Systematic review and critical appraisal of prediction models for diagnosis and prognosis of COVID-19 infection

Laure Wynants, assistant professor[1 2] (0000-0002-3037-122X), Ben Van Calster, associate professor[2 3] (0000-0003-1613-7450), Marc MJ Bonten, professor[4 5] (0000-0002-9095-9201), Gary S Collins, professor[6 7] (0000-0002-2772-2316), Thomas PA Debray[4 8], assistant professor (0000-0002-1790-2719), Maarten De Vos, associate professor[2 9](0000-0002-3482-5145), Maria C. Haller, medical doctor[10 11] (0000-0003-4635-6291), Georg Heinze, associate professor[10] (0000-0003-1147-8491), Karel GM Moons, professor[4 8], Richard D Riley, professor[12] , Ewoud Schuit, assistant professor (0000-0002-9548-3214)[4 8], Luc Smits, professor[1] (0000-0003-0785-1345), Kym IE Snell, lecturer[12] (0000-0001-9373-6591), Ewout W Steyerberg, professor[3] (0000-0002-7787-0122), Christine Wallisch, research fellow[10, 13 14] (0000-0003-3943-6234), Maarten van Smeden, assistant professor[4] (0000-0002-5529-1541)

1 Department of Epidemiology, CAPHRI Care and Public Health Research Institute, Maastricht University, Maastricht, Netherlands

2 KU Leuven, Department of Development and Regeneration, Leuven, Belgium

3 Department of Biomedical Data Sciences, Leiden University Medical Centre (LUMC), Leiden, the Netherlands

4 Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

5 Department of Medical Microbiology, University Medical Centre Utrecht, Utrecht, the Netherlands.

6 Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Musculoskeletal Sciences, University of Oxford, UK

7 NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, United Kingdom NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, United Kingdom

8 Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

9 KU Leuven, Department of Electrical Engineering, ESAT Stadius, Leuven, Belgium

10 Section for Clinical Biometrics; Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, Austria

11 Ordensklinikum Linz, Hospital Elisabethinen, Department of Nephrology, Linz, Austria

12 Centre for Prognosis Research, School of Primary, Community and Social Care, Keele University, UK.

13 Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Institute of Biometry and Clinical Epidemiology, Charitéplatz 1, 10117 Berlin, Germany.

14 Berlin Institute of Health (BIH), Anna-Louisa-Karsch 2, 10178 Berlin, Germany

Corresponding author:

Laure Wynants

Department of Epidemiology

CAPHRI Care and Public Health Research Institute

Maastricht University

Peter Debyeplein 1, 6229 HA Maastricht, Netherlands

T 0032 479 01 03 15

E laure.wynants@maastrichtuniversity.nl

WORD COUNT MAIN TEXT: 2940

Abstract (537 words)

Objective To review and critically appraise published and preprint reports of models that aim to predict either (i) presence of existing COVID-19 infection, or (ii) future complications in individuals already diagnosed with COVID-19. Any models to identify subjects at risk for COVID-19 in the general population were also included.

Design Rapid systematic review and critical appraisal of prediction models for diagnosis or prognosis of COVID-19 infection.

Data sources PubMed, EMBASE via Ovid, Arxiv, medRxiv and bioRxiv until 13th March 2020.

Study selection Studies that developed or validated a multivariable COVID-19 related prediction model. Two authors independently screened titles and abstracts.

Data extraction Data from included studies were extracted independently by at least two authors based on the CHARMS checklist, and risk of bias was assessed using PROBAST. Data were extracted on various domains including the participants, predictors, outcomes, data analysis, and prediction model performance.

Results 1916 titles were screened. Of these, 15 studies describing 19 prediction models were included for data extraction and critical appraisal. We identified three models to predict hospital admission from pneumonia and other events (as a proxy for covid-19 pneumonia) in the general population; nine diagnostic models to detect COVID-19 infection in symptomatic individuals (seven of which were deep learning models for COVID-19 diagnosis utilising computed tomography (CT) results); and seven prognostic models for predicting mortality risk, or length of hospital stay. None of the 15 studies used data on COVID-19 cases outside of China. Predictors included in more than one of the 19 models were: age, sex, comorbidities, C-reactive protein, lymphocyte markers (percentage or neutrophil-to-lymphocyte ratio), lactate dehydrogenase, and features derived from CT images. Reported C-index estimates for the prediction models ranged from 0.73 to 0.81 in those for the general population (reported for all 3 general population models), from 0.81 to > 0.99 in those for diagnosis (reported for 5 of the 9 diagnostic models), and from 0.90 to 0.98 in those for prognosis (reported for 4 of the 7 prognostic models). All studies were rated at high 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, and poor statistical analysis, including high risk of model overfitting. Reporting quality varied substantially between studies. A description of the study population and intended use of the models was absent in almost all reports, and calibration of predictions was rarely assessed.

Conclusion COVID-19 related prediction models for diagnosis and prognosis are quickly entering the academic literature through publications and preprint reports, aiming to support medical decision making in a time where this is needed urgently. Many models were poorly reported and all appraised as high risk of bias. We call for immediate sharing of the individual participant data from COVID-19 studies worldwide to support collaborative efforts in building more rigorously developed and validated COVID-19 related prediction models. The predictors identified in current studies should be considered for potential inclusion in new models. We also stress the need to adhere to methodological standards when developing and evaluating COVID-19 related predictions models, as unreliable predictions may cause more harm than benefit when used to guide clinical decisions about COVID-19 in the current pandemic.

Systematic review registration osf.io/ehc47/

Summary boxes

What is already known on this topic

- The sharp recent increase in COVID-19 infections has put a strain on healthcare systems worldwide, necessitating efficient diagnosis of patients suspected of the infection and prognostication of COVID-19 confirmed cases.
- Viral nucleic acid testing and chest CT are standard methods for diagnosing COVID-19, but are time-consuming.
- Earlier reports suggest that the elderly, patients with comorbidity (COPD, cardiovascular disease, hypertension), and patients presenting with dysapnoea are vulnerable to more severe morbidity and mortality after COVID-19 infection.

What this study adds

- We identified three models to predict hospital admission from pneumonia and other events (as a proxy for covid-19 pneumonia) in the general population.
- We identified nine diagnostic models for COVID-19 detection in symptomatic patients. Seven of these were neural network models based on CT images.
- We identified seven prognostic models, of which five aimed to predict mortality risk in confirmed COVID-19 patients and two aimed to predict a hospital stay of more than 10 days from admission.
- All included studies were appraised at high risk of bias, suggesting concern that the models may be flawed and perform poorly when applied in practice, such that their predictions may be unreliable.

INTRODUCTION

The novel coronavirus (COVID-19) presents a significant and urgent threat to global health. Since the outbreak in early December 2019 in the Hubei Province of the People's Republic of China, more than 318.000 cases have been confirmed in over 160 countries, and over 13.000 people died (up to 22st March).[1] 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 may follow.[2 3] Outbreaks lead to important increases in the demand for hospital beds and shortage of medical equipment, while medical staff themselves may also get infected.

To mitigate the burden on the health care system, while also providing the best possible care for patients, efficient diagnosis and prognosis is needed. Prediction models, which combine multiple predictors (variables or features) to estimate the risk of being infected or experiencing poor outcome of the infection, could assist medical staff in triaging patients when allocating limited healthcare resources. Prediction models, ranging from rule-based scoring systems to advanced machine learning models (deep learning), have already 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.[4] Many of these prediction models are published in open access repositories, ahead of peer-review.

We aimed to systematically review and critically appraise currently available COVID-19 related prediction models, in particular models for diagnosis of COVID-19 in suspected cases or models for prognosis of individuals in confirmed cases. This systematic review was done in collaboration with the Cochrane Prognosis Methods group.

METHODS

We searched PubMed, EMBASE via Ovid, bioRxiv, medRxiv, and arXiv for research on COVID-19 published after 3rd January 2020. We used the publicly available publication list of the COVID-19 Living Systematic Review.[5] This list contains studies on COVID-19 published on PubMed, EMBASE via Ovid, bioRxiv, and medRxiv, and is continuously updated. We validated the list to examine whether it is fit for purpose by comparing it to relevant hits from bioRxiv and medRxiv when combining COVID-19 search terms (covid-19, sars-cov-2, "novel corona", 2019-ncov) with methodological search terms (diagnostic, prognostic, prediction model, machine learning, artificial intelligence, algorithm, score, deep learning, regression). All relevant hits were found on the Living Review list. We supplemented the Living Review list with hits from PubMed searching for “covid-19”, as this was at the moment of our search not included in the Living Review search terms for PubMed. We further supplemented the Living Review list with studies on COVID-19 retrieved from arXiv. The search strings are listed in the Supplementary Material. In addition, we reached out to authors to include studies that were not publicly available at the time of the search.[6 7]

Databases were searched on 13th March 2020. All studies were considered, regardless of language or publication status (preprint or peer reviewed articles). Studies were included if they developed or validated a multivariable model or scoring system, based on individual participant level data, to predict any COVID-19 related outcome in individuals, either to inform diagnosis or prognosis. There was no restriction on setting (e.g., in- or outpatients), prediction horizon, included predictors, or outcomes. Prediction models to detect individuals at risk of developing COVID-19 pneumonia in the general population were also included. Epidemiological studies that aimed at modelling disease transmission or case-fatality rates, diagnostic test accuracy and predictor finding studies were excluded. Titles, abstracts and full texts were screened in duplicate for eligibility by pairs of independent reviewers (from LW, BVC, MvS), and discrepancies were resolved through discussion.

Data extraction of included articles was done by two independent reviewers (from LW, BVC, GSC, TPAD, MCH, GH, KGM, RDR, ES, LS, EWS, KIES, CW and MvS), using a standardized data extraction form based on the CHARMS checklist [8] and Prediction model Risk Of Bias ASsessment Tool (PROBAST) [9]. We sought to extract each model's predictive performance, using whatever measures were presented, including any summaries of calibration (the extent to which predicted risks correspond to observed risks) and discrimination (the extent to which predicted risks discriminate between participants with and without the outcome), as recommended in the TRIPOD statement.[10] Any discrepancies in data extraction were resolved by LW and MvS. Details on data extraction are provided in the Supplementary Material. Reporting of the article considered aspects of PRISMA[11] and TRIPOD [10].

RESULTS

A total of 1914 titles were retrieved through our systematic search (Figure 1). Two additional unpublished studies were made available upon request. These were not yet publicly available at the time of the search and were identified after a call on social media. Out of 1916 titles, 46 studies were retained for abstract and full text screening. Fifteen studies, describing nineteen prediction models, met the inclusion criteria and were selected for data extraction and critical appraisal.[6 7 12-24]

Primary datasets

All of the 15 studies used data on COVID-19 cases from China (see Supplementary Table 1). Based on nine of the 15 studies that reported study dates, data were collected between 8th December 2019 and 3rd March 2020. The duration of follow-up was unclear in most studies, although one reported a median follow-up of 8.4 days.[13] Some Chinese centers provided data to multiple studies, but it was unclear how much these datasets overlapped across our 15 identified studies. One study used U.S. Medicare claims data from 2015 to 2016 to estimate COVID-19 vulnerability,[7] one study used control CT scans from the USA,[19] and one study used simulated data.[12]

All but one study[18] developed prediction models for use in adults. The median age varied between studies (from 38 to 65 years, see Supplementary Table 1), as did the percentage of men (from 35% to 61%).

Among the seven studies that developed prognostic models to predict mortality risk in individuals with confirmed or suspected COVID-19 infection, the percentage of deaths varied between 8% and 59% (See Table 1). This wide variation is in part due to severe sampling bias caused by studies excluding participants who had not experienced the event by the end of the study period (i.e., neither healed nor died). In addition, length of follow-up may have varied between studies, but was not reported.

Among the nine diagnostic model studies, there was only one that reported on prevalence of COVID-19 infection in those suspected of having COVID-19; the prevalence was 19% (development dataset) and 24% (validation dataset).[24] Since the seven imaging studies used either case-control sampling or the method of data collection was unclear, the prevalence in these diagnostic studies may not have been representative of their target population.

In what follows, we give an overview of the 19 prediction models reported in the 15 identified studies (Table 1). Modeling details are provided in Supplementary Table 2, and the availability of models in a format for use in clinical practice is discussed in Box 1.

Models to predict the risk of hospital admission due to COVID-19 pneumonia in the general population

Three models predicted the risk of hospital admission for COVID-19 pneumonia for individuals in the general population, but used admission due to non-tuberculosis pneumonia, influenza, acute bronchitis, or upper respiratory infections as outcomes in a dataset without any COVID-19 cases (see Table 1).[7] Among the predictors were age, sex, previous hospital admissions, comorbidity data, and social determinants of health. The models reported C-index estimates of 0.73, 0.81 and 0.81.

Diagnostic models to detect COVID-19 infection in symptomatic individuals

One study developed a warning score to diagnose COVID-19 in symptomatic adults based on sex, age, fever, highest body temperature between onset and admission, history of close contact with confirmed patients, signs of pneumonia on CT, neutrophil-to-lymphocyte ratio, and meaningful respiratory syndromes (see Table 1).[24] They report a C-index estimate of 0.97. One study developed a decision tree to diagnose severe disease in symptomatic paediatric inpatients based on direct bilirubin and alaninetransaminase.[18] They report an estimated F1 score of 1.00 (indicating 100% observed sensitivity and specificity).

Seven prediction models were proposed to support the diagnosis of COVID-19 or COVID-19 pneumonia (and monitor progression) based on CT images. The reported predictive performance varied widely, with C-index estimates ranging from 0.81 to nearly 1.

Prognostic models for patients diagnosed with COVID-19 infection

We identified seven prognostic models (Table 1). Of these, five estimated mortality risk in suspected or confirmed COVID-19 patients.[12 13 15 16 25] The intended use of these models (i.e., when to use it, in whom to use it, and the prediction horizon (e.g., mortality by what time)) was not clearly described. Two models aimed to predict a hospital stay of more than 10 days from admission.[14] Predictors that were included in more than one prognostic model were age (n=3), features derived from CT-scoring (n=3), C-reactive protein (n=2), lactic dehydrogenase (n=2), and lymphocyte count (n=2) (see Table 1).

Only two studies predicting mortality reported a C-index; they obtained estimates of 0.90 [16] and 0.98 [6]. Only one study evaluated calibration.[6] When applied to new patients, their model yielded probabilities of mortality that were too high for low-risk patients and too low for high-risk patients (calibration slope >1), despite excellent discrimination.[6] One study developed two models to predict a hospital stay of >10 days and reported C-indexes of 0.92 and 0.96.[14]

Risk of bias

All models were at high risk of bias according to assessment with PROBAST (Table 1), which suggests that their predictive performance when used in practice is likely lower than what is reported, and so gives concern that their predictions are unreliable. Details on common causes for risk of bias are given in Box 2 for each type of model.

Six of the fifteen studies had a high risk of bias for the “participants” domain (Table 2), indicating that the participants enrolled in the studies may not be representative for the models’ targeted populations. Unclear reporting on the inclusion of participants prohibited a risk of bias assessment in six studies. Three out of sixteen studies had a high risk of bias for the "predictors" domain, indicating that predictors were not available at the models' intended time of use, not clearly defined, or influenced by the outcome measurement. The diagnostic model studies that used CT imaging predictors were all scored as "unclear" on the "predictors" domain. The publications often lacked clear information on the preprocessing steps (e.g., cropping of images). Moreover, translation of CT image results to predictors is challenging, which makes the PROBAST predictors section hard to complete reliably. Most studies used outcomes that are easy to assess (e.g., death, presence of COVID-19 by laboratory confirmation). Nonetheless, there was reason to be concerned of bias induced by the outcome measurement in five studies, due to the use of subjective or proxy-outcomes.

All studies were at high risk of bias for the "analysis" domain (Table 2). Many studies had small sample sizes (Table 1), leading to an increased risk of overfitting, particularly if complex modeling strategies were used. Two studies did not report the predictive performance of the developed model, and two studies reported the apparent performance only (that is, the performance in the exact same data as was used to develop the model, without adjustment for optimism due to potential overfitting). Only one study assessed calibration (i.e., the extent to which predicted risks corresponded to observed risks).

DISCUSSION

Main findings

In this systematic review of prediction models related to the COVID-19 pandemic, we identified and critically appraised 15 studies that described 19 prediction models for detecting individuals at risk for hospital admission for COVID-19 pneumonia in the general population, for diagnosis of COVID-19 in symptomatic individuals, and for prognosis of COVID-19 infected patients. All models reported good to even excellent predictive performance, but all were appraised as high risk of bias, due to a combination of poor reporting and poor methodological conduct for participant selection, predictor description and statistical methods used. As expected, in these early COVID-19 related prediction model studies, clinical data from COVID-19 patients is still scarce and limited to data from China. With few exceptions, the available sample size and number of events for the outcomes of interest were limited, which is a known problem for building prediction models, increasing the risk of overfitting the prediction model.[26] A high risk of bias implies that these models are likely to perform worse in practice than the performance that is reported by the researchers. Only two studies carried out an external validation on data from other individuals than from which the models was developed, and only one study assessed calibration (i.e., the correspondence between predicted and observed risks).

We reviewed seven studies that used advanced machine learning methodology on chest CT scans to diagnose COVID-19 disease, COVID-19 related pneumonia, or to assist in segmentation of lung images. The predictive performance measures showed a high to almost perfect ability to identify COVID-19, although these models and their evaluations also suffered from a high risk of bias, notably due to an artificial mix of COVID-19 cases and non-cases.

Challenges and opportunities

The main aim of prediction models is to support medical decision making. It is therefore key to identify a target population in which predictions serve a clinical need, and a representative dataset on which the prediction model can be developed and validated. This target population must also be carefully described such that the performance of the developed or validated model can be appraised in context, and users know in which individuals the model can be applied to make predictions. However, the included studies in our systematic review often lacked an adequate description of the study population, which leaves users of these models in doubt of the models' applicability. While we recognize that all studies were done under severe time constraints caused by urgency, we recommend that studies currently in preprint and future studies use and adhere to the TRIPOD reporting guideline[10] to improve the description of their study population as well as their modeling choices. TRIPOD translations (e.g., in Chinese and Japanese) are also available.

A better description of the study population may also help understand the observed variability in the reported outcomes across studies, such as COVID-19 related mortality. The variability in the relative frequencies of the predicted outcomes presents an important challenge to the prediction modeler: a prediction model applied in a setting with a different relative frequency of the outcome may produce predictions that are miscalibrated [27] and may need to be updated before it can safely be applied in that new setting.[28] Indeed, such an update may often be required when prediction models are transported to different healthcare systems, which requires COVID-19 patient data to be available from that system.

Instead of developing and updating predictions in their local setting, Individual Participant Data (IPD) from multiple countries and healthcare systems may facilitate better understanding of the generalizability and implementation prediction models across different settings and populations, and may greatly improve their applicability and robustness in routine care.[29-32]

The evidence base for the development and validation of prediction models related to COVID-19 will quickly increase over the coming months. Together with the increasing evidence from predictor finding studies [33] and open peer review initiatives for COVID-19 related publications,[34] data registries [35-38] are being set up. To maximize the new opportunities and to facilitate IPD meta-analyses, the WHO has recently released a new data platform to encourage sharing of anonymized COVID-19 clinical data.[39] To leverage the full potential of these evolutions, international and interdisciplinary collaboration in terms of data acquisition and model building is crucial.

Limitations of this study

With new publications on COVID-19 related prediction models that are currently quickly 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, studies we reviewed were only available as a preprint, and they might improve after peer review, when entering the official medical literature. Finally, we have also found other prediction models which are currently implemented in clinical practice without scientific publications [40] and web risk calculators launched for use while the scientific manuscript was still under review (and unavailable upon request).[41] These unpublished models naturally fall outside the scope of this review of the literature.

Implications for practice

All nineteen reviewed prediction models were found to have a high risk of bias and evidence from independent external validation of these models is currently lacking. However, the urgency of prediction models to assist in quick and efficient triage of patients in the COVID-19 pandemic may encourage clinicians to implement prediction models without sufficient documentation and validation. Although we cannot let perfect be the enemy of good, earlier studies have shown that models were of limited use in the context of a pandemic,[42] and they may even cause more harm than good.[43] We anticipate that more COVID-19 data on the individual participant level will soon become available. Based on the predictors included in models identified by our review, we encourage researchers to include data on age, sex, comorbidities, C-reactive protein, lymphocyte markers (percentage or neutrophil-to-lymphocyte ratio), lactate dehydrogenase, and potentially features derived from CT images when collecting data or building new models. 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 aiming at developing new models or validating and updating existing ones.

CONCLUSION

Diagnostic and prognostic models for COVID-19 are available and they all appear to show good to excellent discriminative performance. However, their performance estimates are likely to be optimistic and thus misleading, as all identified studies were at high risk of bias. Sharing data and expertise for development, validation and updating of COVID-19 related prediction models is urgently needed.

Box 1. Availability of the models in a format for use in clinical practice

Models to predict hospital admission for COVID-19 pneumonia in the general population. The "CV-19 vulnerability index" to detect hospital admission for COVID-19 pneumonia from other respiratory infections (e.g. pneumonia, influenza), is available as an online tool.[7 44]

Diagnostic models. The "COVID-19 Early Warning Score" to detect COVID-19 infection in adults is available as a score chart in an article.[24] A decision tree to detect severe disease for pediatric COVID-19 confirmed patients is also available in an article.[18]

Diagnostic models based on CT imaging. Three of the seven AI models to assist with diagnosis based on CT results, are available via web applications. [17 20 23 45-47]

Prognostic models. To assist in the prognosis of mortality, a nomogram (a graphic aid to calculate mortality risk),[6] a decision tree,[15] and a CT-based scoring rule are available in the articles.[16]

All other 9 reports did not include any usable equation, format or reference for use of their prediction model.

Because all models were at high risk of bias, we cannot recommend their routine use before they are properly externally validated.

Box 2. Common causes of risk of bias in the 19 reported prediction models.

Models to predict hospital admission for COVID-19 pneumonia in the general population. These models were based on Medicare claims data, and used proxy outcomes to predict hospital admission for COVID-19 pneumonia, in absence of COVID-19 cases.[7]

Diagnostic models. Individuals without COVID-19 were excluded, altering the disease prevalence.[24] Predictors were dichotomized, leading to a loss of information.[18 24]

Diagnostic models based on CT imaging. There was generally poor reporting on which patients CT images were obtained during clinical routine, and it was unclear whether the selection of controls was sampled from the target population (i.e., patients suspected of COVID-19).[17 23] It was often unclear how regions of interest (ROIs) were annotated. Images were sometimes annotated by only one scorer without quality control, or the model output influenced annotation.[19 21 22] Careful description of model specification and subsequent estimation was lacking, challenging the transparency and reproducibility of the models. Every study used a different deep learning architecture, including established and specifically designed ones, without benchmarking the used architecture with respect to others.

Prognostic models. Study participants were often simply excluded because they did not develop the outcome at the end of the study period but were still in follow-up (i.e., in the hospital and neither healed nor died), yielding a highly selected study sample. Only one study accounted for censoring by using Cox regression.[13] Other studies used highly subjective predictors,[16] or the last available predictor measurement from electronic health records was used (rather than the measurement of the predictor value at the time the model is intended to be used).[15] Dichotomization of predictors was often applied which tends to lead to loss of information.[18 24]

ACKNOWLEDGMENTS

We thank the authors who made their work available, by posting it on public registries or sharing it confidentially.

Contributors: LW conceived the study; LW and MvS designed the study; LW, MvS and BVC screened titles and abstracts for inclusion. LW, BVC, GSC, TPAD, MCH, GH, KGM, RDR, ES, LS, EWS, KIES, CW and MvS extracted and analysed data. MD helped interpret the findings on deep learning studies and MB and MCH assisted in the interpretation from a clinical viewpoint. LW and MvS wrote the first draft, which all authors revised for critical content. All authors approved the final manuscript. LW and MvS are the guarantors. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: LW is a post-doctoral fellow of Research Foundation – Flanders (FWO). BVC received support from FWO (grant G0B4716N) and Internal Funds KU Leuven (grant C24/15/037). TD acknowledges financial support from the Netherlands Organization for Health Research and Development (Grant Numbers: 91617050). KGMM gratefully acknowledges financial support from Cochrane Collaboration (SMF 2018). KIES is funded by the National Institute for Health Research School for Primary Care Research (NIHR SPCR). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. GSC was supported by the NIHR Biomedical Research Centre, Oxford, and Cancer Research UK (programme grant: C49297/A27294). The funders played no role in study design, data collection, data analysis, data interpretation, or reporting. The guarantors had full access to all the data in the study, take responsibility for the integrity of the data and the accuracy of the data analysis, and had final responsibility for the decision to submit for publication.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare they have no competing interests.

REFERENCES

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020 doi: 10.1016/s1473-3099(20)30120-1

2. Arabi YM, Murthy S, Webb S. COVID-19: a novel coronavirus and a novel challenge for critical care. Intensive Care Med 2020 doi: 10.1007/s00134-020-05955-1

3. Grasselli G, Pesenti A, Cecconi M. Critical Care Utilization for the COVID-19 Outbreak in Lombardy, Italy: Early Experience and Forecast During an Emergency Response. JAMA 2020 doi: 10.1001/jama.2020.4031

4. Sharing research data and findings relevant to the novel coronavirus (COVID-19) outbreak [Available from: https://wellcome.ac.uk/press-release/sharing-research-data-and-findings-relevant-novel-coronavirus-covid-19-outbreak accessed 22.03.2020.

5. Living Evidence on COVID-19 [Available from: https://ispmbern.github.io/covid-19/living-review/index.html accessed 22.03.2020.]

6. Xie J, Hungerford D, Abrams S, et al. Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19. 2020

7. DeCaprio D, Gartner J, Burgess T, et al. Building a COVID-19 Vulnerability Index. arXiv e-prints 2020. https://ui.adsabs.harvard.edu/abs/2020arXiv200307347D (accessed March 01, 2020).

8. Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014;11(10):e1001744. doi: 10.1371/journal.pmed.1001744

9. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Ann Intern Med 2019;170(1):W1-w33. doi: 10.7326/m18-1377

10. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and ElaborationThe TRIPOD Statement: Explanation and Elaboration. Ann Intern Med 2015;162(1):W1-W73. doi: 10.7326/M14-0698

11. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. PLoS Med 2009;6 doi: 10.1371/journal.pmed.1000100

12. Caramelo F, Ferreira N, Oliveiros B. Estimation of risk factors for COVID-19 mortality - preliminary results. medRxiv 2020:2020.02.24.20027268. doi: 10.1101/2020.02.24.20027268

13. Lu J, Hu S, Fan R, et al. ACP risk grade: a simple mortality index for patients with confirmed or suspected severe acute respiratory syndrome coronavirus 2 disease (COVID-19) during the early stage of outbreak in Wuhan, China. medRxiv 2020:2020.02.20.20025510. doi: 10.1101/2020.02.20.20025510

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

15. Yan L, Zhang H-T, Xiao Y, et al. Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. medRxiv 2020:2020.02.27.20028027. doi: 10.1101/2020.02.27.20028027

16. Yuan M, Yin W, Tao Z, et al. Association of radiologic findings with mortality of patients infected with 2019 novel coronavirus in Wuhan, China. medRxiv 2020:2020.02.22.20024927. doi: 10.1101/2020.02.22.20024927

17. Song Y, Zheng S, Li L, et al. Deep learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) with CT images. medRxiv 2020:2020.02.23.20026930. doi: 10.1101/2020.02.23.20026930

18. Yu H, Shao J, Guo Y, et al. Data-driven discovery of clinical routes for severity detection in COVID-19 pediatric cases. medRxiv 2020:2020.03.09.20032219. doi: 10.1101/2020.03.09.20032219

19. Gozes O, Frid-Adar M, Greenspan H, et al. Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic: Initial Results for Automated Detection & Patient Monitoring using Deep Learning CT Image Analysis. arXiv e-prints 2020. https://ui.adsabs.harvard.edu/abs/2020arXiv200305037G (accessed March 01, 2020).

20. Chen J, Wu L, Zhang J, et al. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study. medRxiv 2020:2020.02.25.20021568. doi: 10.1101/2020.02.25.20021568

21. Xu X, Jiang X, Ma C, et al. Deep Learning System to Screen Coronavirus Disease 2019 Pneumonia. arXiv e-prints 2020. https://ui.adsabs.harvard.edu/abs/2020arXiv200209334X (accessed February 01, 2020).

22. Shan F, Gao Y, Wang J, et al. Lung Infection Quantification of COVID-19 in CT Images with Deep Learning. arXiv e-prints 2020. https://ui.adsabs.harvard.edu/abs/2020arXiv200304655S (accessed March 01, 2020).

23. Wang S, Kang B, Ma J, et al. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). medRxiv 2020:2020.02.14.20023028. doi: 10.1101/2020.02.14.20023028

24. Song C-Y, Xu J, He J-Q, et al. COVID-19 early warning score: a multi-parameter screening tool to identify highly suspected patients. medRxiv 2020:2020.03.05.20031906. doi: 10.1101/2020.03.05.20031906

25. Xie J, Tong Z, Guan X, et al. Critical care crisis and some recommendations during the COVID-19 epidemic in China. Intensive Care Med 2020 doi: 10.1007/s00134-020-05979-7

26. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ 2020;368:m441. doi: 10.1136/bmj.m441

27. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. BMC Med 2019;17(1):1-7.

28. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York, NY: Springer US 2009.

29. Riley RD, Ensor J, Snell KI, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016;353:i3140. doi: 10.1136/bmj.i3140

30. Debray TP, Riley RD, Rovers MM, et al. Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. PLoS Med 2015;12(10):e1001886. doi: 10.1371/journal.pmed.1001886

31. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016;69:245-7. doi: 10.1016/j.jclinepi.2015.04.005

32. Wynants L, Kent DM, Timmerman D, et al. Untapped potential of multicenter studies: a review of cardiovascular risk prediction models revealed inappropriate analyses and wide variation in reporting. Diagn Progn Res 2019;3:6. doi: 10.1186/s41512-019-0046-9

33. Jain V, Yuan J-M. Systematic review and meta-analysis of predictive symptoms and comorbidities for severe COVID-19 infection. medRxiv 2020:2020.03.15.20035360. doi: 10.1101/2020.03.15.20035360

34. Johansson M, Saderi D. Open peer-review platform for COVID-19 preprints. Nature 2020;579(7797):29.

35. Xu K, Cai H, Shen Y, et al. [Management of corona virus disease-19 (COVID-19): the Zhejiang experience]. Zhejiang Da Xue Xue Bao Yi Xue Ban 2020;49(1):0.

36. COVID-19 DATABASE [Available from: https://www.sirm.org/category/senza-categoria/covid-19/ accessed 22.03.2020.]

37. Wynants L, Bouwmeester W, Moons KG, et al. A simulation study of sample size demonstrated the importance of the number of events per variable to develop prediction models in clustered data. J Clin Epidemiol 2015;68(12):1406-14. doi: 10.1016/j.jclinepi.2015.02.002

38. Covid chestxray dataset [Available from: https://github.com/ieee8023/covid-chestxray-dataset accessed 22.03.2020.]

39. Coronavirus disease (COVID-19) technical guidance: Early investigations protocols [Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/early-investigations accessed 22.03.2020.]

40. Infervision [Available from: https://global.infervision.com/ accessed 22.03.2020.]

41. COVID-19 Response Center [Available from: https://surgisphere.com/covid-19-response-center/ accessed 22.03.2020.]

42. Enfield K, Miller R, Rice T, et al. Limited Utility of SOFA and APACHE II Prediction Models for ICU Triage in Pandemic Influenza. Chest 2011;140(4):913A. doi: 10.1378/chest.1118087

43. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. Med Decis Making 2015;35(2):162-9. doi: 10.1177/0272989x14547233

44. The COVID-19 Vulnerability Index (CV19 Index) [Available from: https://closedloop.ai/cv19index/ accessed 22.03.2020.]

45. AI diagnostic system for 2019-nCoV [Available from: http://121.40.75.149/znyx-ncov/index accessed 22.03.2020.]

46. ai.nscc [Available from: https://ai.nscc-tj.cn/thai/deploy/public/pneumonia_ct accessed 22.03.2020.]

47. Discriminating COVID-19 Pneumonia from CT Images [Available from: http://biomed.nscc-gz.cn/server/Ncov2019 accessed 22.03.2020.]

Table 1. Overview of prediction models for diagnosis and prognosis of COVID-19 infection.

						Predictive performance upon validation			
	Study	**Setting**	**Outcome**	**Predictors in final model**	**Sample size: Total number of participants for model development set (number with outcome)**	**Type of validation *1**	**Sample size: Total number of participants for model validation (number with outcome)**	**Performance *1: C-index, sensitivity, specificity, PPV/NPV, calibration slope, other (CI, if reported)**	**Overall risk of bias using PROBAST**
Hospital admission in general population									
	Decaprio, Gartner, et al.	Data from US, general population	Hospital admission for COVID-19 pneumonia (proxy events)*2	Age; sex; number of previous hospital admissions; 11 diagnostic features; interactions between age and diagnostic features	1.5M (unknown)	Training-test split	369,865 (unknown)	C-index: 0.73	High
	Decaprio, Gartner, et al.	Data from US, general population	Hospital admission for COVID-19 pneumonia (proxy events)*2	Age and 500+ features related to diagnosis history	1.5M (unknown)	Training-test split	369,865 (unknown)	C-index: 0.81	High
	Decaprio, Gartner, et al.	Data from US, general population	Hospital admission for COVID-19 pneumonia (proxy events)*2	500+ undisclosed features, including age, diagnostic history, social determinants of health, Charlson comorbidity index	1.5M (unknown)	Training-test split	369,865 (unknown)	C-index: 0.81	High
Diagnosis									
	Song, Xu, et al.	Data from China, COVID-19 suspected cases (inpatients)	COVID-19 diagnosis	Fever; history of close contact; signs of pneumonia on CT; neutrofil-to-lymphocyte ratio; highest body temperature; sex; (age, meaningful respiratory syndromes)	304 (73)	Training-test split	95 (18)	C-index: 0.97 (0.93; 1.00)	High

	Yu, Shao, et al.	Data from China, pediatric inpatients COVID-19 confirmed cases	Severe disease (yes/no) defined based on clinical symptoms	Direct Bilirubin; Alaninetransaminase	105 (8)	Apparent performance only	Not applicable	F1 score: 1.00	High
Diagnostic imaging									
	Gozed, Frid-Adar, et al.	Data from China and USA*3, COVID-19 suspected cases	COVID-19 diagnosis	Not applicable	50 (unknown)	External validation with Chinese cases and U.S. controls.	unclear	C-index: 0.996 (0.989;1.000)	High
	Chen, Wu, et al.	Data from China, COVID-19 pneumonia suspected cases	COVID-19 pneumonia	Not applicable	106 (51)	Training-test split	27 (11)	Sensitvity: 100%, specificity: 82%	High
	Xu, Jiang, et al.	Data from China, target population unclear	COVID-19 diagnosis	Not applicable	509 (110)	Training-test split	90 (30)	Sensitivity: 87%, PPV: 81%	High
	Wang, Kang, et al.	Data from China, target population unclear	COVID-19 diagnosis	Not applicable	259 (79)	internal, other images from same individuals	Not applicable	C-index: 0.81 (0.71,0.84) sensitivity: 83%; specificity: 0.67%	High
	Ying, Zheng, et al.	Data from China, target population unclear	diagnosis of COVID-19 vs healthy controls	Not applicable	123 (61)	Training-test split	51 (27)	AUC: 0.99	High

	Ying, Zheng, et al.	Data from China, target population unclear	diagnosis of COVID-19 vs bacterial pneumonia	Not applicable	131 (61)	Training-test split	57 (27)	AUC: 0.96	High
	Shan, Gao, et al.	Data from China, COVID-19 confirmed cases	Segmentation and quantification of infection regions in lung from chest CT scans.	Not applicable	249 (not applicable)	Training-test split	300 (not applicable)	Dice similarity coefficient 91.6% *4	High
Prognosis									
	Caramelo, Ferreira, et al.	Data from China, target population unclear	Mortality (period unspecified) *5	Age; sex; presence of any comorbidity (hypertension, diabetes, cardiovascular disease, chronic respiratory disease, cancer) *5	Unknown	Not reported	Not applicable	Not reported	High
	Lu, Hu, et al.	Data from China, inpatients at admission suspected or confirmed COVID-19 case	Mortality (12 day)	Age; C-reactive protein	577 (44)	Not reported	Not applicable	Not reported	High
	Qi, Jiang, et al	Data from China, COVID-19 confirmed inpatients at admission	Hospital stay >10 days	6 features derived from CT images *6 (logistic regression model)	26 (20)	5 fold cross-validation	Not applicable	C-index: 0.92	High
	Qi, Jiang, et al	Data from China, COVID-19 confrimed inpatients at admission	Hospital stay >10 days	6 features derived from CT images *6 (random forestl)	26 (20)	5 fold cross-validation	Not applicable	C-index: 0.96	High
	Xie, Hungerford, et al.	Data from China, COVID-19 confirmed inpatients at admission	Mortality (in hospital)	Age, LDH, lymphocyte count, SPO2	299 (155)	External validation (other Chinese center)	130 (69)	C-index: 0.98 (0.96,1.00); calibration slope: 2.5 (1.7,3.7)	High

	Yan, Zhang, et al.	Data from China, inpatients suspected of COVID-19	Mortality (period unspecified)	Lactic dehydrogenase; lymphocyte count; high-sensitivity C-reactive protein	375 (174)	Temporal validation, selecting only severe cases	29 (17)	Sensitivity: 92%; PPV: 95%	High
	Yuan, Yin, et al.	Data from China, target population unclear	Mortality (period unspecified)	Clinical scorings of CT images (zone, left/right, location, attenuation, distribtion of affected parenchyma)	27 (10)	Apparent performance only	Not applicable	C-index: 0.90 (0.87, 0.93)	High

*1 Performance is given for the strongest form of validation reported. This is indicated in the column “type of validation”. When a train-test split was used, performance on the test set is reported. Apparent performance is the performance observed in the development data.

*2 Proxy events used: pneumonia (except from TB), influenza, acute bronchitis, or other specified upper respiratory infections (no COVID-19 pneumonia cases in data).

*3 The development set contains scans from Chinese patients, the testing set contained scans from Chinese cases and controls, and U.S. controls.

*4 Describes similarity between segmentation of the CT scan by a medical doctor and automated segmentation.

*5 Outcome and pedictor data were simulated.

*6 Wavelet-HLH_gldm_SmallDependenceLowGrayLevelEmphasis, wavelet-LHH_glcm_Correlation, wavelet-LHL_glszm_GrayLevelV ariance, wavelet-LLH_glszm_SizeZoneNonUniformityNormalized, wavelet-LLH_glszm_SmallAreaEmphasis, wavelet-LLH_glcm_Correlation.

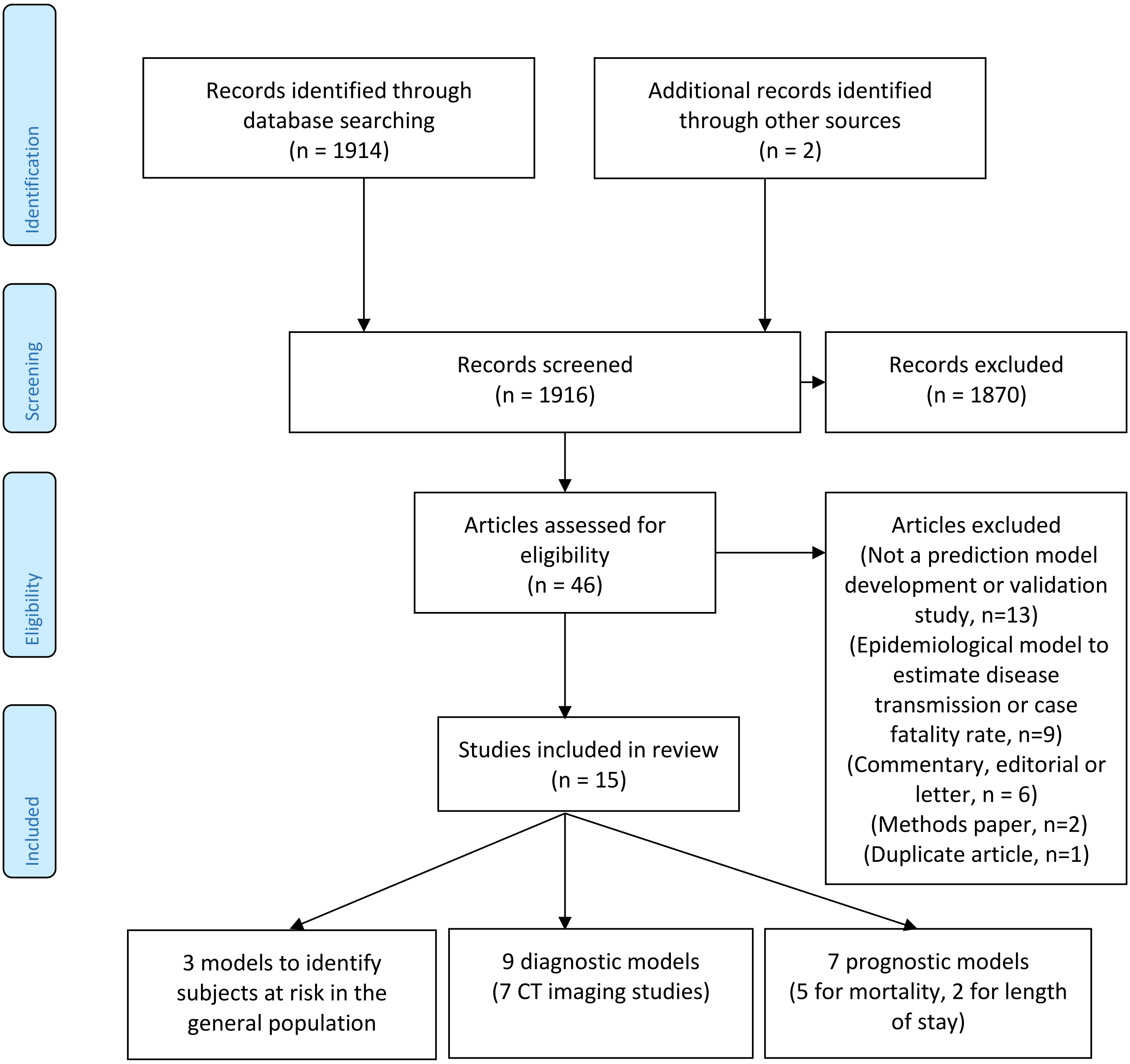
Table 2. Risk of bias assessment (using PROBAST) based on four domains across 15 studies creating prediction models for COVID-19

Authors	Risk of bias: participants	Risk of bias: predictors	Risk of bias: outcome	Risk of bias: analysis
Hospital admission in general population				
DeCaprio, Gartner, et al.	high	low	high	high
Diagnosis				
Song, Xu, et al.	high	unclear	low	high
Yu, Shao, et al.	unclear	unclear	unclear	high
Diagnostic imaging				
Gozed, Frid-Adar, et al.	unclear	unclear	high	high
Chen, Wu, et al.	high	unclear	low	high [*1]
Shan, Gao, et al.	unclear	unclear	high	high [*2]
Wang, Kang, et al.	high	unclear	low	high
Xu, Jiang, et al.	high	unclear	high	high
Ying, Zheng, et al.	unclear	unclear	low	high
Prognosis				
Caramelo, Ferreira, et al.	high	high	high	high
Lu, Hu, et al.	low	low	low	high
Qi, Jiang, et al.	unclear	low	low	high
Xie, Hungerford, et al.	low	low	low	high
Yan, Zhang, et al.	low	high	low	high
Yuan, Yin, et al.	unclear	high	low	high

[*1] Risk of bias high due to not evaluating calibration. If this criterion is not taken into account, analysis risk of bias would have been unclear.

[*2] Risk of bias high due to not evaluating calibration. If this criterion is not taken into account, analysis risk of bias would have been low.

Figure 1. PRISMA flowchart of in- and exclusions.


Identification
Screening
Eligibility
Included
Records identified through database searching (n = 1914)
Additional records identified through other sources (n = 2)
Records screened (n = 1916)
Records excluded (n = 1870)
Articles assessed for eligibility (n = 46)
Articles excluded
(Not a prediction model development or validation study, n=13)
(Epidemiological model to estimate disease transmission or case fatality rate, n=9)
(Commentary, editorial or letter, n = 6)
(Methods paper, n=2)
(Duplicate article, n=1)
Studies included in review (n = 15)
3 models to identify subjects at risk in the general population
9 diagnostic models (7 CT imaging studies)
7 prognostic models (5 for mortality, 2 for length of stay)