Review of chest imaging findings in COVID-19

Weronika Hryniewska

Cele prezentacji

- Jak zrobić dobre review?
- Jaki jest status prac nad obrazami medycznymi z Covid-19?
- Co było najbardziej wartościowego w opublikowanych review?
- Co można by poprawić w opublikowanych review?

Jak zrobić dobre review?

Wyszukiwanie artykułów

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)

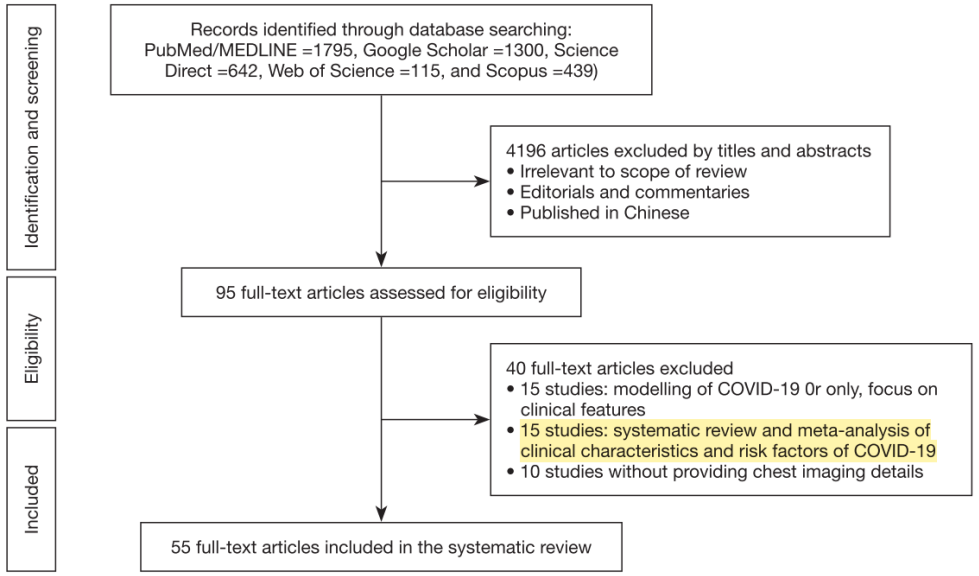
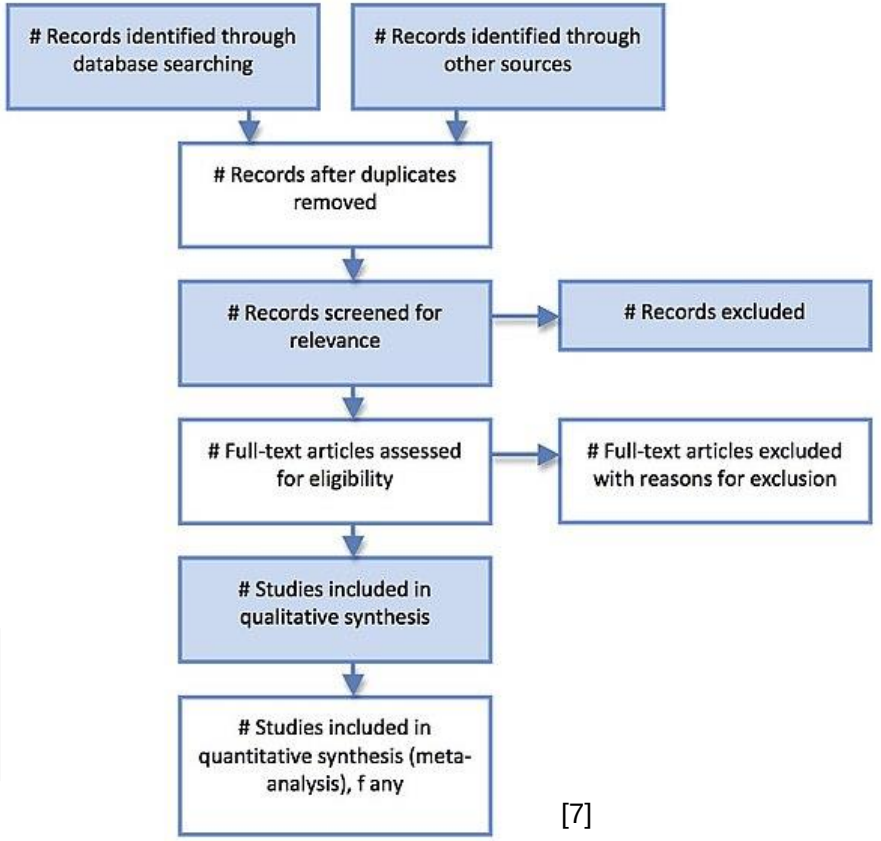


Figure 1 Flow chart showing the selection process of identifying studies that met the inclusion criteria. [4]



[7]

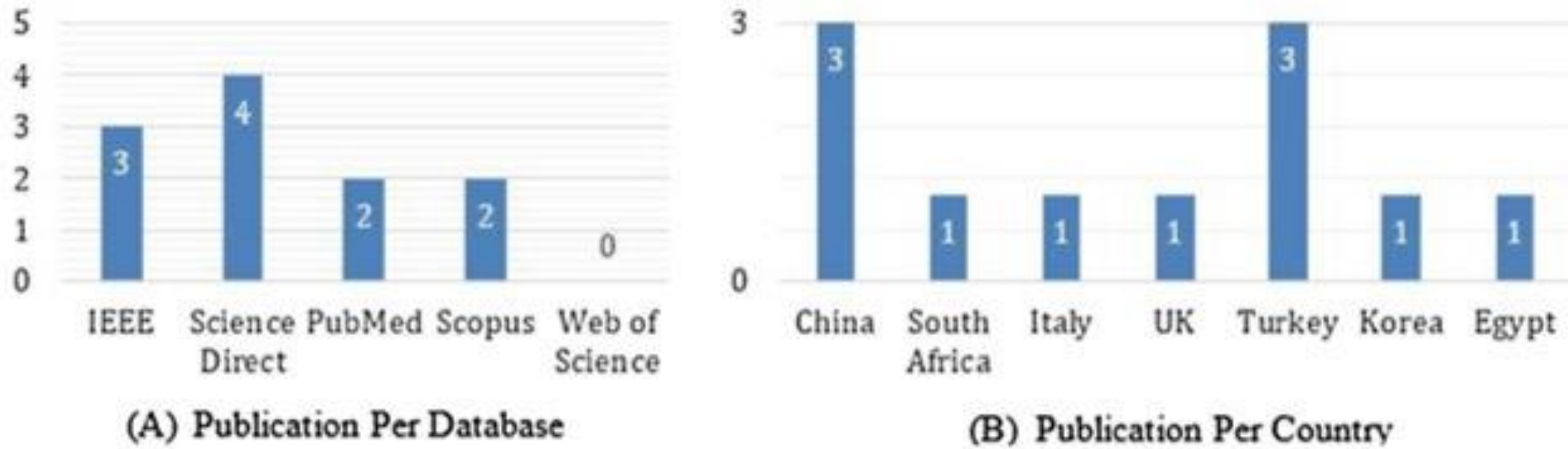


Fig. 2. Statistics of the included studies by databases and countries. [1]

Jakie artykuły brać pod uwagę?

- All studies were considered, regardless of language or publication status (preprint or peer reviewed articles; updates of preprints will only be included and reassessed in future updates after publication in a peer reviewed journal).[5]
- Several studies used open Github or Kaggle data repositories (version or date of access often unclear), and so it was unclear how much these datasets overlapped across. [5]

Grupowanie rozwiązań zaproponowane w artykułach

- Classification of COVID-19 from non-COVID-19
- Classification of COVID-19 from other pneumonia
- Severity assessment of COVID-19:
  - Tang et al. proposed an RF-based model for COVID-19 severity assessment (non-severe or severe). A deep learning method VB-Net is adopted to divide the lung into anatomical sub-regions (e.g., lobes and segments), based on which infection volumes and ratios of each anatomical sub-region are calculated and used as quantitative features to train a RF model. [3]

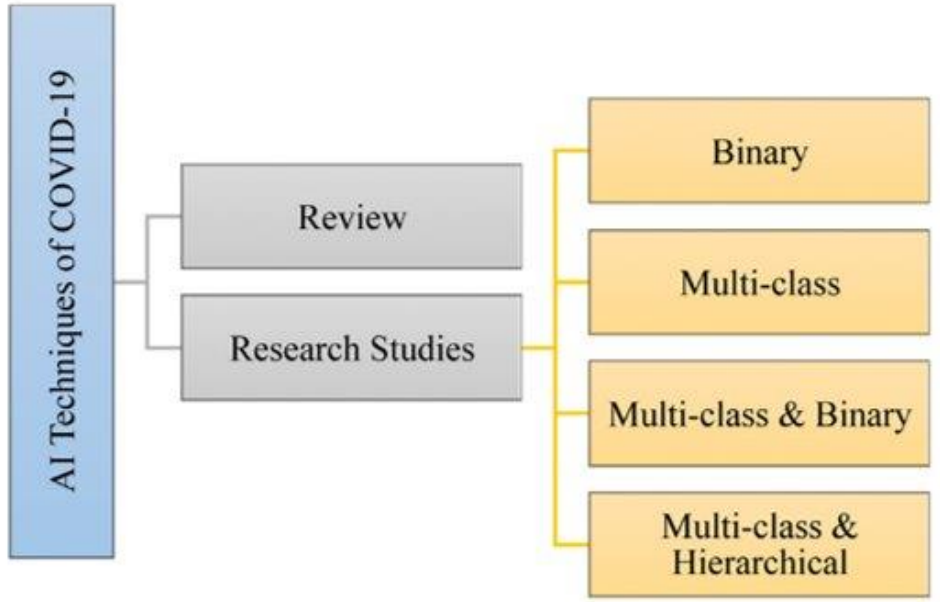


Fig. 3. Taxonomy of research literature on AI techniques used in the detection and classification of COVID-19 medical images. [1]

Table 1 [1]
Summary of the perspectives of works described in research cluster studies.

Ref.	Type of datasets		AI techniques		Case study
	Primary data	Secondary data	Traditional machine learning techniques	Deep learning techniques	
[31]	χ	χ	χ	√	CT scan
[32]	√	χ	√	χ	CT scan
[36]	√	√	√	√	CT scan
[37]	√	√	√	√	X-ray
[7]	χ	√	χ	√	X-ray
[33]	χ	√	χ	√	X-ray
[34]	χ	√	χ	√	X-ray
[40]	χ	√	√	χ	X-ray
[35]	χ	√	√	√	X-ray
[39]	√	√	√	√	X-ray

[31] Li D, et al. False-negative results of real-time reverse-transcriptase poly- merase chain reaction for severe acute respiratory syndrome coronavirus 2: role of deep-learning-based CT diagnosis and insights from two cases. Korean J Radiol 2020;21(4):505–8.
[32] Wallis LA. COVID-19 severity scoring tool for low resourced settings. Afr J Emerg Med 2020, http://dx.doi.org/10.1016/j.afjem.2020.03.002. In press.
[36] Laghi A. Cautions about radiologic diagnosis of COVID-19 infection driven by artificial intelligence. Lancet Digit Health 2020;2(5):e225.
[37] McCall B. COVID-19 and artificial intelligence: protecting health-care workers and curbing the spread. Lancet Digit Health 2020;2(4):e166–7.
[7] Ozturk T, et al. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput Biol Med 2020;103792.
[33] Ucar F, Korkmaz D. COVIDiagnosis-net: deep Bayes-SqueezeNet based diag- nostic of the coronavirus disease 2019 (COVID-19) from X-ray images. Med Hypotheses 2020;109761.
[34] Oh Y, Park S, Ye JC. Deep learning COVID-19 features on CXR using limited training data sets. arXiv preprint arXiv:2004.05758 2020, http://dx.doi.org/ 10.1109/TMI.2020.2993291, 1–1. In press.
[40] Abdel-Basset M, Mohamed R, Elhoseny M, Chakrabortty RK, Ryan M. A hybrid COVID-19 detection model using an improved marine predators algorithm and a ranking-based diversity reduction strategy. IEEE Access 2020;8:79521–40.
[35] To˘gac¸ ar M, Ergen B, Cömert Z. COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches. Comput Biol Med 2020;103805.
[39] Pereira RM, Bertolini D, Teixeira LO, Silla Jr CN, Costa YMG. COVID-19 identi- fication in chest X-ray images on flat and hierarchical classification scenarios. arXiv 2020;105532.

Publication	ML/AI method	Types of data	No of patients	Validation method	Sample size	Accuracy
Ardakani, A. A *et al.*, [28]	Deep Convolutional Neural Network ResNet-101	Clinical, Mamographic	1020, 86	Holdout	1020 CT images of 108 volume of patients with laboratory confirmed Covid-19, 86 CT images of viral and atypical pneumonia patients,	Accuracy: 99.51% Specificity: 99.02%
Ozturk, T. *et al.*, [29]	Convolutional Neural Network DarkCovidNet Architecture	Clinical, Mamographic	127, 43 f, 82 m 500, 500	Cross-validation	127 X-ray images with 43 female and 82 male positive cases 500 no-findings and pneumonia cases of 500	Accuracy: 98.08% on Binary classes Accuracy: 87.02% on Multi-classes
Sun, L *et al.*, [30]	Support Vector Machine	Clinical, laboratory features, Demographics	336, 220	Holdout	336 infected patients with PCR kit, 26 severe/critical cases and 310 non-serious cases and with another related disease79 hypertension, 29diabetes, 17 coronary disease and 7 having history of tuberculosis	Accuracy: 77.5% Specificity: 78.4% AUROC reaches 0.99 training and 0.98 testing dataset
Wu, J. *et al.*, [31]	Random forest Algorithm	Clinical, Demographics	253, 169, 49,24	Cross-validataion	Total of 253 samples from 169 patients suspected with Covid-19 collected from multiple sources. Clinical blood test of 49 patients derived from commercial clinic center. 24 samples infected patient with Covid-19	Accuracy: 95.95% Specificity: 96.95%

TABLE II
RELATED STUDIES WITH MEDICAL IMAGES FOR AI-ASSISTED DIAGNOSIS OF COVID-19

Literature	Modality	Subjects	Task	Method	Result
Ghoshal *et al.* [72]	X-Ray	70 COVID-19 Others (# of subjects not available)	Classification: COVID-19/ Others	CNN	92.9% (Acc.)
Narin *et al.* [9]	X-Ray	50 COVID-19 50 Normal	Classification: COVID-19/ Normal	ResNet50	98.0% (Acc.)
Zhang et al. [74]	X-Ray	70 COVID-19 1008 Others	Classification: COVID-19/ Others	ResNet	96.0% (Sens.) 70.7% (Spec.) 0.952 (AUC)
Wang *et al.* [11]	X-Ray	45 COVID-19 931 Bac. Pneu. 660 Vir. Pneu. 1203 Normal	Classification: COVID-19/ Bac. Pneu./ Vir. Pneu./ Normal	CNN	83.5% (Acc.)
Chen *et al.* [56]	CT	51 COVID-19 55 Others	Classification: COVID-19/ Others	UNet++	95.2% (Acc.) 100% (Sens.) 93.6% (Spec.)
Zheng *et al.* [50]	CT	313 COVID-19 229 Others	Classification: COVID-19/ Others	U-Net CNN	90.7% (Sens.) 91.1% (Spec.) 0.959 (AUC)
Jin *et al.* [69]	CT	496 COVID-19 1385 Others	Classification: COVID-19/ Others	CNN	94.1% (Sens.) 95.5% (Spec.)
Jin *et al.* [57]	CT	723 COVID-19 413 Others	Classification: COVID-19/ Others	UNet++ CNN	97.4% (Sens.) 92.2% (Spec.)
Wang *et al.* [75]	CT	44 COVID-19 55 Vir. Pneu.	Classification: COVID-19/ Vir. Pneu.	CNN	82.9% (Acc.)
Ying *et al.* [70]	CT	88 COVID-19 100 Bac. Pneu. 86 Normal	Classification: COVID-19/ Bac. Pneu./ Normal	ResNet-50	86.0% (Acc.)
Xu *et al.* [76]	CT	219 COVID-19 224 Influ.-A 175 Normal	Classification: COVID-19/ Influ.-A/ Normal	CNN	86.7% (Acc.)
Li *et al.* [55]	CT	468 COVID-19 1551 CAP 1445 Non-pneu.	Classification: COVID-19/ CAP/ Non-pneu.	ResNet-50	90.0% (Sens.) 96.0% (Spec.)
Shi *et al.* [77]	CT	1658 COVID-19 1027 CAP	Classification: COVID-19/CAP	RF	87.9% (Acc.) 90.7% (Sens.) 83.3% (Spec.)
Tang *et al.* [78]	CT	176 COVID-19	Severity assessment	RF	87.5% (Acc.) 93.3% (TPR) 74.5% (TNR)

Bac. Pneu.: Bacterial pneumonia; Vir. Pneu.: Viral pneumonia; Influ.-A: Influenza-A; Non-pneu.: Non- pneumonia

[3]

Study; setting; and outcome	Predictors in final model	Predictive performance on validation				Overall risk of bias using PROBAST
		Sample size: total No of participants for model development set (No with outcome)	Type of validation*	Sample size: total No of participants for model validation (No with outcome)	Performance* (C index, sensitivity (%), specificity (%), PPV/NPV (%), calibration slope, other (95% CI, if reported))	
Update 1						
Abbas et al[47]; data from repositories (origin unspecified), target population unclear; covid-19 diagnosis	Not applicable	137 (unknown)	Training test split	59 (unknown)	C index 0.94, sensitivity 98, specificity 92	High
Apostolopoulos et al[48]; data from repositories (US, Italy); patients with suspected covid-19; covid-19 diagnosis	Not applicable	1427 (224)	10-fold cross validation	Not applicable	Sensitivity 99, specificity 97	High
Bukhari et al[49]; data from Canada and US; patients with suspected covid-19; covid-19 diagnosis	Not applicable	223 (unknown)	Training test split	61 (17)	Sensitivity 98, PPV 91	High
Chaganti et al[50]; data from Canada, US, and European countries; patients with suspected covid-19; percentage lung opacity	Not applicable	631 (not applicable)	Training test split	100 (not applicable)	Correlation§§ 0.98	High
Chaganti et al[50]; data from Canada, US, and European countries; patients with suspected covid-19; percentage high lung opacity	Not applicable	631 (not applicable)	Training test split	100 (not applicable)	Correlation§§ 0.98	High
Chaganti et al[50]; data from Canada, US, and European countries; patients with suspected covid-19; severity score	Not applicable	631 (not applicable)	Training test split	100 (not applicable)	Correlation§§ 0.97	High
Chaganti et al[50]; data from Canada, US, and European countries; patients with suspected covid-19; lung opacity score	Not applicable	631 (not applicable)	Training test split	100 (not applicable)	Correlation§§ 0.97	High
Chowdhury et al[39]; data from repositories (Italy and other unspecified countries), target population unclear; covid-19 *v* "normal"	Not applicable	Unknown	Fivefold cross validation	Not applicable	C index 0.99	High
Chowdhury et al[39]; data from repositories (Italy and other unspecified countries), target population unclear; covid-19 *v* "normal" and viral pneumonia	Not applicable	Unknown	Fivefold cross validation	Not applicable	C index 0.98	High
Chowdhury et al[39]; data from repositories (Italy and other unspecified countries), target population unclear; covid-19 *v* "normal"	Not applicable	Unknown	Fivefold cross validation	Not applicable	C index 0.998	High
Chowdhury et al[39]; data from repositories (Italy and other unspecified countries), target population unclear; covid-19 *v* "normal" and viral pneumonia	Not applicable	Unknown	Fivefold cross validation	Not applicable	C index 0.99	High
Fu et al[51]; data from China, target population unclear; covid-19 diagnosis	Not applicable	610 (100)	External validation	309 (50)	C index 0.99, sensitivity 97, specificity 99	High
Gozes et al[52]; data from China, people with suspected covid-19; covid-19 diagnosis	Not applicable	50 (unknown)	External validation	199 (109)	C index 0.95 (0.91 to 0.99)	High
Imran et al[53]; data from unspecified source, target population unclear; covid-19 diagnosis	Not applicable	357 (48)	Twofold cross validation	Not applicable	Sensitivity 90, specificity 81	High
Li et al[54]; data from China, inpatients with confirmed covid-19; severe and critical covid-19	Severity score based on CT scans	Not applicable	External validation of existing score	78 (not applicable)	C index 0.92 (0.84 to 0.99)	High
Li et al[55]; data from unknown origin, patients with suspected covid-19; covid-19	Not applicable	360 (120)	Training test split	135 (45)	C index 0.97	High
Hassanien et al[56]; data from repositories (origin unspecified), people with suspected covid-19; covid-19 diagnosis	Not applicable	Unknown	Training test split	Unknown	Sensitivity 95, specificity 100	High
Tang et al[57]; data from China, patients with confirmed covid-19; covid-19 severe *v* non-severe	Not applicable	176 (55)	Threefold cross validation	Not applicable	C index 0.91, sensitivity 93, specificity 75	High

[5]

CHARMS

the CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies [8]

Item	Comments and examples
1. Prognostic versus diagnostic prediction model	Define whether the aim is to review models to predict:
	• Future events: prognostic prediction models
	• Current (disease) status: diagnostic prediction models
2. Intended scope of the review	Define intended scope of the review and intended purpose of the models reviewed in it. Examples:
	• Models to inform physicians' therapeutic decision making
	• Models to inform referral to or withholding from invasive diagnostic testing
3. Type of prediction modelling studies (see also Box 1)	Define the type of prediction modelling studies to include. Examples of study types (Box 1):
	• Prediction model development without external validation in independent data
	• Prediction model development with external validation in independent data
	• External model validation, possibly with model updating
4. Target population to whom the prediction model applies	Define the target population relevant to the review scope. Examples:
	• Women with diagnosed breast cancer
	• Healthy adult men in the general population
5. Outcome to be predicted	Define the outcome of interest to be predicted:
	• Specific future event, such as a fatal or non-fatal coronary heart disease
	• Specific diagnostic target disease, such as presence of lung embolism
6. Time span of prediction	Define over what specific time period the outcome is predicted (prognostic models only). Example:
	• Event within a specific time interval, such as event within 3 months, 1 year, or 10 years
7. Intended moment of using the model	The systematic review may focus on models to be used at a specific moment in time. Examples:
	• Models to be used at the moment of diagnosis of a particular disease
	• Models to be used preoperatively to predict the risk of postoperative complications
	• Models to be used in asymptomatic adults to detect undiagnosed type 2 diabetes mellitus

Domain	Key items	General	Applicability	Risk of bias
Source of data	• Source of data (e.g., cohort, case-control, randomised trial participants, or registry data)		X	X
Participants	• Participant eligibility and recruitment method (e.g., consecutive participants, location, number of centres, setting, inclusion and exclusion criteria)	X	X	
	• Participant description	X	X	
	• Details of treatments received, if relevant		X	X
	• Study dates	X	X	
Outcome(s) to be predicted	• Definition and method for measurement of outcome		X	X
	• Was the same outcome definition (and method for measurement) used in all patients?			X
	• Type of outcome (e.g., single or combined endpoints)	X	X	
	• Was the outcome assessed without knowledge of the candidate predictors (i.e., blinded)?			X
	• Were candidate predictors part of the outcome (e.g., in panel or consensus diagnosis)?			X
	• Time of outcome occurrence or summary of duration of follow-up		X	
Candidate predictors (or index tests)	• Number and type of predictors (e.g., demographics, patient history, physical examination, additional testing, disease characteristics)	X		
	• Definition and method for measurement of candidate predictors		X	X
	• Timing of predictor measurement (e.g., at patient presentation, at diagnosis, at treatment initiation)		X	
	• Were predictors assessed blinded for outcome, and for each other (if relevant)?			X
	• Handling of predictors in the modelling (e.g., continuous, linear, non-linear transformations or categorised)			X
Sample size	• Number of participants and number of outcomes/events	X		
	• Number of outcomes/events in relation to the number of candidate predictors (Events Per Variable)			X
Missing data	• Number of participants with any missing value (include predictors and outcomes)	X		X
	• Number of participants with missing data for each predictor			X
	• Handling of missing data (e.g., complete-case analysis, imputation, or other methods)			X
Model development	• Modelling method (e.g., logistic, survival, neural networks, or machine learning techniques)	X		
	• Modelling assumptions satisfied			X
	• Method for selection of predictors for inclusion in multivariable modelling (e.g., all candidate predictors, pre-selection based on unadjusted association with the outcome)			X
	• Method for selection of predictors during multivariable modelling (e.g., full model approach, backward or forward selection) and criteria used (e.g., p-value, Akaike Information Criterion)			X
	• Shrinkage of predictor weights or regression coefficients (e.g., no shrinkage, uniform shrinkage, penalized estimation)		X	X
Model performance	• Calibration (calibration plot, calibration slope, Hosmer-Lemeshow test) and Discrimination (C-statistic, D-statistic, log-rank) measures with confidence intervals		X	
	• Classification measures (e.g., sensitivity, specificity, predictive values, net reclassification improvement) and whether a priori cut points were used			X
Model evaluation	• Method used for testing model performance: development dataset only (random split of data, resampling methods, e.g., bootstrap or cross-validation, none) or separate external validation (e.g., temporal, geographical, different setting, different investigators)			X
	• In case of poor validation, whether model was adjusted or updated (e.g., intercept recalibrated, predictor effects adjusted, or new predictors added)		X	X
Results	• Final and other multivariable models (e.g., basic, extended, simplified) presented, including predictor weights or regression coefficients, intercept, baseline survival, model performance measures (with standard errors or confidence intervals)	X	X	
	• Any alternative presentation of the final prediction models, e.g., sum score, nomogram, score chart, predictions for specific risk subgroups with performance	X	X	
	• Comparison of the distribution of predictors (including missing data) for development and validation datasets			X
Interpretation and Discussion	• Interpretation of presented models (confirmatory, i.e., model useful for practice versus exploratory, i.e., more research needed)	X	X	
	• Comparison with other studies, discussion of generalizability, strengths and limitations	X	X	

Jak napisać dobry artykuł?

TRIPOD

(Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) [6]

TRIPOD Checklist: Prediction Model Development

Section/Topic	Item	Checklist Item	Page
Title and abstract			
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	
Introduction			
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	
Methods			
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	
	5b	Describe eligibility criteria for participants.	
	5c	Give details of treatments received, if relevant.	
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	
	6b	Report any actions to blind assessment of the outcome to be predicted.	
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	
Sample size	8	Explain how the study size was arrived at.	
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	
Statistical analysis methods	10a	Describe how predictors were handled in the analyses.	
	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	
Risk groups	11	Provide details on how risk groups were created, if done.	
Results			
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	
Model development	14a	Specify the number of participants and outcome events in each analysis.	
	14b	If done, report the unadjusted association between each candidate predictor and outcome.	
Model specification	15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	
	15b	Explain how to the use the prediction model.	
Model performance	16	Report performance measures (with CIs) for the prediction model.	
Discussion			
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	
Interpretation	19b	Give an overall interpretation of the results, considering objectives, limitations, and results from similar studies, and other relevant evidence.	
Implications	20	Discuss the potential clinical use of the model and implications for future research.	
Other information			
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	
Funding	22	Give the source of funding and the role of the funders for the present study.	

We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.

Czy Twój zbiór danych jest wystarczająco duży?

- Create bigger dataset than data from China, Italy, and international registries [5]
- Majority of the current reports is dominated by case studies documenting individual institution’s experience of diagnosing and treatment COVID-19 patients.
  Only two studies in this review included more than 1,000 cases, whereas 72,2% of the studies included <100 patients. [4]

Napisz dlaczego te dane są sensowne

- Chest radiographs (X-rays) and chest computed tomography (CT) scans can assist and reveal anomalies indicative of different lung diseases, including COVID-19. CT scan and X-ray tests could be utilised as a primary detection tool to evaluate the severity of COVID-19, monitor the emergency case of infected patients and predict COVID-19 progression. [1]

- Use of CT as a first-line diagnostic or screening tool in COVID-19 is not recommended. Although CXR resembles CT findings (in 28 patients) in these common abnormal lung findings, it is less sensitive than CT in detecting the abnormalities. CXR plays a role in the identification and detection of abnormal lung changes, while chest CT could serve as a complementary role in evaluating potential complications, disease severity and progression rather than a routine diagnostic approach. [4]

- A recent study reported that X-ray shows normal in early or mild disease. In particular, abnormal chest radiographs are found in 69% of the patients at the initial time of admission, and in 80% of the patients some time after during hospitalization [3]

Obróbka wstępna danych

- Their results indicated that pre-processing for normalisation of data helped in the processing of cross-database and significantly improved the accuracy of segmentation (Jaccard similarity coefficients from 0.932 to 0.943, $p < 0.001$). [1]

- often lacked clear information on
the preprocessing steps (eg. cropping of images) [5]

Zweryfikuj możliwość uprzedzeń

PROBAST

Prediction model Risk Of Bias ASsessment Tool [9]

Table 1. Four Steps in PROBAST

Step	Task	When to Complete
1	Specify your systematic review question(s)	Once per systematic review
2	Classify the type of prediction model evaluation	Once for each model of interest in each publication being assessed, for each relevant outcome
3	Assess risk of bias and applicability (per domain)	Once for each development and validation of each distinct prediction model in a publication
4	Overall judgment of risk of bias and applicability	Once for each development and validation of each distinct prediction model in a publication

Table 2. PROBAST: Summary of Step 3—Assessment of Risk of Bias and Concerns Regarding Applicability*

1. Participants	2. Predictors	3. Outcome	4. Analysis
Signaling questions			
1.1. Were appropriate data sources used, e.g., cohort, RCT, or nested case-control study data?	2.1. Were predictors defined and assessed in a similar way for all participants?	3.1. Was the outcome determined appropriately?	4.1. Were there a reasonable number of participants with the outcome?
1.2. Were all inclusions and exclusions of participants appropriate?	2.2. Were predictor assessments made without knowledge of outcome data?	3.2. Was a prespecified or standard outcome definition used?	4.2. Were continuous and categorical predictors handled appropriately?
-	2.3. Are all predictors available at the time the model is intended to be used?	3.3. Were predictors excluded from the outcome definition?	4.3. Were all enrolled participants included in the analysis?
-	-	3.4. Was the outcome defined and determined in a similar way for all participants?	4.4. Were participants with missing data handled appropriately?
-	-	3.5. Was the outcome determined without knowledge of predictor information?	4.5. Was selection of predictors based on univariable analysis avoided?†
-	-	3.6. Was the time interval between predictor assessment and outcome determination appropriate?	4.6. Were complexities in the data (e.g., censoring, competing risks, sampling of control participants) accounted for appropriately?
-	-	-	4.7. Were relevant model performance measures evaluated appropriately?
-	-	-	4.8. Were model overfitting, underfitting, and optimism in model performance accounted for?†
-	-	-	4.9. Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?†
ROB			
Selection of participants	Predictors or their assessment	Outcome or its determination	Analysis
Applicability			
Included participants or setting does not match the review question	Definition, assessment, or timing of predictors does not match the review question	Its definition, timing, or determination does not match the review question	-

RCT = randomized controlled trial; ROB = risk of bias.
* For further details, please see the explanation and elaboration document (27), available at Annals.org, and www.probast.org. Signaling questions are answered as yes, probably yes, probably no, no, or no information. ROB and concerns for applicability are rated as low, high, or unclear.
† Development studies only.

Walidacja

- available form of validation in order of strength [5]:
  - **external (evaluation in an independent database),**
  - internal (bootstrap validation, cross validation, random training test splits, temporal splits),
  - apparent (evaluation by using exactly the same data used for development)

Proces oceny i techniki klasyfikacji porównawczej [1]

Challenge of multiple evaluation criteria (Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag 2009;45(4):427–37.)

- 6 criteria for binary classification
- 8 criteria for multi-class classification
- 4 criteria for multi-labelled classification
- 6 criteria for hierarchical classification

Challenge of criteria trade-off

- The reliability should possess a high rate.
- Time complexity to conduct the output that also need to be low.
- A new approach for the evaluation that handles all conflict criteria and data problems should emerge, and this method should be flexible.

Challenge of criteria importance

- The COVID-19 classification technique simultaneously considers multiple criteria and then assign a suitable weight for all evaluation criteria.
- The experts who are in charge of assigning a score for the COVID-19 classification techniques could assign more weights to different features aside from the ones that acquire less interest than any other criteria. By contrast, experts who aim to make use of benchmarking method in order to address such problems would consider different criteria as the most significant ones.

Table 2
Measures for binary classification using the notation of Table 1.

Measure	Formula	Evaluation focus
Accuracy	$\frac{tp+tn}{tp+fn+fp+tn}$	Overall effectiveness of a classifier
Precision	$\frac{tp}{tp+fp}$	Class agreement of the data labels with the positive labels given by the classifier
Recall (*Sensitivity*)	$\frac{tp}{tp+fn}$	Effectiveness of a classifier to identify positive labels
Fscore	$\frac{(\beta^2+1)tp}{(\beta^2+1)tp+\beta^2 fn+fp}$	Relations between data's positive labels and those given by a classifier
Specificity	$\frac{tn}{fp+tn}$	How effectively a classifier identifies negative labels
AUC	$\frac{1}{2}\left(\frac{tp}{tp+fn}+\frac{tn}{tn+fp}\right)$	Classifier's ability to avoid false classification

Table 3
Measures for multi-class classification based on a generalization of the measures of Table 1 for many classes C_i: tp_i are true positive for C_i, and fp_i – false positive, fn_i – false negative, and tn_i – true negative counts respectively. μ and M indices represent micro- and macro-averaging.

Measure	Formula	Evaluation focus
Average Accuracy	$\frac{\sum_{i=1}^{l}\frac{tp_i+tn_i}{tp_i+fn_i+fp_i+tn_i}}{l}$	The average per-class effectiveness of a classifier
Error Rate	$\frac{\sum_{i=1}^{l}\frac{fp_i+fn_i}{tp_i+fn_i+fp_i+tn_i}}{l}$	The average per-class classification error
$Precision_\mu$	$\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i+fp_i)}$	Agreement of the data class labels with those of a classifiers if calculated from sums of per-text decisions
$Recall_\mu$	$\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i+fn_i)}$	Effectiveness of a classifier to identify class labels if calculated from sums of per-text decisions
$Fscore_\mu$	$\frac{(\beta^2+1)Precision_\mu Recall_\mu}{\beta^2 Precision_\mu+Recall_\mu}$	Relations between data's positive labels and those given by a classifier based on sums of per-text decisions
$Precision_M$	$\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fp_i}}{l}$	An average per-class agreement of the data class labels with those of a classifiers
$Recall_M$	$\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fn_i}}{l}$	An average per-class effectiveness of a classifier to identify class labels
$Fscore_M$	$\frac{(\beta^2+1)Precision_M Recall_M}{\beta^2 Precision_M+Recall_M}$	Relations between data's positive labels and those given by a classifier based on a per-class average

[1]

Table 4
Measures for multi-topic classification; I is the indicator function; $L_i = L_i[1], \ldots, L_i[l]$ denotes a set of class labels for x_i, $L_i[j] = 1$ if C_j is present among the labels and 0, otherwise; L_i^c are labels given by a classifier, L_i^d are the data labels.

Measure	Formula	Evaluation focus
Exact Match Ratio	$\frac{\sum_{i=1}^{n} I(L_i^d = L_i^d)}{n}$	The average per-text exact classification
Labelling Fscore	$\frac{\sum_{i=1}^{n} \frac{2\sum_{j=1}^{l} L_i^c[j] L_i^d[j]}{\sum_{j=1}^{l} (L_i^c[j] + L_i^d[j])}}{n}$	The average per-text classification with partial matches
Retrieval Fscore	$\frac{\sum_{j=1}^{l} \frac{2\sum_{i=1}^{n} L_i^c[j] L_i^d[j]}{\sum_{i=1}^{n} (L_i^c[j] + L_i^d[j])}}{l}$	The average per-class classification with partial matches
Hamming Loss	$\frac{\sum_{i=1}^{n} \sum_{j=1}^{l} I(L_i^c[j] \neq L_i^d[j])}{nl}$	The average per-example per-class total error

Table 5
Measures for hierarchical classification: $C_\downarrow$ means subclasses of class C, $C_\downarrow^c$ denotes subclasses assigned by a classifier; $C_\downarrow^d$ – data class labels; similar notations apply to superclasses, which are denoted by $C_\uparrow$.

Measure	Formula	Evaluation focus
$Precision_\downarrow$	$\frac{\lvert C_\downarrow^c \cap C_\downarrow^d \rvert}{\lvert C_\downarrow^c \rvert}$	Positive agreement on subclass labels w.r.t. the subclass labels given by a classifier
$Recall_\downarrow$	$\frac{\lvert C_\downarrow^c \cap C_\downarrow^d \rvert}{\lvert C_\downarrow^d \rvert}$	Positive agreement on subclass labels w.r.t. the subclass labels given by data
$Fscore_\downarrow$	$\frac{(\beta^2+1) Precision_\downarrow Recall_\downarrow}{\beta^2 Precision_\downarrow + Recall_\downarrow}$	Relations between data's positive subclass labels and those given by a classifier
$Precision_\uparrow$	$\frac{\lvert C_\uparrow^c \cap C_\uparrow^d \rvert}{\lvert C_\uparrow^c \rvert}$	Positive agreement on superclass labels w.r.t. the superclass labels given by a classifier
$Recall_\uparrow$	$\frac{\lvert C_\uparrow^c \cap C_\uparrow^d \rvert}{\lvert C_\uparrow^d \rvert}$	Positive agreement on superclass labels w.r.t. the superclass labels given by data
$Fscore_\uparrow$	$\frac{(\beta^2+1) Precision_\uparrow Precision_\uparrow}{\beta^2 Precision_\uparrow + Precision_\uparrow}$	Relations between data's positive superclass labels and those given by a classifier

[1]

Napisz o temacie, którego jeszcze nie było

- Review of XAI in Covid-19 articles
- No study has provided multi-labelled classification for the detection of COVID-19 medical images. [1]
- ‘Which classification technique is appropriate for such purpose?’ [1]
- Hybrid solution
  - Majority of the literature aimed to investigate hybrid AI techniques by combining deep learning and traditional machine learning. [1]
  - It is suggested to come up with a hybrid classification method applying more potential algorithm on multi-database or hybrid-database consisting of clinical, mammographic, and demographic data, as each type of data has a significant factor that could represent the true identity of the infected patients and deployment of the application in the real world. [2]

Bibliografia

[1] O. S. Albahri *et al.*, “Systematic Review of Artificial Intelligence Techniques in the Detection and Classification of COVID-19 Medical Images in Terms of Evaluation and Benchmarking: Taxonomy Analysis, Challenges, Future Solutions and Methodological Aspects,” *J. Infect. Public Health*, no. June, 2020, doi: 10.1016/j.jiph.2020.06.028.

[2] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, “Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review,” *Chaos, Solitons and Fractals*, vol. 139, 2020, doi: 10.1016/j.chaos.2020.110059.

[3] F. Shi *et al.*, “Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation and Diagnosis for COVID-19,” *IEEE Rev. Biomed. Eng.*, pp. 1–11, 2020, doi: 10.1109/RBME.2020.2987975.

[4] Z. Sun, N. Zhang, Y. Li, and X. Xu, “A systematic review of chest imaging findings in COVID-19,” *Quant. Imaging Med. Surg.*, vol. 10, no. 5, pp. 1058–1079, 2020, doi: 10.21037/QIMS-20-564.

[5] L. Wynants *et al.*, “Prediction models for diagnosis and prognosis of covid-19 infection: Systematic review and critical appraisal,” *BMJ*, vol. 369, 2020, doi: 10.1136/bmj.m1328.

Bibliografia c.d.

[6] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, “Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement,” *Eur. Urol.*, vol. 67, no. 6, pp. 1142–1151, 2015, doi: 10.1016/j.eururo.2014.11.025.

[7] A. Liberati *et al.*, “The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration,” *PLoS Med.*, vol. 6, no. 7, 2009, doi: 10.1371/journal.pmed.1000100.

[8] K. G. M. Moons *et al.*, “Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist,” *PLoS Med.*, vol. 11, no. 10, 2014, doi: 10.1371/journal.pmed.1001744.

[9] R. F. Wolff *et al.*, “PROBAST: A tool to assess the risk of bias and applicability of prediction model studies,” *Ann. Intern. Med.*, vol. 170, no. 1, pp. 51–58, 2019, doi: 10.7326/M18-1376.