Fast Track

Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal

Laure Wynants,1,2 Ben Van Calster,2,3 Gary S Collins,4,5 Richard D Riley,6 Georg Heinze,7 Ewoud Schuit,8,9 Marc M J Bonten,8,10 Darren L Dahly,11,12 Johanna A Damen,8,9 Thomas P A Debray,8,9 Valentijn M T De Jong,8,9 Maarten De Vos,2,13 Paula Dhiman,4,5 Maria C Haller,7,14 Michael O Harhay,15,16 Liesbet Henckaerts,17,18 Pauline Heus,8,9 Michael Kammer,7,19 Nina Kreuzberger,20 Anna Lohmann,21 Kim Luijken,21 Jie Ma,5,6 Gen P Martin,22 David J McLernon,23 Constanza L Andaur Navarro,8,9 Johannes B Reitsma,8,9 Jamie C Sergeant,24,25 Chunhu Shi,26 Nicole Skoetz,19 Luc J M Smits,1 Kym L E Snell,6 Matthew Sperrin,27 René Spijkervt,8,9,28 Ewout W Steyerberg,3 Toshihiko Takada,8 Ioanna Tzoulaki,29,30 Sander M J van Kuijk,31 Bas C T van Bussel,1,32 Iwan C C van der Horst,32 Florien S van Royen,8,9 Jan Y Verbakel,33,34 Christine Wallisch,7,35,36 Jack Wilkinson,22 Robert Wolff,37 Lotty Hooff,8,9 Karel G M Moons,8,9 Maarten van Smeden8

ABSTRACT

OBJECTIVE
To review and appraise the validity and usefulness of published and preprint reports of prediction models for diagnosing coronavirus disease 2019 (covid-19) in patients with suspected infection, for prognosis of patients with covid-19, and for detecting people in the general population at increased risk of covid-19 infection or being admitted to hospital with the disease.

DESIGN
Living systematic review and critical appraisal by the COVID-PRECISE (Precise Risk Estimation to optimise covid-19 Care for Infected or Suspected patients in diverse settings) group.

WHAT IS ALREADY KNOWN ON THIS TOPIC

The sharp recent increase in coronavirus disease 2019 (covid-19) incidence has put a strain on healthcare systems worldwide; an urgent need exists for efficient early detection of covid-19 in the general population, for diagnosis of covid-19 in patients with suspected disease, and for prognosis of covid-19 in patients with confirmed disease.

Viral nucleic acid testing and chest computed tomography imaging are standard methods for diagnosing covid-19, but are time consuming.

Earlier reports suggest that elderly patients, patients with comorbidities (chronic obstructive pulmonary disease, cardiovascular disease, hypertension), and patients presenting with dyspnoea are vulnerable to more severe morbidity and mortality after infection.

WHAT THIS STUDY ADDS

Seven models identified patients at risk in the general population (using proxy outcomes for covid-19).

Thirty three diagnostic models were identified for detecting covid-19, in addition to 75 diagnostic models based on medical images, 10 diagnostic models for severity classification, and 107 prognostic models for predicting, among others, mortality risk, progression to severe disease.

Proposed models are poorly reported and at high risk of bias, raising concern that their predictions could be unreliable when applied in daily practice.

Two prediction models (one for diagnosis and one for prognosis) were identified as being of higher quality than others and efforts should be made to validate these in other datasets.

RESULTS

PubMed and Embase through Ovid, up to 1 July 2020, supplemented with arXiv, medRxiv, and bioRxiv up to 5 April 2020.

STUDY SELECTION

Studies that developed or validated a multivariable covid-19 related prediction model.

DATA EXTRAVCTIONS

At least two authors independently extracted data using the CHARMS (critical appraisal and data extraction for systematic reviews of prediction modelling studies) checklist; risk of bias was assessed using PROBAST (prediction model risk of bias assessment tool).

DATA SOURCES

PubMed and Embase through Ovid, up to 1 July 2020, supplemented with arXiv, medRxiv, and bioRxiv up to 5 April 2020.

STUDY SELECTION

Studies that developed or validated a multivariable covid-19 related prediction model.

DATA EXTRAVCTIONS

At least two authors independently extracted data using the CHARMS (critical appraisal and data extraction for systematic reviews of prediction modelling studies) checklist; risk of bias was assessed using PROBAST (prediction model risk of bias assessment tool).

RESULTS

37421 titles were screened, and 169 studies describing 232 prediction models were included. The review identified seven models for identifying people at risk in the general population; 118 diagnostic models for detecting covid-19 (75 were based on medical imaging, 10 to diagnose disease severity); and 107 prognostic models for predicting mortality risk, progression to severe disease, intensive care unit admission, ventilation, intubation, or length of hospital stay. The most frequent types of predictors included in the covid-19 prediction models are vital signs, age, comorbidities, and image features. Flu-like symptoms are frequently predictive in diagnostic models, while sex, C reactive protein, and lymphocyte counts are frequent prognostic factors. Reported C index estimates from the strongest form of validation available per model ranged from 0.71 to 0.99 in prediction models for the general population, from 0.65 to more than 0.99 in diagnostic models, and from 0.54 to 0.99 in prognostic models. All models were rated at high or unclear risk of bias, mostly because of non-representative selection of control patients, exclusion of patients who had not experienced the event of interest by the end of the study, high risk of model overfitting, and unclear reporting. Many models did not include a description of the target population (n=27, 12%) or care setting (n=75, 32%), and only 11 (5%) were externally
validated by a calibration plot. The Jehi diagnostic model and the 4C mortality score were identified as promising models.

**CONCLUSION**

Prediction models for covid-19 are quickly entering the academic literature to support medical decision making at a time when they are urgently needed. This review indicates that almost all published prediction models are poorly reported, and at high risk of bias such that their reported predictive performance is probably optimistic. However, we have identified two (one diagnostic and one prognostic) promising models that should soon be validated in multiple cohorts, preferably through collaborative efforts and data sharing to also allow an investigation of the stability and heterogeneity in their performance across populations and settings. Details on all reviewed models are publicly available at https://www.covprecise.org/. Methodological guidance as provided in this paper should be followed because unreliable predictions could cause more harm than benefit in guiding clinical decisions. Finally, prediction model authors should adhere to the TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) reporting guideline.

**SYSTEMATIC REVIEW REGISTRATION**


**READERS’ NOTE**

This article is a living systematic review that will be updated to reflect emerging evidence. Updates may occur for up to two years from the date of original publication. This version is update 3 of the original article published on 7 April 2020 (BMJ 2020;369:m1328). Previous updates can be found as data supplements (https://www.bmj.com/content/369/bmj.m1328/related#datasupp). When citing this paper please consider adding the update number and date of access for clarity.

**Introduction**

The novel coronavirus disease 2019 (covid-19) presents an important and urgent threat to global health. Since the outbreak in early December 2019 in the Hubei province of the People’s Republic of China, the number of patients confirmed to have the disease has exceeded 47 million as the disease spread globally, and the number of people infected is probably much higher. More than 1.2 million people have died from covid-19 (up to 3 November 2020). Despite public health responses aimed at containing the disease and delaying the spread, several countries have been confronted with a critical care crisis, and more countries could follow. Outbreaks lead to important increases in the demand for hospital beds and shortage of medical equipment, while medical staff themselves can also become infected. Several regions have had or are experiencing second waves, and despite improvements in testing and tracing, several regions are again facing the limits of their test capacity, hospital resources and healthcare staff.

To mitigate the burden on the healthcare system, while also providing the best possible care for patients, efficient diagnosis and information on the prognosis of the disease are needed. Prediction models that combine several variables or features to estimate the risk of people being infected or experiencing a poor outcome from the infection could assist medical staff in triaging patients when allocating limited healthcare resources. Models ranging from rule based scoring systems to advanced machine learning models (deep learning) have been proposed and published in response to a call to share relevant covid-19 research findings rapidly and openly to inform the public health response and help save lives.

We aimed to systematically review and critically appraise all currently available prediction models for covid-19, in particular models to predict the risk of covid-19 infection or being admitted to hospital with the disease, models to predict the presence of covid-19 in patients with suspected infection, and models to predict the prognosis or course of infection in patients with covid-19. We included model development and external validation studies. This living systematic review, with periodic updates, is being conducted by the international COVID-PRECISE (Precise Risk Estimation to optimise covid-19 Care for Infected or Suspected patients in diverse settings; https://www.covprecise.org/) group in collaboration with the Cochrane Prognosis Methods Group.

**Methods**

We searched the publicly available, continuously updated publication list of the covid-19 living systematic review. We validated whether the list is fit for purpose (online supplementary material) and further supplemented it with studies on covid-19 retrieved from arXiv. The online supplementary material presents the search strings. We included studies if they developed or validated a multivariable model or scoring system, based on individual participant level data, to predict any covid-19 related outcome. These models included three types of prediction models: diagnostic models to predict the presence or severity of covid-19 in patients with suspected infection; prognostic models to predict the course of infection in patients with covid-19; and prediction models to identify people in the general population at risk of covid-19 infection or at risk of being admitted to hospital with the disease.

We searched the database repeatedly up to 1 July 2020 (supplementary table 1). As of the third update (search date 1 July), we only include peer reviewed articles (indexed in PubMed and Embase through Ovid). Preprints (from bioRxiv, medRxiv, and arXiv) that were already included in previous updates of the systematic review remain included in the analysis. Reassessment takes place after publication of a preprint in a peer reviewed journal. No restrictions were made on the setting (eg, inpatients, outpatients, or general population), prediction horizon (how far ahead the model predicts), included predictors,
or outcomes. Epidemiological studies that aimed to model disease transmission or fatality rates, diagnostic test accuracy, and predictor finding studies were excluded. We focus on studies published in English. Starting with the second update, retrieved records were initially screened by a text analysis tool developed using artificial intelligence to prioritise sensitivity (supplementary material). Titles, abstracts, and full texts were screened for eligibility in duplicate by independent reviewers (pairs from LW, BVC, MvS) using EPPI-Reviewer, and discrepancies were resolved through discussion.

Data extraction of included articles was done by two independent reviewers (from LW, BVC, GSC, TPAD, MCH, GH, KMM, RDR, ES, IJMS, EWS, KIES, CW, AL, JM, TT, JAAD, KL, JBR, LHH, CS, MS, MCH, NS, NK, SMvK, JCS, PD, CLAN, RW, GPM, IT, JYV, DLD, JW, FSvR, PH, VMtDj, BCTvB, ICCvdH, DJM, MK, and MvS). Reviewers used a standardised data extraction form based on the CHARMS (critical appraisal and data extraction for systematic reviews of prediction modelling studies) checklist and PROBAST (prediction model risk of bias assessment tool; www.probast.org) for assessing the reported prediction models. We sought to extract each model’s predictive performance by using whatever measures were presented. These measures included any summaries of discrimination (the extent to which predicted risks discriminate between participants with and without the outcome), and calibration (the extent to which predicted risks correspond to observed risks) as recommended in the TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis; www.tripod-statement.org) statement.

Discrimination is often quantified by the C index (C index = 1 if the model discriminates perfectly; C index > 0.5 if discrimination is no better than chance). Calibration is often quantified by the calibration intercept (which is zero when the risks are not systematically overestimated or underestimated) and calibration slope (which is one if the predicted risks are not too extreme or too moderate). We focused on performance statistics as estimated from the strongest available form of validation (in order of strength: external (evaluation in an independent database), internal (bootstrap validation, cross validation, random training test splits, temporal splits), apparent (evaluation by using exactly the same data used for development)). Any discrepancies in data extraction were discussed between reviewers, and remaining conflicts were resolved by LW or MvS.

The online supplementary material provides details on data extraction. Some studies investigated multiple models and some models were investigated in multiple studies (that is, in external validation studies). The unit of analysis was a model within a study, unless stated otherwise. We considered aspects of PRISMA (preferred reporting items for systematic reviews and meta-analyses) and TRIPOD in reporting our study. Details on all reviewed studies and prediction models are publicly available at https://www.covprecise.org/.

**Patient and public involvement**

It was not possible to involve patients or the public in the design, conduct, or reporting of our research. A lay summary of the project’s aims is available on https://www.covprecise.org/project/. The study protocol and preliminary results are publicly available on https://osf.io/ehc67/, medRxiv and https://www.covprecise.org/living-review/.

**Results**

We retrieved 37 412 titles through our systematic search (of which 23 203 were included in the present update; supplementary table 1, fig 1). We included a further nine studies that were publicly available but were not detected by our search. Of 37 421 titles, 444 studies were retained for abstract and full text screening (of which 169 were included in the present update). One hundred sixty nine studies describing 232 prediction models met the inclusion criteria (of which 62 studies and 87 models added since the present update, supplementary table 1). These studies were selected for data extraction and critical appraisal. The unit of analysis was the model within a study: of these 232 models, 208 were unique, newly developed models for covid-19. The remaining 24 analyses were external validations of existing models (in a study other than the model development study). Some models were validated more than once (in different studies, as described below). Many models are publicly available (box 1). A database with the description of each model and its risk of bias assessment can be found on https://www.covprecise.org/.

**Primary datasets**

One hundred seventy four (75%) models used data from a single country (table 1), 42 (18%) models used international data, and for 16 (7%) models it was unclear how many (and which) countries contributed data. Two (1%) models used simulated data and 12 (5%) used proxy data to estimate covid-19 related risks (eg, Medicare claims data from 2015 to 2016). Most models were intended for use in confirmed covid-19 cases (47%) and a hospital setting (51%). The average patient age ranged from 39 to 71 years, and the proportion of men ranged from 35% to 75%, although this information was often not reported. One study developed a prediction model for use in paediatric patients.

Based on the studies that reported study dates, data were collected from December 2019 to June 2020. Some centres provided data to multiple studies and several studies used open Github or Kaggle data repositories (version or date of access often unspecified), and so it was unclear how much these datasets overlapped across our identified studies.

Among the diagnostic model studies, the reported prevalence of covid-19 varied between 7% and 71% (if a cross sectional or cohort design was used). Because 75 diagnostic studies used either case-control sampling or an unclear method of data collection, the
prevalence in these diagnostic studies might not be representative of their target population.

Among the studies that developed prognostic models to predict mortality risk in people with confirmed or suspected infection, the percentage of deaths ranged from 1% to 52%. This wide variation is partly because of substantial sampling bias caused by studies excluding participants who still had the disease at the end of the study period (that is, they had neither recovered nor died). Additionally, length of follow-up varied between studies (but was often not reported), and there is likely to be local and temporal variation in how people were diagnosed as having covid-19 or were admitted to the hospital (and therefore recruited for the studies).

We identified seven models that predicted risk of covid-19 in the general population. Three models from one study used hospital admission for non-tuberculosis pneumonia, influenza, acute bronchitis, or upper respiratory tract infections as proxy outcomes in a dataset without any patients with covid-19.16 Among the predictors were age, sex, previous hospital admission, comorbidities, and social determinants of health. The study reported C indices of 0.73, 0.81, and 0.81. A fourth model used deep learning on thermal...
Diagnostic models to detect covid-19 in patients with suspected infection

We identified 33 multivariable models to distinguish between patients with and without covid-19. Most models targeted patients with suspected covid-19. Reported C index values ranged between 0.65 and 0.99. Calibration was assessed for seven models using calibration plots (including two at external validation), with mixed results. The most frequently included predictors (≥10 times) were vital signs (eg, temperature, heart rate, respiratory rate, oxygen saturation, blood pressure), flu-like signs and symptoms (eg, shiver, fatigue), age, electrolytes, image features (eg, pneumonia signs on CT scan), contact with individuals with confirmed covid-19, lymphocyte count, neutrophil count, cough or sputum, sex, leukocytes, liver enzymes, and red cell distribution width.

Ten studies aimed to diagnose severe disease in patients with covid-19: nine in adults with reported C indices between value of 0.80 and 0.99, and one in children that reported perfect classification of severe disease. Calibration was not assessed in any of the models. Predictors of severe covid-19 used more than once were comorbidities, liver enzymes, C reactive protein, imaging features, lymphocyte count, and neutrophil count.

Seventy five prediction models were proposed to support the diagnosis of covid-19 or covid-19 pneumonia (and some also to monitor progression) based on images. Most studies used CT images or chest radiographs. Others used spectrograms of cough sounds and lung ultrasound. The predictive performance varied considerably, with reported C index values ranging from 0.70 to more than 0.99. Only one model based on imaging was evaluated by use of a calibration plot, and it appeared to be well calibrated at external validation.

Prognostic models for patients with diagnosis of covid-19

We identified 107 prognostic models for patients with a diagnosis of covid-19. The intended use of these models (that is, when to use them, and for whom) was often not clearly described. Prediction horizons varied between one and 37 days, but were often unspecified.

Of these models, 39 estimated mortality risk and 28 aimed to predict progression to a severe or critical disease. The remaining studies used other outcomes (single or as part of a composite) including recovery, length of hospital stay, intensive care unit admission, intubation, (duration of) mechanical ventilation, acute respiratory distress syndrome, cardiac injury and thrombotic complication. One study used data from 2015 to 2019 to predict mortality and prolonged assisted mechanical ventilation (as a non-covid-19 proxy outcome). The most frequently used categories of prognostic factors (for any outcome, included at least 20 times) included age, comorbidities, vital signs, image features, sex, lymphocyte count, and C reactive protein.

---

Table 1 | Characteristics of reviewed prediction models for diagnosis and prognosis of coronavirus disease 2019 (covid-19)

<table>
<thead>
<tr>
<th>Country†</th>
<th>No (%) of models* or median (interquartile range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single country data</td>
<td>174 (75)</td>
</tr>
<tr>
<td>China</td>
<td>97 (43)</td>
</tr>
<tr>
<td>Italy</td>
<td>23 (10)</td>
</tr>
<tr>
<td>United States</td>
<td>17 (7)</td>
</tr>
<tr>
<td>South Korea</td>
<td>10 (4)</td>
</tr>
<tr>
<td>France</td>
<td>5 (2)</td>
</tr>
<tr>
<td>Singapore</td>
<td>4 (2)</td>
</tr>
<tr>
<td>Turkey</td>
<td>4 (2)</td>
</tr>
<tr>
<td>Brazil</td>
<td>3 (1)</td>
</tr>
<tr>
<td>Spain</td>
<td>2 (1)</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>2 (1)</td>
</tr>
<tr>
<td>Other single country</td>
<td>8 (3)</td>
</tr>
<tr>
<td>International (combined) data</td>
<td>42 (18)</td>
</tr>
<tr>
<td>Unknown origin of data</td>
<td>16 (7)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of data used</th>
<th>No (%) of models* or median (interquartile range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proxy (non-covid-19) data</td>
<td>12 (5)</td>
</tr>
<tr>
<td>Simulated data</td>
<td>2 (1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Target setting</th>
<th>No (%) of models* or median (interquartile range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients admitted to hospital</td>
<td>119 (51)</td>
</tr>
<tr>
<td>Patient at triage centre or fever clinic</td>
<td>12 (5)</td>
</tr>
<tr>
<td>Patients in general practice</td>
<td>3 (1)</td>
</tr>
<tr>
<td>Other</td>
<td>23 (10)</td>
</tr>
<tr>
<td>Unclear</td>
<td>25 (12)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Target population</th>
<th>No (%) of models* or median (interquartile range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confirmed covid-19</td>
<td>108 (47)</td>
</tr>
<tr>
<td>Suspected covid-19</td>
<td>84 (36)</td>
</tr>
<tr>
<td>Other</td>
<td>13 (6)</td>
</tr>
<tr>
<td>Unclear</td>
<td>27 (12)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of model</th>
<th>No (%) of models* or median (interquartile range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnostic (covid-19 v not covid-19)</td>
<td>33 (14)</td>
</tr>
<tr>
<td>Diagnostic classification of covid-19 severity</td>
<td>10 (4)</td>
</tr>
<tr>
<td>Diagnostic, imaging data only</td>
<td>75 (32)</td>
</tr>
<tr>
<td>Prognostic</td>
<td>107 (46)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study type</th>
<th>No (%) of models* or median (interquartile range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developed in reviewed study</td>
<td>50 (22)</td>
</tr>
<tr>
<td>Developed and internally validated in reviewed study</td>
<td>112 (48)</td>
</tr>
<tr>
<td>Developed and externally validated in reviewed study</td>
<td>46 (20)</td>
</tr>
<tr>
<td>Externally validated in reviewed study</td>
<td>24 (10)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample size</th>
<th>No (%) of models* or median (interquartile range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size (development)</td>
<td>338 (134-707)</td>
</tr>
<tr>
<td>No of events (development)</td>
<td>69 (37-160)</td>
</tr>
<tr>
<td>Sample size (external validation)</td>
<td>189 (76-312)</td>
</tr>
<tr>
<td>No of events (external validation)</td>
<td>40 (24-122)</td>
</tr>
</tbody>
</table>

*Analysis unit is a model within a study. Some studies investigated multiple models and some models were investigated in multiple studies (that is, in external validation studies).

†A study that uses development data from one country and validation data from another is classified as international.
Studies that predicted mortality reported C indices between 0.68 and 0.98. Four studies also presented calibration plots (including at external validation for three models), all indicating miscalibration or showing plots for integer scores without clearly explaining how these were translated into predicted risks. The studies that developed models to predict progression to a severe or critical disease reported C indices between 0.58 and 0.99. Five of these models also were evaluated by calibration plots, two of them at external validation. Even though calibration appeared good, plots were constructed in an unclear way. Reported C indices for other outcomes varied between 0.54 (admission to intensive care) and 0.99 (severe symptoms three days after admission), and five models had calibration plots (of which three at external validation), with mixed results.

Risk of bias

All models were at high (n=226, 97%) or unclear (n=6, 3%) risk of bias according to assessment with PROBAST, which suggests that their predictive performance when used in practice is probably lower than that reported (fig 2). Therefore, we have cause for concern that the predictions of the proposed models are unreliable when used in other people. Figure 2 and box 2 gives details on common causes for risk of bias for each type of model.

Fig 2 | PROBAST (prediction model risk of bias assessment tool) risk of bias for all included models combined (n=232) and broken down per type of model

<table>
<thead>
<tr>
<th>Risk of bias</th>
<th>Percentage of models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0%</td>
</tr>
<tr>
<td>Unclear</td>
<td>25%</td>
</tr>
<tr>
<td>High</td>
<td>75%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>General population (n=7)</th>
<th>Diagnosis (n=33)</th>
<th>Diagnosis - imaging (n=75)</th>
<th>Diagnosis - severity (n=10)</th>
<th>Prognosis (n=107)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>Overall</td>
<td>Overall</td>
<td>Overall</td>
<td>Overall</td>
</tr>
<tr>
<td>Participants</td>
<td>Participants</td>
<td>Participants</td>
<td>Participants</td>
<td>Participants</td>
</tr>
<tr>
<td>Predictors</td>
<td>Predictors</td>
<td>Predictors</td>
<td>Predictors</td>
<td>Predictors</td>
</tr>
<tr>
<td>Outcome</td>
<td>Outcome</td>
<td>Outcome</td>
<td>Outcome</td>
<td>Outcome</td>
</tr>
<tr>
<td>Analysis</td>
<td>Analysis</td>
<td>Analysis</td>
<td>Analysis</td>
<td>Analysis</td>
</tr>
</tbody>
</table>
adequate data collection and study design, but issues that could have been avoided by conducting a better statistical analysis. Many studies had small to modest sample sizes (table 1), which led to an increased risk of overfitting, particularly if complex modelling strategies were used. In addition, 50 models (22%) were neither internally nor externally validated. Performance statistics calculated on the development data from these models are likely optimistic. Calibration was only assessed for 22 models using calibration plots (10%), of which 11 on external validation data.

We found two models that were generally of good quality, built on large datasets, and had been rated low risk of bias on most domains but with an overall rating of unclear risk of bias, owing to unclear details on one signalling question within the analysis domain (table 2 provides a summary). Jehi and colleagues presented findings from developing a diagnostic model, however, there was substantial missing data and it remains unclear whether the use of median imputation influenced results, and there are unexplained discrepancies between the online calculator, nomogram, and published logistic regression model.\(^\text{141}\) Hence, the calculator should not be used without further validation. Knight and colleagues developed a prognostic model for in-hospital mortality, however, continuous predictors were dichotomised, which reduces granularity of predicted risks (even though the model had a C index comparable with that of a generalised additive model).\(^\text{145}\) The model was also converted into a sum score, but it was unclear how the scores were translated to the predicted mortality risks that were used to evaluate calibration.

**External validation**

Forty-six models were developed and externally validated in the same study (in an independent dataset, excluding random training test splits and temporal splits). In addition, 24 external validations of models were developed for covid-19 or before the covid-19 pandemic in separate studies. However, none of the external validations was scored as low risk of bias, three were rated as unclear risk of bias, and 67 were rated as high risk of bias. One common concern is that datasets used for the external validation were likely not representative of the target population (eg, patients not being recruited consecutively, use of an inappropriate study design, use of unrepresentative controls, exclusion of patients still in follow-up). Consequently, predictive performance could differ if the models are applied in the targeted population. Moreover, only 15 (21%) external validations had

---

**Box 2: Common causes of risk of bias in the reported prediction models**

**Models to predict coronavirus disease 2019 (covid-19) risk in general population**

All of these models had unclear or high risk of bias for the participant, outcome, and analysis domain. All were based on proxy outcomes to predict covid-19 related risks, such as presence of or hospital admission due to severe respiratory disease, in the absence of data of patients with covid-19.\(^\text{156 159}\)

**Diagnostic models**

Ten models (30%) used inappropriate data sources (eg, due to a non-nested case-control design), nine (27%) used inappropriate inclusion or exclusion criteria such that the study data was not representative of the target population, and eight (24%) selected controls that were not representative of the target population for a diagnostic model (eg, controls for a screening model had viral pneumonia). Other frequent problems were dichotomisation of predictors (nine models, 27%), and tests used to determine the outcome (eight models, 24%) or predictor definitions or measurement procedures (seven models, 21%) that varied between participants.

**Diagnostic models based for severity classification**

Two models (20%) used predictor data that was assessed while the severity (the outcome) was known. Other concerns include non-standard or lack of a prespecified outcome definition (two models, 20%), predictor measurements (eg, fever) being part of the outcome definition (two models, 20%) and outcomes being assessed with knowledge of predictor measurements (two models, 20%).

**Diagnostic models based on medical imaging**

Generally, studies did not clearly report which patients had imaging during clinical routine. Fifty-five (73%) used an inappropriate or unclear study design to collect data (eg, a non-nested case-control). It was often unclear (39 models, 52%) whether the selection of controls was made from the target population (that is, patients with suspected covid-19). Outcome definitions were often not defined or determined in the same way in all participants (18 models, 24%). Diagnostic model studies that used medical images as predictors were all scored as unclear on the predictor domain. These publications often lacked clear information on the preprocessing steps (eg, cropping of images). Moreover, complex machine learning algorithms transform images into predictors in a complex way, which makes it challenging to fully apply the PROBAST predictors section for such imaging studies. However, a more favourable assessment of the predictor domain does not lead to better overall judgment regarding risk of bias for the included models. Careful description of model specification and subsequent estimation were frequently lacking, challenging the transparency and reproducibility of the models. Studies used different deep learning architectures, some were established and others specifically designed, without benchmarking the used architecture against others.

**Prognostic models**

Dichotomisation of predictors was a frequent concern (22 models, 21%). Other problems include inappropriate inclusions or exclusions of study participants (18 models, 17%). Study participants were often excluded because they did not develop the outcome at the end of the study period but were still in follow-up (that is, they were in hospital but had not recovered or died), yielding a selected study sample (12 models, 11%). Additionally, many models (16 models, 15%) did not account for censoring or competing risks.
For participant selection, predictor description, and statistical methods used. Models were developed on data from different countries, but the majority used data from a single country. Often, the available sample sizes and number of events for the outcomes of interest were limited. This problem is well known when building prediction models and increases the risk of overfitting the model. A high risk of bias implies that the performance of these models in new samples will probably be worse than that reported by the researchers. Therefore, the estimated C indices, often close to 1 and indicating near perfect discrimination, are probably optimistic. The majority of studies developed new models specifically for COVID-19, but only 46 carried out an external validation, and calibration was rarely assessed. We cannot yet recommend any of the identified prediction models for widespread use in clinical practice, although a few diagnostic and prognostic models originated from studies that were clearly of better quality. We suggest that these models should be further validated in other data sets, and ideally by independent investigators.

### Challenges and opportunities

The main aim of prediction models is to support medical decision making in individual patients. Therefore, it is vital to identify a target setting in which predictions serve a clinical need (eg, emergency department, intensive care unit, general practice, symptom monitoring app in the general population), and a representative dataset from that setting (preferably comprising consecutive patients) on which the prediction model can be developed and validated. This clinical setting and patient characteristics should be described in detail (including timing within the disease course, the severity of disease at the moment of prediction, and the comorbidity), so that readers and clinicians are able to understand if the proposed model could be suited for their population. Unfortunately, the studies included in our systematic review often lacked an adequate description of the target setting and study population, which leaves users of these models in doubt about the models’ applicability. Although we recognise that the earlier studies were done under severe time constraints, we recommend that any studies currently in preprint and all future studies

### Table 2 | Prediction models with unclear risk of bias overall and large development samples

<table>
<thead>
<tr>
<th>Study setting and outcome</th>
<th>Diagnostic models</th>
<th>Prognostic models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jeji et al141†; data from US, patients with suspected COVID-19; COVID-19 diagnosis</td>
<td>Jehi model Development 11 672 (818), external validation 2295 (290)</td>
<td>Knight et al143; data from UK, suspected or confirmed symptomatic inpatients; in-hospital mortality 4C Mortality Score Development 35 463 (11 426), temporal validation 22 361 (6729)</td>
</tr>
<tr>
<td>Predictive performance</td>
<td>Overall risk of bias using PROBAST</td>
<td></td>
</tr>
<tr>
<td>Strongest type of validation reported</td>
<td>Performance†</td>
<td></td>
</tr>
<tr>
<td>Jehi et al141†; data from US, patients with suspected COVID-19; COVID-19 diagnosis</td>
<td>Jehi model Development 11 672 (818), external validation 2295 (290)</td>
<td>External validation, same country, new centres, and later period</td>
</tr>
<tr>
<td>C index 0.84 (95% CI 0.82 to 0.86)</td>
<td>Unclear</td>
<td></td>
</tr>
<tr>
<td>Knight et al143; data from UK, suspected or confirmed symptomatic inpatients; in-hospital mortality</td>
<td>4C Mortality Score Development 35 463 (11 426), temporal validation 22 361 (6729)</td>
<td>Temporal validation</td>
</tr>
<tr>
<td>C index 0.77 (95% CI 0.76 to 0.77)</td>
<td>Unclear</td>
<td></td>
</tr>
</tbody>
</table>

PROBAST=prediction model risk of bias assessment tool; COVID-19=coronavirus disease 2019

†According to PROBAST, a large dataset is at least 10 events per candidate variable (EPV) for model development and at least 100 events for validation. If EPV could not be extracted or calculated from the study report, 100 events for model development was the lower limit to be included in this table.

†Performance from strongest type of validation reported.
Table 3 | External validations with unclear risk of bias and large validation samples

<table>
<thead>
<tr>
<th>Study setting and outcome</th>
<th>Model</th>
<th>Sample size (total no of participants for model validation set (no with outcome))</th>
<th>Predictive performance</th>
<th>Overall risk of bias using PROBAST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnostic models</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jehi et al11; data from US, patients with suspected covid-19; covid-19 diagnosis</td>
<td>Jehi model</td>
<td>Development 11 672 (981); external validation 2295 (290)</td>
<td>External validation, same country, new centres and later period</td>
<td>C index 0.84 (95% CI 0.82 to 0.86)</td>
</tr>
<tr>
<td>Prognostic models</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Luo et al15; data from China, in-patients with confirmed covid-19; in-hospital mortality</td>
<td>CURB-65</td>
<td>1018 (201)</td>
<td>Independent external validation</td>
<td>C index 0.84 (95% CI 0.82 to 0.93)</td>
</tr>
</tbody>
</table>

PROBAST = prediction model risk of bias assessment tool; CURB-65 = confusion, urina, respiratory rate, blood pressure plus age of at least 65 years.

According to PROBAST, a large dataset is at least 10 events per candidate variable for model development and at least 100 events for validation.

should adhere to the TRIPOD reporting guideline12 to improve the description of their study population and guide their modelling choices. TRIPOD translations (eg, in Chinese and Japanese) are also available at https://www.tripod-statement.org.

A better description of the study population could also help us understand the observed variability in the reported outcomes across studies, such as covid-19 related mortality and covid-19 prevalence. The variability in mortality could be related to differences in included patients (eg, age, comorbidities) and interventions for covid-19. The variability in prevalence could in part be reflective of different diagnostic standards across studies.

Covid-19 prediction will often not present as a simple binary classification task. Complexities in the data should be handled appropriately. For example, a prediction horizon should be specified for prognostic outcomes (eg, 30 day mortality). If study participants have neither recovered nor died within that time period, their data should not be excluded from analysis, which some reviewed studies have done. Instead, an appropriate time to event analysis should be considered to allow for administrative censoring.13 Censoring for other reasons, for instance because of quick recovery and loss to follow-up of patients who are no longer at risk of death from covid-19, could necessitate analysis in a competing risk framework.191

We reviewed 75 studies that used only medical images to diagnose covid-19, covid-19 related pneumonia, or to assist in segmentation of lung images, the majority using advanced machine learning methodology. The predictive performance measures showed a high to almost perfect ability to identify covid-19, although these models and their evaluations also had a high risk of bias, notably because of poor reporting and an artificial mix of patients with and without covid-19. Currently, none of these models is recommended to be used in clinical practice. An independent systematic review and critical appraisal (using PROBAST12) of machine learning models for covid-19 using chest radiographs and CT scans came to the same conclusions, even though they focused on models that met a minimum requirement of study quality based on specialised quality metrics for the assessment of radiomics and deep-learning based diagnostic models in radiology.192

A prediction model applied in a new healthcare setting or country often produces predictions that are miscalibrated193 and might need to be updated before it can safely be applied in that new setting.13 This requires data from patients with covid-19 to be available from that system. Instead of developing and updating predictions in their local setting, individual participant data from multiple countries and healthcare systems might allow better understanding of the generalisability and implementation of prediction models across different settings and populations. This approach could greatly improve the applicability and robustness of prediction models in routine care.194-198

The evidence base for the development and validation of prediction models related to covid-19 will continue to increase over the coming months. To leverage the full potential of these evolutions, international and interdisciplinary collaboration in terms of data acquisition, model building and validation is crucial.

Study limitations
With new publications on covid-19 related prediction models rapidly entering the medical literature, this systematic review cannot be viewed as an up-to-date list of all currently available covid-19 related prediction models. Also, 80 of the studies we reviewed were only available as preprints. These studies might improve after peer review, when they enter the official medical literature; we will reassess these peer reviewed publications in future updates. We also found other prediction models that are currently being used in clinical practice without scientific publications,199 and web risk calculators launched for use while the scientific manuscript is still under review (and unavailable on request).200 These unpublished models naturally fall outside the scope of this review of the literature. As we have argued extensively elsewhere,201 transparent reporting that enables validation by independent researchers is key for predictive analytics, and clinical guidelines should only recommend publicly available and verifiable algorithms.

Implications for practice
All reviewed prediction models were found to have an unclear or high risk of bias, and evidence from independent external validations of the newly
developed models is still scarce. However, the urgency of diagnostic and prognostic models to assist in quick and efficient triage of patients in the covid-19 pandemic might encourage clinicians and policymakers to prematurely implement prediction models without sufficient documentation and validation. Earlier studies have shown that models were of limited use in the context of a pandemic, and they could even cause more harm than good. Therefore, we cannot recommend any model for use in practice at this point.

The current oversupply of insufficiently validated models is not useful for clinical practice. Moreover, predictive performance estimates obtained from different populations, settings, and types of validation (internal vs external) are not directly comparable. Future studies should focus on validating, comparing, improving, and updating promising available prediction models. The models by Knight and colleagues and Jehi and colleagues are good candidates for validation studies in other data. We advise Jehi and colleagues to make all model equations available for independent validation. Such external validations should assess not only discrimination, but also calibration and clinical utility (net benefit) in large datasets collected using an appropriate study design. In addition, these models’ transportability to other countries or settings remains to be investigated. Owing to differences between healthcare systems (eg, Chinese and European) and over time in when patients are admitted to and discharged from hospital, as well as the testing criteria for patients with suspected covid-19, we anticipate most existing models will be miscalibrated, but researchers could attempt to update and adjust the model to the local setting.

Most reviewed models used data from a hospital setting, but few are available for primary care and the general population. Additional research is needed, including validation of any recently proposed models not yet included in the current update of the living review (eg, Clift et al). The models reviewed to date predicted the covid-19 diagnosis or assess the risk of mortality or deterioration, whereas long term morbidity and functional outcomes remain understudied and could be a target outcome of interest in future studies developing prediction models.

When creating a new prediction model, we recommend building on previous literature and expert opinion to select predictors, rather than selecting predictors in a purely data driven way. This is especially important for datasets with limited sample size. Frequently used predictors included in multiple models identified by our review are vital signs, age, comorbidities, and image features, and these should be considered when appropriate. Flu-like symptoms should be considered in diagnostic models, and sex, C reactive protein, and lymphocyte counts could be considered as prognostic factors.

By pointing to the most important methodological challenges and issues in design and reporting of the currently available models, we hope to have provided a useful starting point for further studies, which should preferably validate and update existing ones. This living systematic review has been conducted in collaboration with the Cochrane Prognosis Methods Group. We will update this review and appraisal continuously to provide up-to-date information for healthcare decision makers and professionals as more international research emerges over time.

Conclusion
Several diagnostic and prognostic models for covid-19 are currently available and they all report moderate to excellent discrimination. However, these models are all at high or unclear risk of bias, mainly because of model overfitting, inappropriate model evaluation (eg, calibration ignored), use of inappropriate data sources and unclear reporting. Therefore, their performance estimates are probably optimistic and not representative for the target population. The COVID-PRECISE group does not recommend any of the current prediction models to be used in practice, but one diagnostic and one prognostic model originated from higher quality studies and should be (independently) validated in other datasets. For details of the reviewed models, see https://www.covprecise.org/. Future studies aimed at developing and validating diagnostic or prognostic models for covid-19 should explicitly describe the concerns raised and follow existing methodological guidance for prediction modeling studies, because unreliable predictions could cause more harm than benefit in guiding clinical decisions. Prediction model authors should adhere to the TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) reporting guideline. Finally, sharing data and expertise for the validation and updating of covid-19 related prediction models is urgently needed.

AUTHOR AFFILIATIONS
1 Department of Epidemiology, CAPHRI Care and Public Health Research Institute, Maastricht University, Peter Debyeplein 1, 6229 HA Maastricht, Netherlands
2 Department of Development and Regeneration, KU Leuven, Leuven, Belgium
3 Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, Netherlands
4 Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Musculoskeletal Sciences, University of Oxford, Oxford, UK
5 NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, UK
6 Centre for Prognosis Research, School of Primary, Community and Social Care, Keele University, Keele, UK
7 Section for Clinical Biometrics, Centre for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, Austria
8 Julius Center for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht University, Utrecht, Netherlands
9 Cochrane Netherlands, University Medical Centre Utrecht, Utrecht University, Utrecht, Netherlands
10 Department of Medical Microbiology, University Medical Centre Utrecht, Utrecht, Netherlands
11 HRB Clinical Research Facility, Cork, Ireland
12 School of Public Health, University College Cork, Cork, Ireland
submit for publication. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: LW, BVC, LH, and MDV acknowledge specific funding for this work from Internal Funds KU Leuven, KOOR, and the COVID-19 Fund. LW is a postdoctoral fellow of Research Foundation-Flanders (FWO) and receives support from ZonMw (grant 10430012010001). BVC received support from FWO (grant G084716N) and Internal Funds KU Leuven (grant C24/15/037). TAPD acknowledges financial support from the Netherlands Organisation for Health Research and Development (grant 91617050). VMD was supported by the European Union and MvS wrote the first draft of the manuscript. The funders played no role in study design, data collection, data analysis, data interpretation, or reporting.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: support from Internal Funds KU Leuven, KOOR, and the COVID-19 Fund for the submitted work; no competing interests with regards to the submitted work; LW discloses support from Research Foundation-Flanders; BVC reports personal fees as a statistics editor for The BMJ (since 2009), consultancy fees for Roche for giving meta-analysis teaching and advice in October 2018, and personal fees for delivering in-house training courses at Barts and the London School of Medicine and Dentistry, and the Universities of Aberdeen, Exeter, and Leeds, all outside the submitted work, MS coauthored the editorial on the original article.

Ethical approval: Not required.

Data sharing: The study protocol is available online at https://osf.io/ehc47/. Detailed extracted data on all included studies are available on https://www.covprecise.org/.

The lead authors affirm that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

Dissemination to participants and related patient and public communities: The study protocol is available online at https://osf.io/ehc47/.

Provenance and peer review: Not commissioned; externally peer reviewed.

This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt and build upon this work for non-commercial use, provided the original work is properly cited. See: http://creativecommons.org/licenses/by-nc/4.0/.


Imran A, Posokhova I, Qureshi HN, et al. AI4COVID-19: AI enabled preliminary diagnosis for covid-19 from cough samples via an


Web appendix: Supplementary material