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 Schuij,8,9 Elena Albu,2 Banafsheh Arshi,1 Vanesa Bellou,10 Marc M J Bonten,11
Darren L Dahly,12,13 Johanna A Damen,8,9 Thomas P A Debray,8,14 Valentin M T de Jong,8,9
Maarten De Vos,8,15 Paula Dhiman,4,5 Joie Ensor,8 Shan Gao,8 Maria C Haller,7,16
Mohammed Hudda,21 Kevin Jenniskens,8,9 Michael Kammer,22 Nina Kreuzberger,23
Anna Lohmann,24 Brooke Levis,6 Kim Luijken,25 Jie Ma,5 Glen P Martin,25 David J McLernon,26
Constanza L Andaur Navarro,8,9 Johannes B Reitsma,8,9 Jamie C Sergeant,27,28 Chunhu Shi,29
Nicole Skoetz,22 Luc J M Smits,1 Kym I E Snell,6 Matthew Sperrin,30 René Spijker,8,9,31
Ekwot W Steyerberg,3 Yoshihiko Takada,8,32 Ioanna Tzoulaki,10,33 Sander M J van Kuijk,34
Bas C T van Bussel,1,35 Iwan C C van der Horst,35 Kelly Reeve,36 Florien S van Royen,37
Jan Y Verbakel,37,38 Christine Wallisch,7,39,40 Jack Wilkinson,24 Robert Wolff,41 Lotty Hooft,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 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 or dying 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.

DATA SOURCES PubMed and Embase through Ovid, up to 17 February 2021, supplemented with arXiv, medRxiv, and bioRxiv up to 5 May 2020.

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

WHAT IS ALREADY KNOWN ON THIS TOPIC

Recurrent peaks in covid-19 incidence have put a strain on healthcare systems worldwide; a need exists for efficient early risk stratification in the general population, and for prognosis of covid-19 in patients with confirmed disease. Viral nucleic acid testing, chest computed tomography imaging, and antigen tests are standard methods for diagnosing covid-19, and their availability has made covid-19 diagnostic models less relevant.

WHAT THIS STUDY ADDS

Of models with a low risk of bias, four identify patients at risk in the general population; one assists in patient triage at the emergency department; and three estimate prognosis in patients admitted to hospital with covid-19.

Calibration of these models is likely to vary over time and across settings.

There is an oversupply of models and external validations at high risk of bias, raising concern that predictions could be unreliable when these models are applied in daily practice.

DATA EXTRACTION

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

126 978 titles were screened, and 412 studies describing 731 new prediction models or validations were included. Of these 731,125 were diagnostic models (including 75 based on medical imaging) and the remaining 606 were prognostic models for either identifying those at risk of covid-19 in the general population (13 models) or predicting diverse outcomes in those individuals with confirmed covid-19 (593 models). Owing to the widespread availability of diagnostic testing capacity after the summer of 2020, this living review has now focused on the prognostic models. Of these, 29 had low risk of bias, 32 had unclear risk of bias, and 545 had high risk of bias. The most common causes for high risk of bias were inadequate sample sizes (n=408, 67%) and inappropriate or incomplete evaluation of model performance (n=338, 56%). 381 models were newly developed, and 225 were external validations of existing models. The reported C indexes varied between 0.77 and 0.93 in development studies of existing models. The reported C indexes varied between 0.77 and 0.93 in development studies with low risk of bias, and between 0.56 and 0.78 in external validations with low risk of bias. The Qcovid models, the PRIEST score, Carr’s model, the ISARIC4C Deterioration model, and the Xie model showed adequate predictive performance in studies at low risk of bias. Details on all reviewed models are publicly available at https://www.covprecise.org/.

CONCLUSION

Prediction models for covid-19 entered the academic literature to support medical decision making at unprecedented speed and in large numbers. Most published prediction model studies were poorly reported and at high risk of bias such that their reported predictive performances are probably

BMJ: first published as 10.1136/bmj.m1328 on 7 April 2020. Downloaded from http://www.bmj.com/ on 1 December 2022 by guest. Protected by copyright.
optimistic. Models with low risk of bias should be validated before clinical implementation, preferably through collaborative efforts to also allow an investigation of the heterogeneity in their performance across various populations and settings. 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 modellers 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 the final version of a living systematic review that has been updated over the past two years to reflect emerging evidence. This version is update 4 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 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 500 million as the disease spread globally, and the number of people infected is probably much higher. More than 6 million people have died from covid-19 (up to 24 May 2022).\(^1\) Despite public health responses aimed at containing the disease and delaying the spread, countries have been confronted with repeated surges disrupting health services\(^2-6\) and society at large. More recent outbreaks of the omicron variant led to important increases in the demand for test capacity, hospital beds, and medical equipment, while medical staff members also increasingly became infected themselves.\(^5\) While many national governments have now put an end to covid-19 restrictions, scientists warn that endemic circulation of SARS-CoV-2, perhaps with seasonal epidemic peaks, is likely to have a continued important disease burden.\(^7\) In addition, virus mutations can be unpredictable, and lack of effective surveillance or adequate response could enable the emergence of new epidemic or pandemic covid-19 patterns.\(^7\)\(^8\) To mitigate the burden on the healthcare system, while also providing the best possible care for patients, reliable prognosis of covid-19 remains important to inform decisions regarding shielding, vaccination, treatment, and hospital or intensive care unit (ICU) admission. 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.

The outbreak of covid-19 was accompanied by a surge of scientific evidence.\(^7\) The speed with which evidence about covid-19 has accumulated is unprecedented. To provide an overview of available prediction models, a living systematic review, with periodic updates, was 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. Initially, the review included diagnostic and prognostic models. Owing to the current availability of testing for covid-19 infections, we restricted the focus to prognostic models in this new update. Hence our aim was to systematically review and critically appraise available prognostic models for detecting people in the general population at increased risk of covid-19 infection or being admitted to hospital or dying with the disease, and models to predict the prognosis or course of infection in patients with a covid-19 diagnosis. We included all prognostic model development and external validation studies.

**Methods**
We searched the publicly available, continuously updated publication list of the covid-19 living systematic review.\(^10\) 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 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 or dying with the disease. Diagnostic models to predict the presence or severity of covid-19 in patients with suspected infection were included up to update 3 only, which can be found in the data supplements.

We searched the database repeatedly up to 17 February 2021 (supplementary table 1). As of the third update (search date 1 July 2020), 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 remained included in the analysis. Reassessment took place after publication of a preprint in a peer reviewed journal and replaced the original assessment. 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, and predictor finding studies, were excluded. We only included 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, and KGMM) 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, KGMM, RDR, ES, LJMS, KIES, CW, AL, JM, TT, JAD, KL, JBR, LH, CS, MS, MCH, NS, NK, SMvK, JC, PD, CLAN, RW, GPM, JT, JYV, DLD, JW, FSwR, PH, VMTdJ, BCTvB, ICCvdH, DJM, MK, BL, EA, SG, BA, JH, KJ, SG, KR, JE, MH, VB, 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; https://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; https://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 assessed graphically using calibration plots or 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).
Three hundred and eighty one unique prognostic models were developed in the included studies. Eighty (21%) of these models were presented as a model equation including intercept and regression coefficients. Thirty nine (10%) models were only partially presented (eg, intercept or baseline hazard were missing). The remaining 262 (69%) did not provide the underlying model equation.

One hundred and sixty one (42%) were available in a tool to facilitate use in clinical practice (in addition to or instead of a published equation). Sixty models (16%) were presented as a nomogram, 35 (9%) as a web calculator, 30 (8%) as a sum score, nine (2%) as a software object, five (1%) as a decision tree or set of predictions for subgroups, and 22 (6%) in other usable formats (6%).

All these presentation formats make predictions readily available for use in the clinic. However, because most of these prognostic models were at high or uncertain risk of bias, we do not recommend their routine use before they are externally validated, ideally by independent investigators in other data than used for their development.

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
Severe covid-19 survivors and lay people participated by discussing their perspectives, providing advice, and acting as partners in writing a lay summary of the project’s aims and results (available at https://www.covprecise.org/project/), thereby taking part in the implementation of knowledge distribution. Owing to the initial emergency situation of the covid-19 pandemic, we did not involve patients or the public in the design and conduct of this living review in March 2020, but the study protocol and preliminary results were immediately publicly available on https://osf.io/ehc47/, medRxiv, and https://www.covprecise.org/living-review/.

Results
We identified 126 969 records through our systematic search, of which 89 566 were identified in the present search update (supplementary table 1). Of these, 310 studies describing 606 prognostic models or validations of prognostic models are included in the current analysis: 13 prognostic models for developing covid-19 in the general population and 593 prognostic models for predicting outcomes in patients with covid-19 diagnoses. The results from previous updates, including diagnostic models, are available as supplementary material. A database with the description of each model and validation and its risk of bias assessment can be found on https://www.covprecise.org/.

Of the 606 prognostic models, 381 were unique, newly developed models for covid-19, and 225 were external validations of existing models in a study other than the model development study. These external validations include external validations of newly developed covid models, as well as prognostic scores predating the covid-19 pandemic. Some models were validated more than once (in different studies, as described below). One hundred and fifty eight (41%) newly developed models were publicly available in a format for use in practice (box 1).

Primary datasets
Five hundred and fifty six (92%) developed or validated models used data from a single country (table 1), 39 (6%) used international data, and for 11 (2%) models it was unclear how many (and which) countries contributed data. Thirty (0.5%) models used simulated data and 21 (3%) 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 (83%) and a hospital setting (82%). The average patient age ranged from 38 to 84 years, and the proportion of men ranged from 1% to 95%, although this information was often not reported.

Based on the studies that reported study dates, data were collected from December 2019 to October 2020. Some centres provided data to multiple studies and it was unclear how much these datasets overlapped across identified studies.

The median sample size for model development was 414, with a median number of 74 individuals experiencing the event that was predicted. The mortality risk in patients admitted to hospital ranged from 8% to 46%. This wide variation is partly due to differences in length of follow-up between studies (which was often not reported), local and temporal variation in diagnostic criteria, admission criteria and treatment, as well as selection bias (eg, excluding participants who had neither recovered nor died at the end of the study period).

Models to predict covid-19 related risks in the general population
We identified 13 newly developed models aiming to predict covid-19 related risks in the general population. Five models predicted hospital admission for covid-19, three predicted mortality, one predicted development of severe covid-19, and four predicted an insufficiently defined covid-19 outcome. Eight of these 13 general
Table 1 | Characteristics of reviewed prediction models for 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>556 (92)</td>
</tr>
<tr>
<td>China</td>
<td>200 (33)</td>
</tr>
<tr>
<td>US</td>
<td>80 (13)</td>
</tr>
<tr>
<td>Italy</td>
<td>59 (10)</td>
</tr>
<tr>
<td>UK</td>
<td>53 (9)</td>
</tr>
<tr>
<td>Spain</td>
<td>40 (7)</td>
</tr>
<tr>
<td>South Korea</td>
<td>28 (5)</td>
</tr>
<tr>
<td>Mexico</td>
<td>21 (3)</td>
</tr>
<tr>
<td>Turkey</td>
<td>15 (2)</td>
</tr>
<tr>
<td>Norway</td>
<td>10 (2)</td>
</tr>
<tr>
<td>Other single country</td>
<td>50 (8)</td>
</tr>
<tr>
<td>International (combined) data</td>
<td>39 (6)</td>
</tr>
<tr>
<td>Unknown origin of data</td>
<td>11 (2)</td>
</tr>
</tbody>
</table>

Type of data used

- Pandemic data: 582 (96)
- Proxy (non-covid-19) data: 21 (3)
- Simulated data: 3 (0.5)

Target setting

- Patients admitted to hospital: 496 (82)
- Patient at triage centre or fever clinic: 6 (1)
- Patients in general practice: 6 (1)
- Other: 54 (9)
- Unclear: 47 (8)

Target population

- Confirmed covid-19: 502 (83)
- Suspected covid-19: 46 (8)
- Other: 25 (4)
- Unclear: 33 (5)

Type of model

- Prognostic model to predict future risk of covid-19 in general population: 13 (2)
- Prognostic models for outcomes in patients with covid-19: 593 (98)

Analysis done in reviewed study

- Development only: 96 (16)
- Development and internal validation: 185 (31)
- Development and external validation: 100 (17)
- External validation only: 225 (37)

Sample size

- Sample size (development): 414 (172-1505)
- No of events (development): 74 (36-207)
- Sample size (external validation): 314 (127-516)
- No of events (external validation): 62 (26-115)

*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 investigated in multiple studies (that is, in external validation studies).

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).

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).

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).

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).

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).

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).

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).

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).

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).

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).

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).

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).

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).

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).

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).

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).

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).

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).

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

The studies reported C indexes between 0.49 and 1, with a median of 0.81 (interquartile range 0.75-0.89). The median C index was 0.83 for the mortality models, 0.83 for progression models, and 0.77 for ICU admission models. Researchers showed calibration plots for only 152 of the 593 models (26%, of which 102 at external validation). The calibration results were mixed, with several studies indicating inaccurate risk predictions (examples in Xie et al, Barda et al, and Zhang et al). Plots were sometimes constructed in an unclear way, hampering interpretation (examples in Guo et al, Gong et al, and Knight et al).

Risk of bias

Seven newly developed prognostic models and 22 external validations of prognostic models were at low risk of bias (n=29, 5%). Most newly developed models and external validations were at unclear (n=32, 5%) or high (n=54.5%, 90%) risk of bias according to assessment with PROBAST, which suggests that the predictive performance when used in practice is probably lower than what is reported (fig 2). Figure 2 and box 2 give details on common causes for risk of bias.

Three hundred and eighty four (63%) of the 606 models and validations had a low risk of bias for the participants domain. One hundred and seven models (32%) had an unclear risk of bias for the participants domain, which indicates that the participants enrolled in the studies might not be representative of the models’ targeted populations. Unclear reporting on the inclusion of participants led to an unclear risk of bias assessment in 115 models (19%). Three hundred and eighty six models (64%) had a low risk of bias for the predictor domain, while 193 (32%) had an unclear risk of bias and 27 had a high risk of bias (4%). High risk of bias for the predictor domain indicates that predictors were not available at the models’ intended time of use, not clearly defined, or influenced by the outcome measurement. Most studies used outcomes that are easy to assess (eg, all cause death), and hence 353 (58%) were rated at low risk of bias.
bias for the outcome domain. Nonetheless, there was cause for concern about bias induced by the outcome measurement in 106 models (17%), for example, due to the use of proxy outcomes (eg, hospital admission for non-covid-19 severe respiratory infections). One hundred and forty seven (24%) had an unclear risk of bias due to opaque or ambiguous reporting. In contrast to the participant, predictor, and outcome domains, the analysis domain was problematic for most of the 606 models and validations. Overall, 530 (87%) were at high risk of bias for the analysis domain, and the reporting was insufficiently clear to assess risk of bias in the analysis in 42 (7%). Only 34 (6%) were at low risk of bias for the analysis domain.

**Newly developed models at low risk of bias**

We found seven newly developed models at low risk of bias (table 2). All had good to excellent discrimination, but calibration varied, highlighting the need of local and temporal recalibration.

The four Qcovid models predict hospital admission and death with covid-19 in the general population in the UK, separately for men and women.231 The models use age, ethnic group, deprivation, body mass index, and a range of comorbidities as predictors. The models showed underestimated risks for high risk patients at external validation, which was remedied by recalibrating the model.231

The PRIEST score predicts 30 day death or organ support in patients with suspected or confirmed covid-19 presenting at the emergency department. The triage score is based on NEWS2 (national early warning score 2 consisting of respiratory rate, oxygen saturation, heart rate, systolic blood pressure, temperature, consciousness, air, or supplemental oxygen), age, sex, and performance status (ranging from bed-bound to normal performance). Its external validation in UK emergency departments showed reasonable calibration, but potential heterogeneity in calibration across centres was not examined.
The analysis domain was the most problematic domain: 87% (n=530) of newly developed models and validations were at high risk of bias, compared to 18% (n=107), 4% (n=27), and 17% (n=106) for the participant, predictor, and outcome domains. One hundred and fifty one (25%) models had low risk of bias on all domains except analysis, indicating adequate data collection and study design, but issues that could have been avoided by conducting a better statistical analysis. The most frequent problem was insufficient sample size (n=408, 67%). Small to modest sample sizes and numbers of events (table 1) led to an increased risk of overfitting, particularly if complex modelling strategies were used. Not properly accounting for overfitting or optimism was also common (n=250, 41%). Ninety six models (16%) were neither internally nor externally validated. If done, internal validation was sometimes not correctly executed (ie, not all modelling steps were repeated). Performance statistics from these models are likely optimistic. Moreover, evaluation of discrimination and calibration was often incomplete, or done with inappropriate statistics (n=338, 56%). Calibration was only assessed for 156 models using calibration plots (26%), of which 106 (17%) on external validation data. Inappropriate handling of missing data was common (n=290, 48%). One hundred and twenty seven conducted a complete case analysis (21%), 205 (34%) did not mention how missing data was handled.

### Models to predict covid-19 risk in general population versus prognostic models in patients with covid-19

The 593 prognostic models for patients with covid-19 were more often at high risk of bias than the 13 general population models (90% (n=536) v 69% (n=9)). This difference was mainly due to the analysis domain (88% (n=521) v 69% (n=9) at high risk of bias). The median sample size for model development in patients with covid-19 was 397 (71 events), compared to 1.6 million (1867 events) for general population models. The median sample size for external validation was 299 (42 events), compared to >1 million (1303 events) for general population models. Hence, more models had an inadequate sample size for the chosen analysis strategy (69% (n=407) v 8% (n=1)), and more were at risk of overfitting and optimism (42% (n=248 v 15% (n=2)). The outcome domain was more problematic for the general population models than for the models for patients with covid-19, with 62% (n=8) versus 17% (n=98) at high risk of bias in this domain. This difference was caused using proxy outcomes (n=8, 62%)—for example, hospital admission due to severe respiratory disease other than covid-19. For the participant and predictor domains, the risk of bias was comparable (fig 2).

### Development and external validation

External validations were more often at low risk of bias than newly developed models (10%, (n=22/225) v 2% (n=7/381)). The statistical analysis domain was the most problematic domain for model development as well as for external validation studies, with 93% (n=353) and 79% (n=177) at high risk of bias for this domain, respectively. The most common causes of high risk of bias were the same for both types (small sample size, inappropriate evaluation of predictive performance, and inappropriate handling of missing data), except for overfitting and optimism, which is not a concern at external validation.

Carr’s model268 and the ISARIC 4C deterioration model268 predict deterioration in covid-19 patients admitted to hospital. The composite outcomes for both models included ICU admission and death, while the ISARIC 4C model also adds ventilatory support. Both models had comparable performance but included different predictors, all typically available at admission. Carr and colleagues supplemented NEWS2 with age, laboratory and physiological parameters (supplemental oxygen flow rate, urea, oxygen saturation, C reactive protein, estimated glomerular filtration rate, neutrophil count, neutrophil-lymphocyte ratio). Gupta and colleagues268 developed a model including age, sex, nosocomial infection, Glasgow coma scale score, peripheral oxygen saturation at admission, breathing room air or oxygen therapy, respiratory rate, urea concentration, C reactive protein concentration, lymphocyte count, and presence of radiographic chest infiltrates. Carr’s model was validated internationally261 and the ISARIC4C Deterioration model was validated regionally within the UK.268 For both models, calibration varied across settings.

### External validations at low risk of bias

We identified 225 external validations in dedicated (ie, not combined with the development of the model) external validation studies. Only 22 were low risk of bias, although all 22 validations came from the same study using single-centre UK data (table 3).269 This validation study included 411 patients, of which 180 experienced a deterioration in health, and 115 died. In this study, the Carr model and NEWS2 performed best to predict deterioration, while the Xie model and REMS performed best to predict mortality. Both the Carr model (a preprint version that differs slightly from the Carr model reported above) and the Xie model showed slight miscalibration.

NEWS2 and REMS were also validated in other dedicated validation studies. NEWS2 obtained C indexes between 0.65 and 0.90,141 203 214 233 245 280 281 303 319 340 REMS obtained C indexes between 0.74 and 0.88.91 233 319 These studies were too heterogeneous and biased to meta-analyse: they used varying outcome definitions (mortality, ICU admission, various composites, with time horizons varying from 1 to 30 days), from different populations (Italy, UK, Norway, China), and were at high or unclear risk of bias.

### Discussion

In this systematic review of prognostic prediction models related to the covid-19 pandemic, we identified and critically appraised 606 models described in 310 studies. These prognostic models can be divided into models to predict the risk of developing covid-19 or having an adverse disease course in the general population (n=13), and models to support the prognosis of patients with covid-19 (n=593). Most
Table 2 | Prediction models for covid-19 with low risk of bias

<table>
<thead>
<tr>
<th>Study; setting; and outcome</th>
<th>Model</th>
<th>Sample size (total No of participants (No with outcome))</th>
<th>Predictive performance</th>
<th>Strongest type of validation reported</th>
<th>C index (95% CI)*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General population models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clift et al231; data from UK, men registered at GP, death with covid-19</td>
<td>Qcovid mortality (male)</td>
<td>Development 3 047 693 (1867); external validation 1 097 268 (744)</td>
<td>External validation, new centres, same country</td>
<td>0.93 (0.92 to 0.94)</td>
<td></td>
</tr>
<tr>
<td>Clift et al231; data from UK, women registered at GP, death with covid-19</td>
<td>Qcovid mortality (female)</td>
<td>Development 3 035 409 (2517); external validation 1 075 788 (978)</td>
<td>External validation, new centres, same country</td>
<td>0.93 (0.92 to 0.94)</td>
<td></td>
</tr>
<tr>
<td>Clift et al231; data from UK, men registered at GP; hospital admission for covid-19 or death with covid-19</td>
<td>Qcovid hospital admission (male)</td>
<td>Development 3 047 693 (5962); external validation 1 097 268 (2076)</td>
<td>External validation, new centres, same country</td>
<td>0.86 (0.85 to 0.87)</td>
<td></td>
</tr>
<tr>
<td>Clift et al231; data from UK, women registered at GP; hospital admission for covid-19 or death with covid-19</td>
<td>Qcovid hospital admission (female)</td>
<td>Development 3 035 409 (4814); external validation 1 075 788 (1627)</td>
<td>External validation, new centres, same country</td>
<td>0.85 (0.84 to 0.86)</td>
<td></td>
</tr>
</tbody>
</table>

**Models for patients with covid-19**

<table>
<thead>
<tr>
<th>Study; setting; and outcome</th>
<th>Model</th>
<th>Sample size (total No of participants (No with outcome))</th>
<th>Predictive performance</th>
<th>Strongest type of validation reported</th>
<th>C index (95% CI)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carr et al231; data from UK, China and Norway; patients admitted to hospital with confirmed covid-19; 14 day ICU admission or death</td>
<td>Carr model</td>
<td>Development 1 276 (389); external validation 6 237 (1308)</td>
<td>External validation, new centres, different countries</td>
<td>0.79 (not reported)</td>
<td></td>
</tr>
<tr>
<td>Goodacre et al269; data from UK; patients with suspected symptoms of covid-19 at the emergency department; 30 day death or organ support</td>
<td>PRIEST score</td>
<td>Development 11 773 (2440); external validation 9 118</td>
<td>External validation, new centres, same country</td>
<td>0.80 (0.79 to 0.81)</td>
<td></td>
</tr>
<tr>
<td>Gupta et al268; data from the UK; hospitalised symptomatic suspected or confirmed cases; ventilatory support, critical care, or in-hospital death</td>
<td>ISARIC4C Deterioration model</td>
<td>Development 66 705 (28 140); external validation 8 239 (3 784)</td>
<td>External validation, new centres, same country</td>
<td>0.77 (0.76 to 0.78)</td>
<td></td>
</tr>
</tbody>
</table>

*Performance from strongest type of validation reported.

studies reported moderate to excellent predictive performance, but only seven newly developed models and 22 external validations of existing models were at low risk of bias. From these, we identified eight models, all developed for prognosis of covid-19, with adequate performance and low risk of bias at model development (four Qcovid models,231 the PRIEST model,262 the ISARIC4C deterioration model,268 and Carr’s model81) or external validation (Xie’s model19 269). We suggest that these models should be further validated within other datasets and settings, and ideally by independent investigators, to investigate which models maintain a robust performance over time and in varying settings.

Most of the 606 models were appraised to have high or uncertain risk of bias owing to a combination of poor reporting and poor methodological conduct. 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.438 Other common causes for bias were not adequately accounting for missing data, using techniques that do not account for optimism in performance estimates, ignoring model calibration, and inappropriate model validation. 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 indexes, often indicating near perfect discrimination, are probably optimistic. For most of these models, no independent external validations with a low risk of bias were performed, even though most were publicly available in a format usable in clinical practice.

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

Table 3 | External validations with low risk of bias from Gupta et al269

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Model</th>
<th>C index (95% CI)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 day deterioration</td>
<td>NEWS2*51</td>
<td>0.78 (0.73 to 0.83)</td>
</tr>
<tr>
<td>10 day deterioration</td>
<td>Carr (preprint)*123</td>
<td>0.56 (0.50 to 0.62)</td>
</tr>
<tr>
<td>14 day deterioration</td>
<td>Carr (preprint)*123</td>
<td>0.56 (0.74 to 0.82)</td>
</tr>
<tr>
<td>14 day deterioration</td>
<td>Guo*57</td>
<td>0.67 (0.61 to 0.73)</td>
</tr>
<tr>
<td>In-hospital deterioration</td>
<td>Zhang (preprint)*</td>
<td>0.74 (0.69 to 0.79)</td>
</tr>
<tr>
<td>In-hospital deterioration</td>
<td>Galloway*19</td>
<td>0.72 (0.68 to 0.77)</td>
</tr>
<tr>
<td>In-hospital deterioration</td>
<td>TAKIC*53</td>
<td>0.70 (0.65 to 0.75)</td>
</tr>
<tr>
<td>In-hospital deterioration</td>
<td>Colombi*</td>
<td>0.69 (0.63 to 0.74)</td>
</tr>
<tr>
<td>In-hospital deterioration</td>
<td>Huang*39</td>
<td>0.67 (0.10 to 0.73)</td>
</tr>
<tr>
<td>In-hospital deterioration</td>
<td>Shi*33</td>
<td>0.61 (0.56 to 0.66)</td>
</tr>
<tr>
<td>In-hospital deterioration</td>
<td>MEWS*24</td>
<td>0.60 (0.54 to 0.65)</td>
</tr>
<tr>
<td>12 day mortality</td>
<td>Lu*45</td>
<td>0.72 (0.67 to 0.76)</td>
</tr>
<tr>
<td>30 day mortality</td>
<td>CURB-65*35</td>
<td>0.75 (0.70 to 0.80)</td>
</tr>
<tr>
<td>30 day mortality</td>
<td>Bello-Chavolla*16</td>
<td>0.66 (0.60 to 0.72)</td>
</tr>
<tr>
<td>In-hospital mortality</td>
<td>REMS*34</td>
<td>0.76 (0.71 to 0.81)</td>
</tr>
<tr>
<td>In-hospital mortality</td>
<td>Xie*19</td>
<td>0.76 (0.69 to 0.82)</td>
</tr>
<tr>
<td>In-hospital mortality</td>
<td>Hu*35</td>
<td>0.74 (0.68 to 0.79)</td>
</tr>
<tr>
<td>In-hospital mortality</td>
<td>Caramelo*44</td>
<td>0.71 (0.66 to 0.76)</td>
</tr>
<tr>
<td>In-hospital mortality</td>
<td>Zhang (death)*123</td>
<td>0.70 (0.65 to 0.76)</td>
</tr>
<tr>
<td>In-hospital mortality</td>
<td>qSOFA*177</td>
<td>0.60 (0.55 to 0.65)</td>
</tr>
<tr>
<td>In-hospital mortality</td>
<td>Yan*26</td>
<td>0.58 (0.49 to 0.67)</td>
</tr>
</tbody>
</table>

NEWS2=rational early warning score; TAKIC=therapeutic study in pre-ICU patients admitted with covid-19; MEWS=modified early warning score; REMS=rapid emergency medicine score; qSOFA=quick sequential (sepsis-related) organ failure assessment; CURB-65=confusion, urea, respiratory rate, blood pressure plus age of at least 65 years; CI=confidence interval.

*Performance from strongest type of validation reported.
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 is suited for their population. However, 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 researchers adhere to the TRIPOD reporting guideline15 16 to improve the description of their study population and guide their modelling choices. TRIPOD translations (eg, in Chinese) 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. The variability in mortality could be related to differences in included patients (eg, age, comorbidities) but also in interventions for covid-19.

In this living review, inadequate sample size to build a robust model or to obtain reliable performance statistics was one of the most prevalent shortcomings. We recommend researchers should make use of formulas and software that have been made available in recent years to calculate the required sample size to build or externally validate models.439-442 The current review also identified that ignoring missing data and performing a complete case analysis is still very common. As this leads to reduced precision and can introduce bias in the estimated model, we recommend researchers address missing data using appropriate techniques before developing or validating a model.443 444 When creating a new prediction model, we recommend building on previous literature and expert opinion to select predictors, rather than selecting predictors purely based on data.17 This recommendation is especially important for datasets with limited sample size.444 To temper optimism in estimated performance, several internal validation strategies can be used—for example, bootstrapping.17 446 We also recommend that researchers should evaluate model performance in terms of correspondence between predicted and observed risk, preferably using flexible calibration plots17 447 in addition to discrimination.

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 neither recovered nor died within that 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.17 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.448

A prediction model applied in a new healthcare setting or country often produces predictions that are miscalibrated447 and might need to be updated before it can safely be applied in that new setting.17 This requires data from patients with covid-19 to be available from that setting. In addition to 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.446 449-452

The covid-19 pandemic has been characterised by an unprecedented speed of data accumulation worldwide. Unfortunately, much of the work done to analyse all these data has been ill informed and disjointed. As a result, we have hundreds of similar models, and very few independent validation studies comparing their performance on the same data. To leverage the full potential of prediction models in emerging pandemics and quickly identify useful models, international and interdisciplinary collaboration in terms of data acquisition, model building, model validation, and systematic review is crucial.

Study limitations
With new publications on covid-19 related prediction models entering the medical literature in unprecedented numbers and at unprecedented speed, this systematic review cannot be viewed as an up-to-date list of all currently available covid-19 related prediction models. It does provide a comprehensive overview of all prognostic model developments and validations from the first year of the pandemic up to 17 February 2021. Also, 69 of the studies we reviewed were only available as preprints. Some of these studies might enter the official medical literature in an improved version, after peer review. We reassessed peer reviewed publications of preprints included in previous updates that have been published before the current update. We also found other prediction models have been used in clinical practice without scientific publications,453 and web risk calculators launched for use while the scientific manuscript is still under review (and unavailable on request).455 These unpublished models naturally fall outside the scope of this review of the literature. As we have argued extensively elsewhere,455 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
This living review has identified a handful of models developed specifically for covid-19 prognosis with good predictive performance at external validation, and with model development or external validation at low risk of bias. The Qcovid models231 were built to prognosticate hospital admission and mortality
risks in the general population. The PRIEST model was proposed to triage patients at the emergency department.262 The ISARIC4C Deterioration model,268 Carr model,269 and Xie model19 were developed to predict adverse outcomes in hospitalised patients (ventilatory support, critical care or death, ICU admission or death, and death, respectively). Since the search date, these models have been validated temporally and geographically, which demonstrated that care should be taken when using these models in policy or clinical practice.231 268 456-460 Differences that care should be taken when using these models temporally and geographically, which demonstrated the search date, these models have been validated admission or death, and death, respectively). Since (ventilatory support, critical care or death, ICU predict adverse outcomes in hospitalised patients (ventilatory support, critical care or death, ICU and most models developed specifically for covid-19 were not validated independently. This oversupply of insufficiently validated models is not useful for clinical practice. Moreover, the urgency of diagnostic and prognostic models to assist in quick and efficient triage of patients in an emergent pandemic might encourage clinicians and policymakers to prematurely implement prediction models without sufficient documentation and validation. Inaccurate models could even cause more harm than good.461 By pointing to the most important methodological challenges and issues in design and reporting, we hope to have provided a useful starting point for future studies and future epidemics.

Conclusion
Several prognostic models for covid-19 are currently available and most report moderate to excellent discrimination. However, many of these models are at high or unclear risk of bias, mainly because of model overfitting, inappropriate model evaluation (eg, calibration ignored), and inappropriate handling of missing data. Therefore, their performance estimates are probably optimistic and might not be representative for the target population. We found that the Qcovid models can be used for risk stratification in the general population, while the PRIEST model, ISARIC4C Deterioration model, Carr’s model, and Xie’s model are suitable for prognostication in a hospital setting. The performance of these models is likely to vary over time and differ between regions, necessitating further validation and potentially updating before implementation. For details of the reviewed models, see https://www.covprecise.org/. Sharing data and expertise for the validation and updating of covid-19 related prediction models is still needed.
The authors and patient partners will distribute our findings and summarising the results for a general audience; and survivors of critical covid-19 for their help in interpreting the study. 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 no discrepancies from the study as planned have been explained.

Dissemination to participants and related patient and public communities: The authors and patient partners will distribute this information through their institutions and on social media to provide an opportunity for public dialogue and as an example of how what we learn as the result of a new disease changes and improves over time. 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 no discrepancies from the study as planned have been explained.

Data sharing: 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 terms of the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, provided the original work is properly cited. See: http://creativecommons.org/licenses/by/4.0/.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare: support from Commercial Use of Data and Meta-analyses in Critical Care (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: Original data are available online on https://osf.io/ehc47/.

Funding: 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 no discrepancies from the study as planned have been explained.


1. COVID-19 has disrupted essential health services globally, affecting the provision of primary care and hospital services.
2. Machine learning and artificial intelligence (AI) systems have been developed to assist in the diagnosis and management of COVID-19.
3. These systems use CT images and clinical data to predict disease severity, aiding in the triage of patients.
4. A machine learning-based CT radiomics model, trained on 601 COVID-19 cases, achieved 91.8% accuracy in predicting COVID-19 from chest CT images.
5. The model can be used to screen for COVID-19 and detect disease progression, with potential future applications in monitoring patients.
6. The development and validation of such prediction models are crucial for optimizing healthcare resources and improving patient outcomes.
7. Further research is needed to validate these models in different populations and settings, to ensure their generalizability.
8. The integration of AI systems into clinical practice requires careful consideration of ethical, privacy, and regulatory issues to ensure responsible and equitable use.


Prognostic Risk Score System for COVID-19 Inpatients: A Multi-Center peerj.9885


Web appendix: Supplementary material