan, China (Feb 1 to March 3, 2020). CT images and radiological reports were obtained from three fever clinics (Tiangou Hospital in Wuhan, China, Xianning Central Hospital in Xianning, China and The Second Xiangya Hospital in Changsha, China). | **Inclusion criteria COVID-19 group:** positive diagnosis confirmed by RT-PCR; pairs of CT scans available, patients older than 14 years; **inclusion criteria non-COVID-19 group:** double negative RT-PCR test results, with or without positive CT findings and scanning available, patients older than 14 years. RT-PCR was used to confirm COVID-19 patients. |
| Harrison (16)                                                            | Data from 9 hospitals in Hunan Providence, China (January 6 to April 1, 2020) and the Rhode Island Hospital was used.       | **Inclusion criteria for pneumonia group:** CT scans with report impression containing the word pneumonia; **exclusion criteria:** no chest CT scan, no clear signs of pneumonia or findings suggestive of another diagnosis on chest CT scan; **inclusion criteria for COVID-19 group:** patients with positive diagnosis by RT-PCR and available CT; **exclusion criteria:** no abnormal findings on chest CT scans. RT-PCR was used to confirm COVID-19 patients. |
| Mei (28)                                                                 | A dataset of the presenting patients was acquired from 18 medical centers in 13 provinces in China (January 17 to March 3, 2020). | **Inclusion criteria** COVID-19 exposure, fever, RT-PCR test, chest CT scan. RT-PCR was used to confirm COVID-19 patients. |
| Murphy (20)                                                              | A publicly available CXR dataset (Radiological Society of North America pneumonia detection challenge) was used for retraining. The test set was selected from Jeroen Bosch Hospital in the Netherlands (March 4 and April 6, 2020). For fine-tuning the system, an additional training set was acquired from Bernhoven Hospital in the Netherlands. These were combined with images from | **Inclusion criteria** Individuals suspected of having COVID-19 who presented to the emergency department with respiratory symptoms; all patients underwent laboratory measurements, chest radiographic imaging, and RT-PCR testing. RT-PCR was used to confirm COVID-19 patients. |
| Author | Methodology | Inclusion Criteria | Exclusion Criteria |
|--------|-------------|--------------------|-------------------|
| Wehbe (17) | DeepCOVID-XR was used, which is a weighted ensemble of deep neural networks. | The study sample included consecutive patients from over 20 sites (including hospitals, standalone emergency departments, and urgent care facilities) across the Northwestern Memorial Health Care (NMHC) System in the United States (February to April 2020). | Inclusion criteria: Patients included adults ≥18 years of age with either 1) a documented RT-PCR test result for SARS-CoV-2 (whether positive or negative), 2) a diagnosis of COVID-19 by International Classification of Diseases (ICD-10) code, or 3) a COVID-19 ‘definitive positive’ flag in the electronic health record; COVID-19 positivity was defined as 1) any single positive RT-PCR result during the associated clinical encounter, 2) a diagnosis of COVID-19 by ICD-10 code, or 3) COVID-19 “definitive positive” flag in the electronic health record. |
| Zhou Min (27) | A Deep learning model (the Trinary scheme) was used to discriminate influenza pneumonia lesions from novel coronavirus pneumonia lesions. | Data from 11 hospitals in China was used that met patient selection criteria. This included data from 4 novel coronavirus pneumonia designated hospitals (January 11 to February 23, 2020) and 5 hospitals for influenza pneumonia patients (May 2015 to February, 2020). The external validation set included patient data from 8 hospitals. | Inclusion criteria for COVID-19 pneumonia group: confirmed positive RT-PCR and underwent CT scans within 4 days, exclusion criteria - negative pneumonia findings on CT; inclusion criteria for influenza pneumonia group: confirmed RT-PCR and underwent CT scans within four days, exclusion criteria - negative pneumonia findings on CT. |
| Castiglioni (21) | A Deep learning classifier was used for the diagnosis of COVID-19. An ensemble of ten convolutional neural networks was then used for training and validation. | The training and validation set was composed of CXRs from the Hospital San Gerardo, Monza and the IRCCS Policlinico San Donato in Italy (March 1 to March 13, 2020). | Inclusion criteria: patients with clinical suspicion of COVID-19 (based on the referring physicians judgment for patients admitted at the emergency department, taking into consideration onset of symptoms and blood tests); these patients underwent CXR and RT-PCR testing. |
| Wang Hongmei (23) | A BigBiGAN-based architecture was used to train and extract high-dimensional deep learning features of COVID-19 versus non-COVID-19 pneumonia lesions. | Patient data from the First Affiliated Hospital of University of Science and Technology of China and The Lu’An Affiliated Hospital of Anhui Medical University in China were selected (January 18–February 29, 2020). Patients from Stanford University Hospital (February 1–May 30, 2020) served as external validation. | Inclusion criteria for COVID-19 group: patients obtained RT-PCR tests to determine COVID-19 status; only those patients who tested positive or negative on at least two RT-PCR tests were included; patients obtained chest CT with or without contrast at time of diagnosis; COVID-19 pneumonia with underlying lung diseases (e.g. lung cancer) were included; inclusion criteria for non-COVID-19 pneumonia, as above, patients were included who were clinically suspected to have a viral source of infection; exclusion criteria: Tuberculosis, fungal, or bacterial pneumonia patients. |
| Author (Year) | Model Description | Data Collection | Inclusion Criteria | Case Collection |
|--------------|-------------------|-----------------|-------------------|-----------------|
| Chiu (5)     | COV19NET was used, which is a DL AI model for the detection of COVID-19 on frontal-view CXR. The algorithm used 2 separate pre-trained networks: (1) an Imagenet pre-trained network based on SE-ResNeXt-50-32x4d tuned on the full ChestX-ray14 data set and (2) the same SEResNeXt-50-32x4d network pre-trained on Imagenet. The study data sets consisted of one publicly available CXR data set (ChestX-ray14) and two cohort data sets. The training cohort was selected from Queen Mary Hospital (December 31, 2019 and February 15, 2020) and cases across 4 local hospitals in Hong Kong: cases from Queen Mary Hospital, Ruttonjee, Queen Elizabeth and Pamela Youde Nethersole Eastern Hospitals (through March 2, 2020). | Inclusion criteria: patients who presented to the Hospital warranting RT-PCR testing and COVID-19 RT-PCR confirmed cases across 4 local hospitals | Cases were collected from Guangzhou Eighth People’s Hospital of China (January 6 to April 14, 2020), the Affiliated Hospital of Hebei University of China and the First Affiliated Hospital of Guangzhou Medical University of China. |
| Yang Yanhong (24) | ResUNet was used, which is a DL network using a UNet encoder/decoder backbone, in combination with residual block. | Inclusion criteria for COVID-19 group: patients who tested positive by nucleic acid amplification testing for nasal and pharyngeal swab specimens and were finally confirmed by clinicians and CT imaging; inclusion criteria for non-COVID-19 group: patients with tuberculosis, pneumonia and other pulmonary infections | Nucleic acid amplification testing was used to confirm COVID-19 patients. |

**Legend:** AI: Artificial Intelligence, CT: Computed Tomography, CO-RADS: COVID-19 Reporting and Data System, RT-PCR: Reverse transcription polymerase chain reaction,
| Original study authors’ overall conclusions | Sensitivity (%) Confidence Interval (CI) (95%) | Specificity Confidence Interval (95%) | Area Under the Receiver Operator Characteristic curve | Accuracy | Positive Predictive Value | Negative Predictive Value | Others |
|---------------------------------------------|----------------------------------------------|---------------------------------------|-----------------------------------------------|----------|--------------------------|--------------------------|--------|
| Lessmann (26)                               | Internal Test set: Eight observer mean sensitivity CO-RADS 5: 61.4% ± 7.9% AI sensitivity optimal threshold: 85-7% (CI 73-1-98.2%) External Test set: AI optimal threshold: 82% (CI 69-7-94.3%) | Internal Test Set: Eight observer mean specificity CO-RADS 5: 99-7% (CI 79-6-100%) AI specificity optimal threshold: 89-8% (CI 79-6-100%) External Test Set: AI specificity optimal threshold: 80-5% (CI 67-9-93.1%) | Internal test set AI: 0.95 (CI 0.91-0.98) External test set AI: 0.88 (CI 0.84-0.93) | n.a. | n.a. | n.a. | n.a. |
| Zhang Ran (18)                              | AI using a high-sensitivity operating threshold of 0.4: 88% (CI 87-89%) Three radiologists with no threshold: 42%, 68%, and 90% | AI using a high-sensitivity operating threshold of 0.4: 79% (CI 77-80%) Three radiologists with no threshold: 96%, 85%, and 55% | For the 500 sampled chest radiographs, CV19-Net: 0.94 (CI 0.93-0.96) Radiologists: 0.85 (CI 0.81-0.88) | n.a. | n.a. | n.a. | AI test performance data available according to different age groups and gender |
| Wang (22)                                   | Performance for the identification of changes in lesion burden between two scans: AI: 0.962 (CI 0.947-1.000) Radiologist 1: 0.904 (CI 0.872-0.951) Radiologist 2: 1.000 (CI 1.000-1.000) Radiologist 3: 0.981 (CI 0.974-1.000) | Performance for the identification of changes in lesion burden between two scans: AI: 0.875 (CI 0.833-0.923) Radiologist 1: 0.979 (CI 0.971-1.000) Radiologist 2: 0.875 (CI 0.833-0.923) Radiologist 3: 0.938 (CI 0.917-0.974) | Performance for the identification of changes in lesion burden between two scans: AI: 0.920 (CI 0.900-0.950) Radiologist 1: 0.940 (CI 0.925-0.962) Radiologist 2: 0.940 (CI 0.925-0.962) Radiologist 3: 0.960 (CI 0.950-0.988) | n.a. | n.a. | n.a. | AI-supported triage improved the efficiency of scan-to-fever-clinician triage at each hospital in the study, ranging from 18.77 min for Tianyou Hospital up to 198.28 min for the Second Xiangya Hospital |
| Author     | Methodology                                                                                   | Radiologists’ Performance | AI Model Performance | AI Model: 95% (CI 83-100%, p = 0.001) | Six radiologists’ average: 79% (CI 64-89%, p = 0.001) | Radiologists assisted by the AI model: 88% (CI 74-95%, p < 0.001) | AI model: 96% (CI 88-99%, p = 0.002) | Six radiologists’ average: 88% (CI 78-94%, p = 0.002) | Radiologists assisted by the AI model: 91% (CI 82-96%, p < 0.001) | AI model: 96% (CI 90-98%, p < 0.001) | Six radiologists: 85% (CI: 77-90%, p < 0.001) | Radiologists assisted by the AI model: average of 90% (CI 83-94 %, p < 0.001) |
|------------|-----------------------------------------------------------------------------------------------|-----------------------------|----------------------|--------------------------------------|---------------------------------------------------------------|--------------------------------------------------------------------------------|--------------------------------------|---------------------------------------------------------------|--------------------------------------------------------------------------------|--------------------------------------|---------------------------------------------------------------|--------------------------------------------------------------------------------|
| Harrison   | AI augmentation significantly improved radiologists’ performance in distinguishing COVID-19 from pneumonia of other origin | Higher measures of accuracy, sensitivity, and specificity for AI when compared with a radiologist-only approach | AI model: 95% (CI 83-100%, p = 0.001) | Six radiologists’ average: 79% (CI 64-89%, p = 0.001) | Radiologists assisted by the AI model: 88% (CI 74-95%, p < 0.001) | AI model: 96% (CI 88-99%, p = 0.002) | Six radiologists’ average: 88% (CI 78-94%, p = 0.002) | Radiologists assisted by the AI model: 91% (CI 82-96%, p < 0.001) | AI model: 96% (CI 90-98%, p < 0.001) | Six radiologists: 85% (CI: 77-90%, p < 0.001) | Radiologists assisted by the AI model: average of 90% (CI 83-94 %, p < 0.001) |
| Mei (28)   | The joint model, the CT model and the clinical model performed equally well in sensitivity compared to the senior thoracic radiologist but showed statistically significant improvement in sensitivity compared to the thoracic radiology fellow | Senior thoracic radiologist: 74.6% (CI 66.4-81.7%, p = 0.0501) | Thoracic radiology fellow: 56.0% (CI 47.1-64.5%, p = 1 × 10^-10) | AI joint model: 84.3% (CI 77.1-90.0%) | Senior thoracic radiologist: 93.8% (CI 88.5-97.1%, p = 0.005) | Thoracic radiology fellow: 90.3% (CI 84.3-94.6%, p = 0.090) | AI joint model: 82.8% (CI 75.6-88.5%) | Senior thoracic radiologist: 84.0% (CI 80.0-88.4%) | Thoracic radiology fellow: 73.0% (CI 68.3-77.8%) | AI joint model: 85.1% (CI 79.3-89.5%) | Senior thoracic radiologist: 81.9% (CI 75.6-88.5%) | Senior thoracic radiologist: 84.6% (CI 79.8-88.4%) | Thoracic radiology fellow: 73.8% (CI 68.3-78.9%) | AI joint model: 83.5% (CI 78.6-87.7%) |
| Murphy (20)| Comparable performance of an AI system in the detection of COVID-19 disease on chest radiographs with that of six independent readers | At 85% sensitivity, the AI system outperformed all readers for detection of COVID-19 | AI system: area under the ROC curve of 0.81 | At 85% sensitivity, AI specificity decreased to 61% (CI 48-72%) | AI system: area under the ROC curve of 0.81 | Consensus of all six readers PPV: 72% | | At fixed operating point sensitivity of 75%: | AI system: 77% | Consensus of all six readers: 78% | |
| Study | Methodology | Results | Comparison | Notes |
|-------|-------------|---------|------------|-------|
| Wehbe (17) | Detection of COVID-19 disease on chest radiographs by the AI algorithm DeepCOVID-XR | AI: 71% (CI 63-79%, p = 0.78) | Similar performance to that of experienced thoracic radiologists in consensus | n.a. |
| | Similar performance to that of experienced thoracic radiologists in consensus | Consensus reading: 70% (CI 62-78%, p = 0.78) | | n.a. |
| | AI: 92% (CI 87-90%, p = 0.29) | Consensus reading: 89% (CI 84-94%, p = 0.29) | AI: 82% Consensus reading: 81% | n.a. |
| | AI: 0.88 (CI 0.84-0.92, p = 0.13) | Consensus reading: 0.85 (CI 0.80-0.89, p = 0.13) | | n.a. |
| | AI algorithm: 300 images analysed in 18 minutes | Expert radiologist: 300 images analysed in 2.5-3.5 hours | | n.a. |
| Zhou Min (27) | The trinary scheme achieved specialist-level performance on patient-level classification | Test set: Trinary scheme: 88.9% (CI 86.6-91.2%) | Test set: Trinary scheme: 93.4% (CI 91.3-95.5%) | n.a. |
| | | External validation set: Trinary scheme: 86% (CI 83-89%) | External validation set: Trinary scheme: 77.2% (CI 74.6-80%) | n.a. |
| | Test set: Trinary scheme: 89.2% (CI 87.0-91.5%) | Test set: Trinary scheme: 91.7% | Test set: Trinary scheme: 91.7% | n.a. |
| | External validation set: Trinary scheme: 88% (CI 86-90%) | | | n.a. |
| | | 5 specialists mean: 71.2% (CI 68.8-73.6%) | 5 residents mean: 57.2% (CI 51.3-63.1%) | n.a. |
| | | 5 residents mean: 76.8% (CI 74.8-78.8%) | | n.a. |
| | | Test set: Trinary scheme: 95% | | n.a. |
| | | External validation set: Trinary scheme: 87% | | n.a. |
| | | Radiologists: Specialist group: 80.2% Resident group: 67% | | n.a. |
| Castiglioni (21) | This preliminary experience based on ten CNNs trained on a limited training dataset shows an interesting potential of deep learning for COVID-19 diagnosis | Deep learning model (independent dataset): 0.80 (CI 0.72-0.86, p < 0.005) | Deep learning model (independent dataset): 0.81 (CI 0.73-0.87, p < 0.005) | n.a. |
| | | Radiologist 1 (independent dataset, Centre 1): 0.64 (CI 0.52-0.74) | Radiologist 1 (independent dataset, Centre 1): 0.78 (CI 0.61-0.90) | n.a. |
| | | Radiologist 2 (independent dataset, Centre 2): 0.64 (CI 0.52-0.74) | Radiologist 2 (independent dataset, Centre 2): 0.86 (CI 0.71-0.95) | n.a. |
| | | Deep learning model (independent dataset): 0.81 (CI 0.73-0.87) | Deep learning model (independent dataset): 0.81 (CI 0.73-0.87) | n.a. |
| Name             | Method                                                                 | Results                                                                 | Training dataset                                                                 | External validation dataset                                                                 |
|------------------|------------------------------------------------------------------------|------------------------------------------------------------------------|----------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------|
| Wang Hongmei     | Human expert diagnostic performance improved using a combined deep learning-radiomics model | 3 radiologists’ average on two test datasets: 74·7%  
Radiologists’ average using deep learning and radiomics features: 91·2% |                                                                                   |                                                                                             |
|                  |                                                                        | 3 radiologists’ average on two test datasets: 80·3%  
Radiologists’ average using deep learning and radiomics features: 91·9% | Training dataset:  
Linear classifier: 0·98  
(91·8% sensitivity, 93·4% specificity) vs. 0·91  
(sensitivity 80·0%, specificity 87·2%) |                                                                                             |
|                  |                                                                        |                                                                        |                                                                                   |                                                                                             |
|                  |                                                                        |                                                                        |                                                                                   |                                                                                             |
| Chiu             | AI model outperformed radiologists in a prospectively collected COVID-19 screening validation cohort for the detection of COVID-19, and also in distinguishing COVID-19 on CXR from non-COVID-19 respiratory infectious agents | DL algorithm: 0·85 (p = 0·01)  
RT-PCR as reference standard  
Radiologists: 0·49 (p = 0·01)  
RT-PCR as reference standard | Training dataset:  
Linear classifier: 0·84  
(sensitivity 75·7%, specificity 76·8%) vs. 0·86  
(sensitivity 76·5%, specificity 80·9%) |                                                                                             |
|                  |                                                                        |                                                                        |                                                                                   |                                                                                             |
| Yang Yanhong     | A deep learning algorithm-based AI model improved radiologists’ performance in distinguishing COVID-19 from other pulmonary infections using chest CT images | Radiologists’ average: 0·895 (CI 0·845-0·934)  
Radiologists’ average with AI assistance: 0·942 (CI 0·896-0·968)  
Only AI: 0·918 | Final model:  
AI: 0·903  
Test set:  
Human readers only: 0·963  
AI-supported reading: 0·966 |                                                                                             |
|                  |                                                                        |                                                                        |                                                                                   |                                                                                             |
|                  |                                                                        |                                                                        |                                                                                   |                                                                                             |

Legend: Legend: AI: Artificial Intelligence, CT: Computed Tomography, CXR: Chest X-Ray, CO-RADS: COVID-19 Reporting and Data System, RT-PCR: Reverse transcription polymerase chain reaction, AUC: area under the curve, ROC: receiver operating characteristic, CI: Confidence Interval
Table 3: Overview of datasets used for statistical analysis

| Study            | Large Validation datasets | Large Testing datasets | non-classified |
|------------------|---------------------------|------------------------|----------------|
|                  | Training dataset          | (internal) Validation  |                 |
|                  |                           | dataset                 |                |
|                  |                           | External Validation    |                |
|                  |                           | dataset                 |                |
|                  |                           | (internal) Testing      |                |
|                  |                           | dataset                 |                |
|                  |                           | External Testing        |                |
|                  |                           | dataset                 |                |
|                  |                           |                        |                |
| Lessmann (26)    | No                        | No                     | 104            |
|                  |                           | No                      | 262            |
|                  |                           | No                      |                |
| Zhang Ran (18)   | No                        | No                     | 2,193          |
|                  |                           | No                      | 500            |
|                  |                           | No                      |                |
| Wang (22)        | No                        | No                     | 910            |
|                  |                           | No                      |                |
| Harrison (16)    | 830 & 637 & 396           | No                     | 119 & 275      |
|                  | 237 & 274 & 395           | No                      | 395            |
|                  |                           | No                      |                |
| Mei (28)         | No                        | No                     | 279            |
|                  |                           | No                      |                |
| Murphy (20)      | No                        | No                     | 454            |
|                  |                           | No                      |                |
| Welbe (17)       | 11,796                    | No                     | 300            |
|                  | 2,992                     | No                      | 2,214          |
|                  | 3,931                     | No                      |                |
|                  | 1,100                     | No                      |                |
| Zhou Min (27)    | 275                       | 33                      | 34             |
|                  | 107                       | No                      |                |
| Castiglioni (21) | No                        | 500                     | 110            |
|                  |                           | No                      |                |
| Wang Hongmei (23)| 148                       | 18                      | 18             |
|                  | 32                        | No                      |                |
| Chiu (5)         | No                        | 248 & 182 & 115         | No             |
|                  | 182                      | No                      |                |
| Yang Yanhong (24)| No                        | No                      | 185            |
|                  |                           | No                      |                |
| number of papers involved | 4 | 6 | 3 | 9 | 2 | 1 |

Legend: The sample sizes for patients for the different datasets in this figure may show discrepancies to the ones reported in table 1a) Study characteristics. This is due to original studies subdividing datasets for different purposes in the datasets in the context of sampling and splitting data. In addition, sample sizes may vary for the statistical analysis, where performance measurements are not always consistently provided for all datasets in the original studies. Numbers highlighted in red did not provide related performance measurements.
Table 4: Overview of statistical analysis results (Min/Max, Median, IQR)

|                          | # papers | # scenarios | Minimum | Median | IQR | Maximum |
|--------------------------|----------|-------------|---------|--------|-----|---------|
| Sensitivity for Human readers in Testing | 9        | 58          | 42%     | 71-1%  | 22-3% | 100%    |
| Specificity for Human readers in Testing  | 8        | 54          | 26%     | 85%    | 21-8% | 100%    |
| AUC for Human readers in Testing         | 5        | 13          | 0.73    | 0.85   | 0.07 | 0.963   |
| Sensitivity for AI models in Testing     | 10       | 27          | 60%     | 85%    | 11-3% | 95%     |
| Specificity for AI models in Testing     | 10       | 27          | 61%     | 86-5%  | 10-9% | 96%     |
| AUC for AI models in Testing             | 10       | 27          | 0.8     | 0.92   | 0.09 | 0.97    |
| Sensitivity for AI supported human readings in Testing | 3        | 13          | 81%     | 91-3%  | 8-8%  | 98%     |
| Specificity for AI supported human readings in Testing | 2        | 9           | 78%     | 91-1%  | 6-1%  | 99%     |
| AUC for AI supported human readings in Testing | 1        | 1           | 0.966*  | 0.966* | n.a.  | 0.966*  |
| Sensitivity of Human Readers, AI models and AI supported readings in Testing | 10       | 98          | 42%     | 81-5%  | 19-4% | 100%    |
| Specificity of Human Readers, AI models and AI supported readings in Testing | 10       | 90          | 26%     | 87%    | 15-9% | 100%    |
| AUC of Human Readers, AI models and AI supported readings in Testing | 10       | 41          | 0.73    | 0.9    | 0.11 | 0.97    |
| Sensitivity for Human readers in Validation | 3        | 8           | 49%     | 64-2%  | 46-4% | 100%    |
| Specificity for Human readers in Validation | 3        | 8           | 72%     | 82-2%  | 18-7% | 97-9%   |
| AUC for Human readers in Validation      | 2        | 5           | 0.61    | 0.64   | 0.121 | 0.802   |
| Sensitivity for AI models in Validation  | 4        | 20          | 71-2%   | 79-4%  | 14-8% | 96-3%   |
| Specificity for AI models in Validation  | 4        | 20          | 57-9%   | 83%    | 9-6%  | 89-2%   |
| AUC for AI models in Validation          | 3        | 19          | 0.71    | 0.87   | 0.05  | 0.98    |
| Sensitivity of Human Readers and AI models Validation | 4        | 28          | 49%     | 78-4%  | 18-1% | 100%    |
| Specificity of Human Readers and AI models in Validation | 4        | 28          | 57-9%   | 81%    | 12%   | 97-9%   |
| AUC of Human Readers and AI models in Validation | 3        | 24          | 0.61    | 0.855  | 0.092 | 0.98    |
| Sensitivity of Human Readers, AI models and AI supported readings in Validation | 12       | 126         | 42%     | 80-8%  | 18-5% | 100%    |
| Specificity of Human Readers, AI models and AI supported readings in Testing and Validation | 12       | 188         | 26%     | 85-5%  | 14-5% | 100%    |
| AUC of Human Readers, AI models and AI supported readings in Testing and Validation | 11       | 65          | 0.61    | 0.87   | 0.325 | 0.98    |

Legend: IQR: Interquartile range, *one value only, n.a.: not applicable
**Figure 1a: QUADAS-2 assessment results for included studies**

| Study                | Patient Selection | Index Test | Reference Standard | Flow and Timing | Data Management |
|----------------------|-------------------|------------|--------------------|-----------------|-----------------|
| Lessmann (26)        | Low               | Low        | High               | Low             | Low             |
| Mei (28)             | Low               | Unclear    | Low                | High            | Low             |
| Wang Hongmei (23)    | Low               | Low        | High               | Low             | Low             |
| Castiglioni (21)     | Low               | Low        | Low                | Low             | Low             |
| Welbe (17)           | Low               | Low        | Low                | High            | Low             |
| Harrison (16)        | Low               | Low        | Low                | Low             | Low             |
| Zhang Ran (18)       | High              | Low        | Low                | Low             | Low             |
| Murphy (20)          | Unclear           | Low        | Low                | High            | Low             |
| Chiu (5)             | Low               | Low        | Unclear            | Unclear         | Low             |
| Yang Yanhong (24)    | Low               | Low        | Low                | High            | Low             |
| Wang (22)            | Low               | Unclear    | Low                | Low             | Low             |
| Zhou Min (27)        | Low               | Low        | High               | Low             | Unclear         |

**Legend:** QUADAS 2 assessment results for included studies, indicating low (green colour), high (orange colour) or unclear (blue colour) risk of bias for the relevant domains.
Figure 1b: QUADAS-2 assessment results for excluded studies

| Study       | Patient Selection | Index Test | Reference Standard | Flow and Timing | Data Management |
|-------------|-------------------|------------|--------------------|-----------------|-----------------|
| Dorr (25)   | Unclear           | Unclear    | High               | High            | High            |
| Javor (11)  | High              | Unclear    | High               | Low             | Unclear         |
| Ni (10)     | High              | Low        | Low                | High            | Low             |
| Wang Zheng (19) | High      | High       | High               | High            | Low             |
| Yang Shuyi (13) | High    | Unclear    | High               | High            | Low             |
| Xie (12)    | High              | Low        | High               | Low             | Low             |
| Krishnamoorthy (15) | High     | High       | Low                | High            | Low             |
| Sukhija (14) | High              | High       | Low                | High            | Low             |

Legend: QUADAS 2 assessment results for excluded studies, indicating low (green colour), high (orange colour) or unclear (blue colour) risk of bias for the relevant domains.
Figure 2: Study selection

Records identified from Databases (n = 1261)
- Embase: 371
- PubMed: 447
- Scopus: 443

Records excluded (n = 1236)
Duplicates: 497
Publication date before 2019: 129
Review, summary, survey, opinion or not full article: 199
Not COVID-19 or imaging focused: 159
No comparison with human readers: 197
No AI methodology: 4
Focus on segmentation: 13
Focus on life survival/disease progression: 33
Focus on treatment: 1
Not in English: 4

Records screened (n = 1270)

Full text articles assessed for eligibility (n = 34)

Records excluded (n = 22)
Incomplete comparison/no relevant results (i.e. sensitivity, specificity) available (n = 14)
Evaluated as high risk of bias studies with QUADAS-2 quality assessment (n=8)

Studies included in systematic review (n = 12)
Figure 3: Overall results visualization

| Original study authors' overall conclusions | Sensitivity | Specificity | AUC |
|-------------------------------------------|-------------|-------------|-----|
| Lessmann (26)                             |             |             | n.a.|
| Zhang Ran (18)                            |             |             | n.a.|
| Wang (22)                                 |             |             | n.a.|
| Harrison (16)                             |             |             | n.a.|
| Mei (28)                                  |             |             | n.a.|
| Murphy (20)                               |             |             | n.a.|
| Wöhbe (17)                               |             |             | n.a.|
| Zhou Min (27)                             |             |             | n.a.|
| Castiglioni (21)                          |             |             | n.a.|
| Wang Hongmei (23)                         |             |             | n.a.|
| Chiu (5)                                  |             |             | n.a.|
| Yang Yanhong (24)                         |             |             | n.a.|

**Legend:** Red colour indicates “Better performance of AI”, blue colour indicates “Better performance of human readers”, orange colour indicates “Increased performance of human readers with AI support”, grey colour indicates “Comparable performance of AI and human readers to differentiate COVID-19”, green colour indicates “Comparable performance of AI and human readers for lesion changes”, no colour and “n.a.” indicates “Values in comparison not indicated”
Figure 4: Forest plots with performance overviews

a) Testing dataset performance human readers: Sensitivity

Legend: Brackets behind scenario names indicate sample sizes, squares with green outlines indicate sensitivity, with 95% CI ranges where reported.
b) Testing dataset performance human readers: Specificity

Legend: Brackets behind scenario names indicate sample sizes, squares with blue outlines indicate specificity, with 95% CI ranges where reported.
c) Testing dataset performance human readers: AUC

Legend: Brackets behind scenario names indicate sample sizes, squares with pink outlines indicate AUC values, with 95% CI ranges where reported.
d) Testing dataset performance AI models: Sensitivity

Legend: Brackets behind scenario names indicate sample sizes, triangles with green outlines indicate sensitivity values, with 95% CI ranges where reported.
e) Testing dataset performance AI models: Specificity

Legend: Brackets behind scenario names indicate sample sizes, triangles with blue outlines indicate specificity values, with 95% CI ranges where reported.
f) Testing dataset performance AI models: AUC

Legend: Brackets behind scenario names indicate sample sizes, triangles with pink outlines indicate AUC values, with 95% CI ranges.

AI model (104)
AI model (262)
CV19Net model high sens threshold (2,193)
DNN final model (119)
DNN model trained on equal split of datasets (95)
DNN model trained on third split of datasets (275)
CNN model with CT only (279)
Joint model with clinical data only (279)
MLP model with clinical data only (279)
AI model with CT only (279)
AI model cutoff 0 (654)
AI model cutoff 1 (654)
AI model cutoff 2 (654)
DeepCOVID-XR with RT-PCR reference standard (117)
DeepCOVID-XR with Radiologist's consensus reference standard (117)
DL Plain scheme (34)
DL ternary scheme (34)
AI model (110)
DL features linear classifier (16)
DL features linear classifier on whole lung (16)
Radiomics features linear classifier (16)
DL & Radiomics features linear classifier (16)
DL features lasso classifier (16)
DL features lasso classifier on whole lung (16)
Radiomics features lasso classifier (16)
DL & Radiomics features lasso classifier (16)
DL model (160)

0.70 0.75 0.80 0.85 0.90 0.95 1.00 1.05 1.10

Lessmen (26)
Zhang Ran (18)
Harrison (16)
Mei (28)
Murphy (20)
Wehbe (17)
Zhou Min (27)
Castiglioni (21)
Wang Hongmei (23)
Yang Yanhong (24)
g) Testing dataset performance AI supported human readings: Sensitivity, Specificity and AUC

Legend: Brackets behind scenario names indicate sample sizes, circles with green outlines indicate sensitivity, circles with blue outlines indicate specificity and circles with pink outlines indicate AUC value, with 95% CI ranges where reported
h) Validation datasets performance human readers: Sensitivity, Specificity and AUC

Legend: Brackets behind scenario names indicate sample sizes, squares with green outlines indicate sensitivity, squares with blue outlines indicate specificity and squares with pink outlines indicate AUC value, with 95% CI ranges where reported.
i) Validation datasets performance AI models: Sensitivity, Specificity and AUC

Legend: Brackets behind scenario names indicate sample sizes, triangles with green outlines indicate sensitivity, triangles with blue outlines indicate specificity and triangles with pink outlines indicate AUC value, with 95% CI ranges where reported.
**Figure 5: Box plots with performance overview comparisons**

**a) Comparison of testing datasets: Sensitivity**

Legend: Boxes show the lower quartile, median and upper quartile. Whiskers adjacent to the boxes represent the minimum and maximum values. Numbers on the right indicate sample sizes, with numbers in brackets indicating the references to the relevant studies.
b) Comparison of testing datasets: Specificity

Legend: Boxes show the lower quartile, median and upper quartile. Whiskers adjacent to the boxes represent the minimum and maximum values. Numbers on the right indicate sample sizes, with numbers in brackets indicating the references to the relevant studies.
c) Comparison of testing datasets: AUC

Legend: Boxes show the lower quartile, median and upper quartile. Whiskers adjacent to the boxes represent the minimum and maximum values. Numbers on the right indicate sample sizes, with numbers in brackets indicating the references to the relevant studies.
d) Comparison of validation datasets: Sensitivity

Legend: Boxes show the lower quartile, median and upper quartile. Whiskers adjacent to the boxes represent the minimum and maximum values. Numbers on the right indicate sample sizes, with numbers in brackets indicating the references to the relevant studies.
e) Comparison of validation datasets: Specificity

**Legend:** Boxes show the lower quartile, median and upper quartile. Whiskers adjacent to the boxes represent the minimum and maximum values. Numbers on the right indicate sample sizes, with numbers in brackets indicating the references to the relevant studies.
f) Comparison of validation datasets: AUC

Legend: Boxes show the lower quartile, median and upper quartile. Whiskers adjacent to the boxes represent the minimum and maximum values. Numbers on the right indicate sample sizes, with numbers in brackets indicating the references to the relevant studies.
g) Comparison between testing and validation datasets: Sensitivity

Legend: Boxes show the lower quartile, median and upper quartile. Whiskers adjacent to the boxes represent the minimum and maximum values. Numbers on the right indicate sample sizes, with numbers in brackets indicating the references to the relevant studies.
h) Comparison between testing and validation datasets: Specificity

Legend: Boxes show the lower quartile, median and upper quartile. Whiskers adjacent to the boxes represent the minimum and maximum values. Numbers on the right indicate sample sizes, with numbers in brackets indicating the references to the relevant studies.
i) Comparison between testing and validation datasets: AUC

Legend: Boxes show the lower quartile, median and upper quartile. Whiskers adjacent to the boxes represent the minimum and maximum values. Numbers on the right indicate sample sizes, with numbers in brackets indicating the references to the relevant studies.
j) Comparison between testing and external validation datasets, only comparable AI models: Sensitivity, Specificity and AUC

Legend: Boxes show the lower quartile, median and upper quartile. Whiskers adjacent to the boxes represent the minimum and maximum values. Numbers on the right indicate sample sizes, with numbers in brackets indicating the references to the relevant studies.
Highlights

- There may be an added value of AI model supported imaging-based COVID-19 detection
- Studies reported comparable or better performance of AI or AI-supported readings
- There was lower variability of diagnostic performance for AI than for human readers
- Our systematic review shows heterogeneity of data characteristics and risks of bias
- There is a variety of applied methodologies and statistical analysis limitations