Prediction Models for Mortality in Patients With Moderate to Severe ARDS Treated in the ICU

A Systematic Review and Meta-Analysis

Katrijn Daenen, MD; Sara C. M. Stoof, MD, PhD; Hugo van Willigen, MD; Anders Boyd, PhD; Virgil A. S. H. Dalm, MD, PhD; Diederik A. M. P. J. Gommers, MD, PhD; Eric C. M. van Gorp, MD, PhD; Abraham J. Valkenburg, MD, PhD; Henrik Endeman, MD, PhD; and Jilske A. Huijben, MD, PhD

BACKGROUND: Mortality prediction models have been developed for patients in the ICU, but infrequently are targeted for specific conditions. Because ARDS is characterized by high morbidity and mortality, ARDS-specific models for outcome prediction could be valuable for informing patients and relatives, for clinical decision-making, for targeted interventions, and for research.

RESEARCH QUESTION: What are the available prediction models for moderate to severe ARDS and what is their capacity to predict mortality?

STUDY DESIGN AND METHODS: In this systematic review and meta-analysis, we searched for eligible studies in PubMed MEDLINE, Embase, PsycINFO, Web of Science, Scopus, CINAHL, Cochrane Library, and Google Scholar databases up to March 11, 2024. We included studies that developed or validated multivariable prediction models for mortality in moderate to severe ARDS, applied within 24 hours after ICU admission. Calibration, discrimination, and clinical usefulness were summarized across models. The pooled area under the receiving operating characteristic curve (AUC) was calculated with random effects models both overall and in subgroups of models and study type (development or validation). Heterogeneity was evaluated using the I^2 statistic.

RESULTS: Of the 7455 screened articles, 14 were included, evaluating 20 unique models. Discrimination was reported for all models, whereas calibration was reported in 16 models. The pooled AUC was 0.782 (95% CI, 0.748-0.817) with an I^2 of 99.5% ($P < .0001$). In subgroup analysis, the pooled AUC for the Sequential Organ Failure Assessment (SOFA) score was 0.802 (95% CI, 0.719-0.885), the age, plateau, and Pao_2 to Fio_2 ratio score was 0.724 (95% CI, 0.643-0.805), the Acute Physiology and Chronic Health Evaluation (APACHE) II score was 0.667 (95% CI, 0.613-0.721), and all other scores were 0.813 (95% CI, 0.774-0.852; $P = .0001$ for subgroup differences). The pooled AUC was higher for derivation vs validation studies (0.816 [95% CI, 0.760-0.872] vs 0.767 [95% CI, 0.725-0.809]; $P = .17$ for subgroup differences).

INTERPRETATION: Substantial variability in discrimination exists across the included models, with calibration frequently unreported. Although models developed specifically for this patient population demonstrate superior performance, general disease severity models like APACHE and SOFA are validated more extensively. Presently, no extensively validated prediction model exists showing good discrimination and calibration for moderate to severe ARDS.

CLINICAL TRIAL REGISTRY: International Prospective Register of Systematic Reviews; No.: CRD42022342893; URL: https://www.crd.york.ac.uk/prospero/

CHEST Critical Care 2025; 3(2):100132

KEY WORDS: ARDS; ICU; mortality; prediction models; systematic review

Take-Home Points

Study Question: What are the available prediction models for moderate to severe ARDS and what is their capacity to predict mortality?

Results: Of the 7,455 screened articles, 14 were included, evaluating 20 unique models. Discrimination was reported for all models, whereas calibration was reported in 16 models. The pooled area under the receiving operating characteristic curve (AUC) was 0.782 (95% CI, 0.748-0.817) with an I^2 of 99.5% ($P < .0001$). In subgroup analysis, the pooled AUC for the Sequential Organ Failure Assessment (SOFA) score was 0.802 (95% CI, 0.719-0.885); the AUC for the age, plateau, and Pao_2 to Fio_2 ratio score was 0.724 (95% CI, 0.643-0.805); the AUC for the Acute Physiology and Chronic Health Evaluation (APACHE) II score was 0.667 (95% CI, 0.613-0.721); and the AUC all other scores was 0.813 (95% CI, 0.774-0.852; $P = .0001$ for subgroup differences). The pooled AUC was higher for derivation vs validation studies (0.816; 95% CI, 0.760-0.872] vs 0.767 [95% CI, 0.725-0.809]; $P = .17$ for subgroup differences).

Interpretation: Our findings show that substantial variability in discrimination exists across the included models, with calibration frequently unreported. Although models developed specifically for this patient population demonstrate superior performance, general disease severity models like APACHE and SOFA are validated more extensively. Presently, no extensively validated prediction model exists showing good discrimination and calibration for moderate to severe ARDS.

In daily practice in the ICU, early detection of clinical deterioration and adverse outcomes relies on monitoring vital parameters, biomarkers, and clinical scores. Numerous models have leveraged this information with the aim of predicting outcomes to support clinical decision-making and to inform patients and relatives on prognosis.[1-4] Alongside improvements in clinical ICU management, accurate outcome prediction models also are valuable in research. The information from these models can be used to correct for differences in treatment populations when using causal inference in observational studies (eg, heterogeneity of treatment effect[5]), or to account for disease susceptibility when stratifying randomization in a randomized controlled trial.

Currently, only the Acute Physiology and Chronic Health Evaluation (APACHE) IV and Sequential Organ Failure Assessment (SOFA) scores been have truly implemented in ICU settings.[6] However, these models have been developed for patients in the ICU in general and are not tailored specifically to patients with ARDS. ARDS is characterized by an acute onset of inflammatory hypoxemic respiratory failure and has various causes. It is a major cause of death in the ICU,[7,8] and no improvement in mortality rate has occurred over the past decades.[9,10] Taken together, these models might perform less accurately in this patient population.

Current outcome prediction efforts in ARDS have focused primarily on identifying subgroups with differential outcomes, mainly relying on respiratory parameters or inflammatory biomarkers. The widely accepted Berlin classification based on the Pao_2 to Fio_2 ratio is one example. Nevertheless, this scoring system bears limited predictive accuracy for mortality.[11,12] Another example is the identification of hyperinflammatory and hypoinflammatory ARDS subphenotypes, revealing significant differences in mortality rates and treatment responses.[13,14] Although these subgroup classifications are helpful, individualized mortality prediction might be more beneficial for patients. Such bedside-applicable models could enable clinicians to personalize care and to deliver more accurate prognostic information. To find accurate prediction models that can be applied directly in the ICU,

ABBREVIATIONS: APACHE = Acute Physiology and Chronic Health Evaluation; APPS = Age, Plateau, and Pao_2 to Fio_2 Ratio Score; AUC = area under the receiving operating characteristic curve; POSTCARDS = Predicting Outcome and Stratification of Severity in ARDS; ROB = risk of bias; SOFA = Sequential Organ Failure Assessments

AFFILIATIONS: From the Department of Intensive Care (K. D., S. C. M. S., D. A. M. P. J. G., H. E., and J. A. H.), the Department of Viroscience (K. D. and E. C. M. v. G.), the Department of Immunology (V. A. S. H. D.), the Division of Allergy & Clinical Immunology (V. A. S. H. D.), Department of Internal Medicine (E. C. M. v. G.), Erasmus University Medical Center, Rotterdam; the Department of Medical Microbiology and Infection Prevention (H. v. W.), the Department of Infectious Diseases (A. B.), Amsterdam University Medical Center, University of Amsterdam; the Department of Infectious Diseases (A. B.), Public Health Service of Amsterdam; the HIV Monitoring Foundation (A. B.), the Department of Intensive Care (H. E.), OLVG Amsterdam, Amsterdam; and the Department of Anesthesiology and Intensive Care (A. J. V.), Isala Clinics, Zwolle, The Netherlands.

CORRESPONDENCE TO: Katrijn Daenen, MD; email: k.daenen@erasmusmc.nl

DOI: https://doi.org/10.1016/j.chstcc.2025.100132

and given the critical role of ventilator parameters in these models, we focused our systematic review on mechanically ventilated patients with moderate to severe ARDS. To our knowledge and to date, no overarching analysis of these models has been conducted for this specific patient population. The aim of this study was to review systematically the performance of available prediction models for mortality, as well as to estimate pooled performance statistics for these models, in patients with moderate to severe ARDS.

Study Design and Methods

Study Design

We conducted a systematic review and meta-analysis. Methods and results were reported following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines. The study protocol was registered in the International Prospective Register of Systematic Reviews under registration number CRD42022342893.

Search Strategy

A comprehensive systematic literature search was conducted with assistance of an experienced research librarian using the databases MEDLINE via OVID, EMBASE.com, Web of Science, SCOPUS, PsycINFO via OVID, CINAHL via EBSCOhost, the Cochrane Central Register of Controlled Trials and Google scholar up to March 2024 (e-Appendix 1). The following search terms, alone and in various combinations, were used: *acute respiratory distress syndrome*, *prediction/prognosis*, *mortality/survival*, and *model/nomogram/algorithm/regression*. The recommendations proposed by Bramer et al[15] were followed for identifying prediction model studies for systematic reviews.

Eligibility Criteria

Studies that developed or validated, or both, a multivariable prediction model for mortality in adult patients (ie, aged $\geq$ 18 years) with moderate to severe ARDS in the ICU were included. The definition of moderate or severe ARDS needed to align with either the Berlin criteria or the American European Consensus Conference definition.[11,16] Both pulmonary and extrapulmonary causes of ARDS were included.[17] Studies solely with patients receiving extracorporeal membrane oxygenation were excluded. Pediatric and animal studies, as well as those not published in English, also were excluded.

The outcomes of the prediction model were restricted to hospital or ICU mortality (or survival, if stated as such). Predictors needed to be collected within the first 24 hours after ICU admission. We classified studies as those involving model development (ie, creating a prediction model that may or may not include internal validation[18]) or external validation of a model (ie, evaluating the performance of an existing model using cohorts that were not used for model development[18,19]).

Study Selection

Reference lists of eligible articles were screened to identify additional articles for inclusion. After the initial search, the research librarian removed duplicates. Two reviewers (K. D., H. v. W.) then independently screened all studies by title and abstract and removed irrelevant articles, resolving discrepancies through discussion. The remaining studies were assessed for eligibility using the full text by 2 reviewers (K. D., S. C. M. S.) independently. Discrepancies in final inclusion were resolved between reviewers, with a third reviewer consulted if necessary. EndNote-20 (Clarivate) was used to manage search results.

Data Extraction

From each study, we extracted data on items that were recommended in the Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies checklist.[20] Data were collected on the following performance measure(s) for each model: overall performance, calibration, discrimination, and clinical usefulness. Overall performance is described as the distance between the predicted and the actual outcomes, expressed through the Brier score, R^2 statistic, or overall correctness.[19,21] Calibration measures the consistency between the predicted probability of having an outcome, as modeled, compared with the observed probability of the outcome and typically can be evaluated using calibration plots or the Hosmer-Lemeshow test.[21] Discrimination was defined as the extent to which a model can discriminate patients with and without the outcome, generally expressed by the area under the receiving operating characteristic curve (AUC).[21] Clinical usefulness evaluates the model's impact on decision-making at specific thresholds, considered with some classification measures.[22] Less common measures, such as the C-index, accuracy rate, Youden index, log-rank test, and (positive or negative) predictive values also were collected, when reported. For each statistic,

the 95% CI, SDs, or SEs were collected. One author (K. D.) extracted all data. Data then were verified on a random sample of one-third of the studies by a second author (S. C. M. S.). We contacted the corresponding authors of the eligible studies if data were unclear.

Feasibility Assessment

Using prespecified criteria established by the authors, we also assessed the feasibility of each model based on the following characteristics to enhance actionability in the clinical context: timing, resources, additional costs, and personnel requirements. We defined timing as the time points of data collection for predictors, whereas resources involve the materials needed for data collection. Additional costs indicate any expenses made besides standard ICU care. These were recorded as either yes or no, while further specifying the type of expenses (eg, CT scan). Personnel requirements indicate which health care professionals are necessary for collecting the variables needed for the model.

Risk of Bias Assessment

The Prediction Model Risk of Bias Assessment Tool was used to assess the risk of bias (ROB) of the included studies.[23] Assessment of the selected studies was performed independently by 2 authors (K. D., J. A. H.), and any conflicting results were resolved by a third reviewer (S. C. M. S.).

Statistical Analysis

The AUC was the only consistently reported statistic; hence, it was used in further meta-analysis. Individual AUC and the SE of the AUC were used for each model. If the SE was not reported, it was calculated directly from the SD or was approximated using the 95% CI. If no measure of variation was available, the model was excluded from the meta-analysis.

The pooled AUC with 95% CI initially was calculated using a fixed-effects model with the inverse variance method. The rma function of the metafor package in R software (R Foundation for Statistical Computing) was used to estimate the model. Heterogeneity was tested for using the I^2 statistic, with higher values representing increased heterogeneity between models, and the null hypothesis of no between-study heterogeneity was tested using the Q statistic. If significant heterogeneity was present, the pooled AUC was calculated using a random-effects model with restricted maximum likelihood estimators for τ^2. Subgroup analyses were carried out on the prediction model (ie, APACHE II; Age, Plateau, and Pao_2 to FIO_2 Ratio Score [APPS]; SOFA) and the study type (ie, development or validation), whereas differences in AUC were assessed using a test for subgroup differences with the metagen function in the meta packages. Given the small numbers of studies, sources of heterogeneity were not evaluated using meta-regression.

Further evaluation of whether certain combinations of models were contributing to excessive heterogeneity was performed using the graphical display of study heterogeneity plot.[24] A total of 10,000 combinations were selected randomly to produce the graph using the gosh() function in the metafor package. The influence of outlying models then was assessed using the method developed by Viechtbauer and Cheung[25] in the metafor package. We also assessed for small study effects using the Egger test of small study bias[26] with $P < .05$ indicating significant bias. Analyses were performed using R version 4.2.3 software.

Results

Description of Studies

The initial literature search resulted in a total 7,455 articles. After removing duplicates and screening titles and abstracts, 228 articles remained. After full-text screening, 14 articles fulfilled eligibility criteria and were included in this systematic review and meta-analysis (Fig 1). Seven studies used retrospective data and 7 studies used prospective data. Most studies (n = 8) were conducted in Europe. Eight were single-center studies and 6 were multicenter studies, with most having a sample size of between 50 and 500 patients. Notably, 3 of the included studies reported an a priori sample size calculation,[27-29] and 2 of them were able to include a sample large enough for adequate precision.[27,28] The American European Consensus Conference criteria were used most frequently to define ARDS (n = 7), whereas 2 studies used both Berlin and American European Consensus Conference criteria. Eleven of 14 studies included patients with ARDS with divergent causes. All models included mortality or survival as outcome measures, assessed at either ICU discharge, day 28, or hospital discharge. Nine articles were classified as development studies[28-36] and 5 articles were classified as validation studies.[37-41] In total, these studies evaluated 20 unique prediction models (Table 1).[28-41]

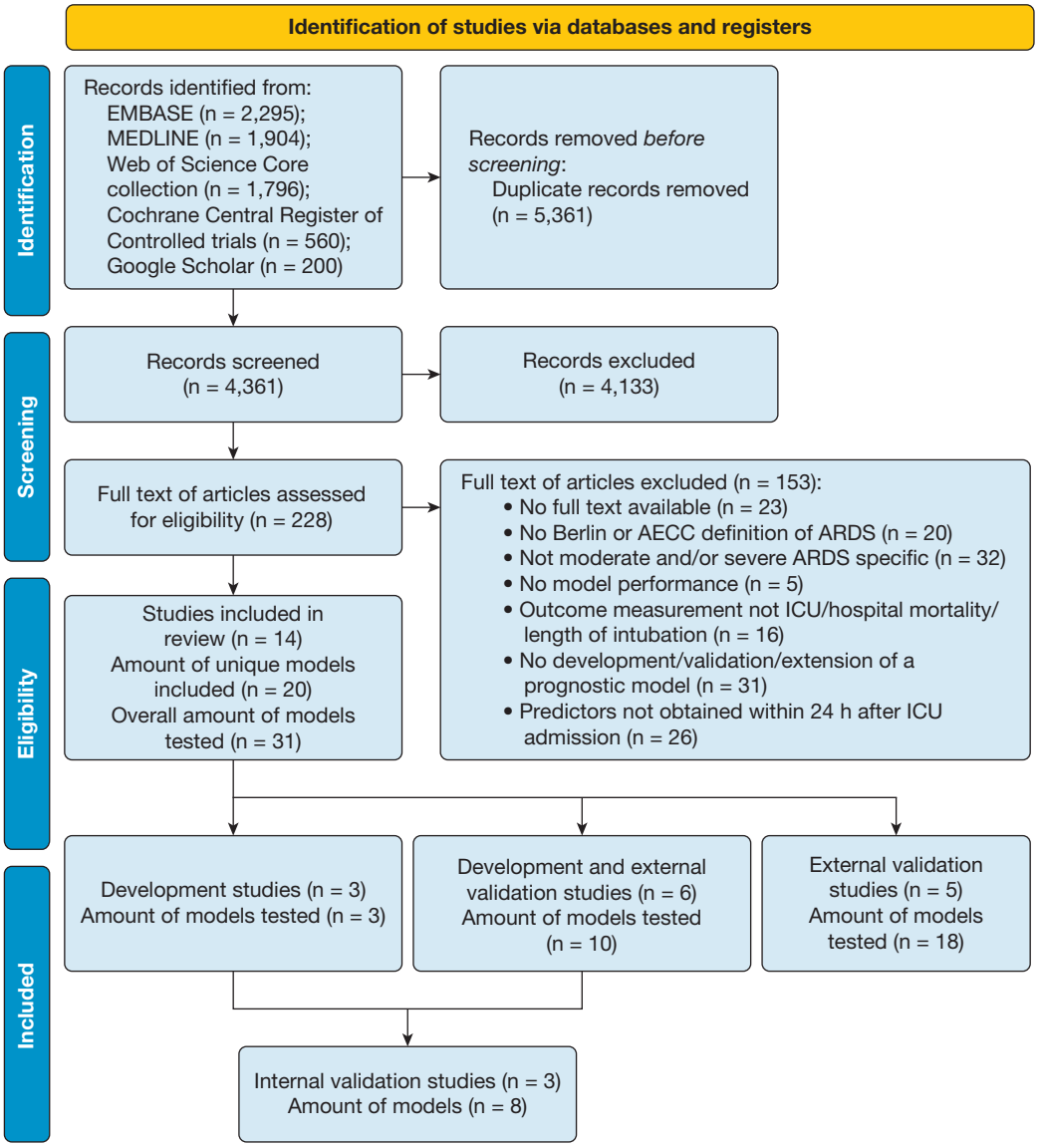

Figure 1 – *Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram for selected studies. AECC = American-European Consensus Conference; EMBASE = Excerpta Medica Database; MEDLINE = Medical Literature Analysis and Retrieval System Online.*

Description of Models Included in the Studies

Development Studies: A total of 13 prediction models were included.[28-36] The most common method used to build the model was logistic regression, with 1 study using machine learning techniques.[35] The most frequently used predictor was age (n = 10), followed by Pa_{O_2} to F_{IO_2} ratio (n = 6) and plateau pressure (n = 6) (Fig 2). Eight development models underwent internal validation using Monte Carlo simulation, bootstrapping, or *n*-fold cross-validation.[30,33-36] Additionally, 7 development models were tested in an external validation cohort, as reported in the development study.[28,30,34-36] Model performance mainly was described using discrimination, with AUC values ranging from 0.72 to 0.95 (Table 2). The highest AUC values were observed with the ARDS score (0.95[34]) and the model by Swaroopa et al[32] (0.94). Six models reported calibration measures: the Hosmer-Lemeshow goodness of fit (n = 5), the intercept of a calibration plot (n = 1), or the slope of a calibration plot (n = 1).[31,33-35] The Predicting Outcome and Stratification of Severity in ARDS (POSTCARDS) model exhibited strong discrimination and calibration in both the development and validation cohorts.[35]

Validation Studies: Eighteen models were externally validated,[28,31,35-41] of which the APACHE II score[28,31,36,38,39] was validated most frequently (n = 6), followed by the SOFA score (n = 3)[31,37,38] and the APPS (n = 2)[39,40] (Table 3).[28,31,35-41] Model

TABLE 1] Overview of Included Studies

Study	No. of Models	Country of Inclusion	Uni-center or Multi-center	Study Design	Sample Size (Calculation)	Study Interval	ARDS		Predicted Outcome	Deaths, %	Development or Validation Model	Model, Method, or Technique	Internal Validation Technique
							Criteria	Cause					
Villar et al (2011)[36]	2	Spain	Multi-center	Retro-spective cohort	220, adhered to power calculation	1999-2005	AECC	All causes	ICU mortality	33.6	Development, internal and external validation	Risk tertiles	Monte Carlo simu-lation test
Villar et al (2019)[30]	3	Spain	Multi-center	Pro-spective trial	1,200	2004-2017	AECC, Berlin	All causes	All-cause death in ICU	37.4	Development, internal and external validation	Logistic regression	Boot-strapp-ing
Villar et al (2016)[28]	2	Spain	Multi-center	Pro-spective observa-tional	600, adhered to power calculation	2008-2015	AECC, Berlin	All causes	All-cause death in hospital	44.3	Development and external validation	Risk tertiles	NR
Türe et al (2005)[31]	5	Turkey	Unicenter	Pro-spective observa-tional	206	1998-2002	Pao_2 to FIo_2 ratio < 150 torr[a]	All causes	ICU mortality	52.4	Development and external validation	Logistic regression	NR
Swaroopa et al (2016)[32]	1	India	Unicenter	Pro-spective observa-tional	30	NR	AECC	NR	28-d mortality	34.6	Development	NR	NR
Sharma et al (2016)[29]	1	India	Unicenter	Pro-spective observa-tional	64, power calculation n = 65	2010-2012	AECC	All causes	28-d mortality	56.2	Development	Multivariable Cox regression	NR
Rocco et al (2001)[33]	1	United States	Unicenter	Retro-spective cohort	111	1990-1998	AECC	All causes	In-hospital mortality	52.3	Development	Logistic regression	NR
Puhr et al (2021)[37]	2	Germany	Unicenter	Retro-spective cohort	53[b]	2020-2021	Berlin	COVID-19	In-hospital mortality	43.4	External validation	Multivariable Cox regression	NR
Monchi et al (1998)[34]	1	France	Unicenter	Retro-spective cohort	259	1992-1995	AECC	NR	In-hospital mortality	65	Development and external validation	Stepwise logistic regression	NR
Lin et al (2010)[38]	5	Taiwan	Multi-center	Retro-spective cohort	135	1996-2006	AECC	Sepsis	In-hospital mortality	65	External validation	Multiple logistic regression	NR

(Continued)

TABLE 1] *(Continued)*

Study	No. of Models	Country of Inclusion	Uni-center or Multi-center	Study Design	Sample Size (Calculation)	Study Interval	ARDS		Predicted Outcome	Deaths, %	Development or Validation Model	Model, Method, or Technique	Internal Validation Technique
							Criteria	Cause					
Hwang et al (2020)[39]	2	Korea	Unicenter	Retro-spective cohort	116	2015-2016	Berlin	All causes	In-hospital mortality	72.4	External validation	Validation APPS	NR
Bos et al (2016)[40]	2	The Netherlands	Multi-center	Pro-spective observa-tional	439	2011-2013	AECC	All causes	All-cause in-hospital-mortality	43	External validation	Validation APPS	NR
Villar and González-Martín (2023)[35]	2	Spain	Multi-center	Pro-spective observa-tional	1,303	2008-2018	Berlin	All causes	All-cause ICU mortality	37.4	Development, internal and external validation	XGboost, RF, logistic regression, SPIRES	5-fold cross-valida-tion
Sanchez et al (2023)[41]	2	Mexico	Unicenter	Retro-spective cohort	115	2020-2021	Berlin	COVID-19	30-d survival	53	External validation	Cox regre-ssion	NR

This table shows an overview of the extracted information of all 14 included studies in our systematic review. AECC = American European Consensus Conference; APPS = Age, Plateau, and Pao_2 to FIo_2 Ratio Score; NR = not reported; RF = random forest; SPIRES = Stratification for Identification of Prognostic Categories in the ARDS Score; XGBoost = extreme gradient boosting.

[a]This study did not mention the Berlin or AECC criteria, but was included because a Pao_2 to FIo_2 ratio of < 150 torr was used as inclusion criterion, which is equivalent to < 150 mm Hg.[28]

[b]Sensitivity analysis with 53 patients with moderate to severe ARDS.

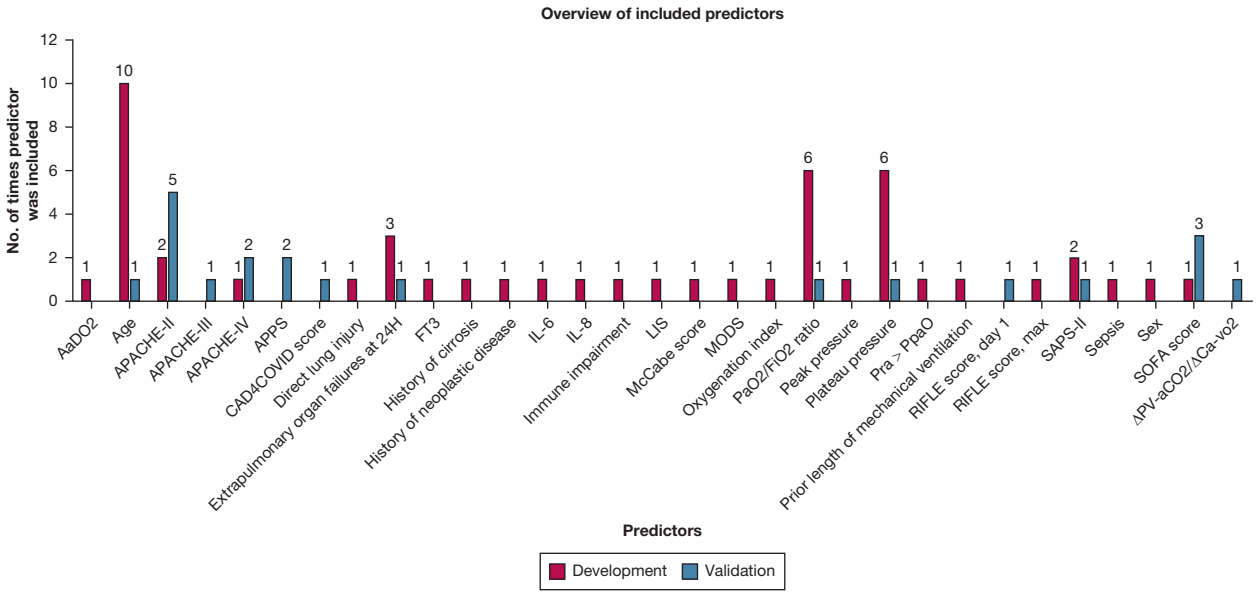

Figure 2 – *Bar graph showing overview of predictors. $AaDO_2$ = alveolar-arterial oxygen tension difference; APACHE = Acute Physiology and Chronic Health Evaluation; APPS = age, plateau, and Pao_2 to Fio_2 ratio score; CAD4COVID = Computer-Aided Detection for Coronavirus Disease (software tool developed to assist in diagnosing COVID-19 based on radiograph); fT3 = free triiodothyronine; LIS = lung injury score; MODS = multiple organ dysfunction score; PpaO = pulmonary artery occlusion pressure; Pra = right atrial pressure; PV-aCO2/Ca-vO2 = venoarterial carbon dioxide tension difference to the arteriovenous oxygen content difference ratio; RIFLE = risk, injury, failure, loss of kidney function, and end-stage kidney disease; SAPS = simplified acute physiology score; SOFA = sequential organ failure assessment.*

performance was described most often using discrimination and AUCs ranged from 0.55 to 0.88. The highest AUC was achieved by the stratification for identification of prognostic categories in the ARDS score model.[35] Measures of calibration were reported for 10 externally validated models: the Hosmer-Lemeshow goodness of fit (n = 9) and the intercept and slope of a calibration plot (n = 1).[31,35,38-40]

Model Feasibility

Most prediction models use variables collected over the entire first 24 hours after ICU admission and rely on commonly available clinical measurements or blood samples (Table 4). Exceptions include the ARDS score, which requires right heart catheterization,[34] and the model of Swaroopa et al,[32] which uses specific biomarkers. Overall, most variables can be collected during standard care by ICU staff, without additional costs.

Pooled AUC for Discrimination

Of the 31 models reporting an AUC measure, the fixed-effects model revealed an I^2 of 99.2% with significant heterogeneity ($P < .0001$), and hence a random-effects model was used. The pooled AUC was 0.782 (95% CI, 0.748-0.817) with an I^2 of 99.5% ($P < .0001$) (Fig 3). In subgroup analysis, the pooled AUC for the SOFA score was 0.802 (95% CI, 0.719-0.885), the APPS score was 0.724 (95% CI, 0.643-0.805), the APACHE II score was 0.667 (95% CI, 0.613-0.721), and all other scores were 0.813 (95% CI, 0.774-0.852; $P = .0001$ for subgroup differences). The pooled AUC was higher for the derivation vs validation studies (0.816 [95% CI, 0.760-0.872] vs 0.767 [95% CI, 0.725-0.809], respectively; $P = .17$ for subgroup differences).

No study subsets gave rise to excessive heterogeneity, according to the graphical display of study heterogeneity plot (e-Fig 1). Also no studies appeared to influence the pooled AUC. From the Egger's test for small study bias, smaller studies (ie, those with lower SE) exhibited significantly lower AUCs ($P = .0009$) (e-Fig 2).

ROB Assessment Using the Prediction Model Risk of Bias Assessment Tool

The ROB assessment using the Prediction Model Risk of Bias Assessment Tool is visualized in Figure 4, with detailed data available in e-Table 1. In the participant and outcome domain, 1 study showed a high ROB,[32] whereas no bias was observed in the predictors domain. Seven studies showed a high ROB in both the analysis domain and overall assessment, mostly because of issues such as not handling missing data, lacking calibration measures, and not accounting for overfitting and optimism.[28,29,31-33,37,41]

TABLE 2] Overview Performance Development Models

Study	Predictors in Final Model	Model Name and Type	Evaluation of Model Performance										
			Calibration, Hosmer-Lemeshow Goodness of Fit/Intercept/ Slope	Discrimination				Clinical Usefulness					
				AUC	Kaplan-Meier Curve	Youden Index, %	Concordance Statistic	Specificity	Sensitivity	PLR	NLR	PPV	NPV
Villar et al (2011)[36]	Age, Pplat, Pao_2 to FIo_2 ratio	Risk tertiles model	NR	*D*, 0.73; *V*, 0.81	NR	NR	NR	NR	NR	NR	NR	NR	NR
Villar et al (2019)[30]	Age, Pao_2 to FIo_2 ratio	Enrichment model	NR	0.74	NR	NR	NR	NR	NR	NR	NR	NR	NR
	Age, Pplat, Pao_2 to FIo_2 ratio	Enrichment model	NR	0.81	NR	NR	NR	NR	NR	NR	NR	NR	NR
	Age, Pplat, Pao_2 to FIo_2 ratio, OF	Enrichment model	NR	0.86 (0.84-0.88)	NR	NR	NR	NR	NR	NR	NR	NR	NR
Villar et al (2016)[28]	Age, Pplat, Pao_2 to FIo_2 ratio	APPS: 9-point score	NR	*D,* 0.76 (0.70-0.81); *V,* 0.80 (0.75-0.85)	Present $P < .001$	NR	NR	NR	NR	NR	NR	NR	NR
Ture et al (2005)[31]	Age, fT3	Logit model	9.86	0.723 (0.052)	NR	34.4	NR	58.0	76.4	NR	NR	NR	NR
	APACHE II score, sex	Logit model	14.03	0.861 (0.036)	NR	54.2	NR	76.0	78.2	NR	NR	NR	NR
	SOFA score, age	Logit model	20.91	0.891 (0.033)	NR	60.4	NR	84.0	76.4	NR	NR	NR	NR
Swaroopa et al (2016)[32]	APACHE II, IL-6, IL-8	NR	NR	0.94	NR	NR	NR	NR	NR	NR	NR	NR	NR
Sharma et al (2016)[29]	OF, SAPS II score, Ppeak	Logistic regression model	NR	NR	Present	NR	NR	NR	NR	NR	NR	NR	NR
Rocco et al (2001)[33]	Age, MODS, LIS	Logistic regression model	0.76	NR	NR	NR	NR	73.6	62.1	NR	NR	NR	NR

(Continued)

TABLE 2] *(Continued)*

Study	Predictors in Final Model	Model Name and Type	Evaluation of Model Performance									
			Calibration, Hosmer-Lemeshow Goodness of Fit/Intercept/Slope	Discrimination				Clinical Usefulness					
				AUC	Kaplan-Meier Curve	Youden Index, %	Concordance Statistic	Specificity	Sensitivity	PLR	NLR	PPV	NPV
Monchi et al (1998)[34]	Cirrhosis, PRA > Ppao, direct lung injury, MV, SAPS II score, McCabe score, OI	ARDS score	*D*, 0.84; *V*, 0.72	*D*, 0.95; *V*, 0.92	NR	NR	NR	NR	NR	NR	NR	NR	NR
Villar and González-Martín (2023)[35,a]	Age, Pplat at 24 h, Pao_2 to FIO_2 ratio, OF, neoplastic disease, immunosuppression, Pplat at BL	POSTCARDS model (7 variables)	*V* RF: intercept, 0.18 (−0.12 to 0.48); slope, 1.13 (0.87-1.39); XGB: intercept, 0.02 (−0.30 to 0.35); slope, 0.95 (0.74-1.16); LR: intercept 0.05 (−0.27 to 0.37); slope, 1.10 (0.85-1.34)	*D* RF, 0.87 (0.82-0.91); XGB, 0.86 (0.81-0.90); LR, 0.87 (0.82-0.91); *V* RF, 0.89 (0.85-0.92); XGB, 0.90 (0.86-0.93); LR, 0.91 (0.87-0.94)	NR	NR	*V* RF, 0.89 (0.85-0.92); XGB, 0.90 (0.86-0.93); LR, 0.91 (0.87-0.94)	*D* RF, 0.80; XGB, 0.80; LR, 0.82; *V* RF, 0.69; XGB, 0.81; LR, 0.84	*D* RF, 0.81; XGB, 0.79; LR, 0.79; *V* RF, 0.90; XGB, 0.88; LR, 0.85	NR	NR	*D* RF, 0.72; XGB, 0.71; LR, 0.73; *V* RF, 0.64; XGB, 0.73; LR, 0.76	*D* RF, 0.88; XGB, 0.87; LR, 0.87; *V* RF, 0.94; XGB, 0.92; LR, 0.90

This table provides an overview of the reported performance measures of the models in the development studies. Certain studies developed and tested multiple models, resulting in performance measures reported in rows underneath each other. APACHE = Acute Physiology and Chronic Health Evaluation; AUC = area under the receiving operating characteristic curve, BL = baseline; *D* = derivation; fT3 = free triiodothyronine; LIS = lung injury score; LR = logistic regression; MODS = Multiple Organ Dysfunction Score; MV = mechanical ventilation; NLR = negative likelihood ratio; NPV = negative predictive value; NR = not reported; OF = extrapulmonary organ failure; OI = oxygenation index; PLR = positive likelihood ratio; Ppao = pulmonary artery occlusive pressure; Ppeak = peak pressure; POSTCARDS = Predicting Outcome and Stratification of Severity in ARDS; Pplat = plateau pressure; PPV = positive predictive value; Pra = right atrial pressure; RF = random forest; SAPS = Simplified Acute Physiology Score; SOFA = Sequential Organ Failure Assessment; *V* = validation; XGBoost = extreme gradient boosting.

[a]In this study, a model was developed using 3 distinct methods: logistic regression analysis following variable selection by a genetic algorithm, random forest, and extreme gradient boosting machine learning techniques.

TABLE 3] Overview Performance Validation Models

Study	Predictors in Final Model	Model Type	Evaluation of Model Performance												
			Calibration, Hosmer-Lemeshow Goodness of Fit	Discrimination					Clinical Usefulness						Overall Performance: Overall Correctness
				AUC	Discriminative Value	Kaplan-Meier Curve	Younden Index, %	Concordance Statistic	Specificity	Sensitivity	PLR	NLR	PPV	NPV	
Villar et al (2011)[36]	APACHE II	Scoring system	NR	*D*, 0.70; *V*, 0.62	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR
Villar et al (2016)[28]	APACHE II	Scoring system	NR	*D*, 0.63 (0.57-0.70); *V*, 0.66 (0.60-0.72)	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR
Türe et al (2005)[31]	APACHE II	Scoring system	45.32	0.85 (0.041)	NR	NR	56.2	NR	78	78.2	NR	NR	78.3	82.2	NR
	SOFA	Scoring system	35.63	0.86 (0.037)	NR	NR	54.0	NR	74.0	80.0	NR	NR	78.5	90.0	NR
Puhr et al (2021)[37]	SOFA	Scoring system	NR	0.77 (0.64-0.89)	7.5	NR	0.46	NR	0.50	0.96	NR	NR	NR	NR	NR
	CAD4COVID	Scoring system	NR	0.55 (0.39-0.72)	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR
Lin et al (2010)[38]	APACHE II	Scoring system	4.088 (df, 8) ($P = .849$)	0.751 (0.042)	NR	NR	0.39	NR	83	57	NR	NR	NR	NR	70.0
	APACHE III	Scoring system	9.191 (df, 8) ($P = .326$)	0.785 (0.039)	NR	NR	0.51	NR	74	77	NR	NR	NR	NR	75.5
	APACHE IV	Scoring system	2.414 (df, 8) ($P = .966$)	0.792 (0.038)	NR	NR	0.51	NR	76	75	NR	NR	NR	NR	75.5
	RIFLE-D1	Scoring system	5.711 (df, 2) ($P = .058$)	0.687 (0.047)	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR
	SOFA	Scoring system	6.691 (df, 8) ($P = .570$)	0.740 (0.042)	NR	NR	0.42	NR	91	51	NR	NR	NR	NR	71.0
Hwang et al (2020)[39]	APPS	Scoring system	$P = .636$	0.711 (0.609-0.813)	NR	Yes	NR	0.70 (0.60-0.81)	NR	NR	NR	NR	NR	NR	NR
	APACHE II	Scoring system	NR	0.624 (0.513-0.736)	NR	Yes	NR	0.62 (0.51-0.74)	NR	NR	NR	NR	NR	NR	NR
Bos et al (2016)[40]	APPS	Scoring system	$P < .001$	0.62 (0.56-0.67)	NR	NR	NR	NR	0.56	0.63	1.43	0.66	NR	NR	NR
	APACHE IV	Scoring system	NR	0.66 (0.61-0.71)	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR

(Continued)

TABLE 3] *(Continued)*

Study	Predictors in Final Model	Model Type	Evaluation of Model Performance											
				Discrimination					Clinical Usefulness						
			Calibration, Hosmer-Lemeshow Goodness of Fit	AUC	Discriminative Value	Kaplan-Meier Curve	Younden Index, %	Concordance Statistic	Specificity	Sensitivity	PLR	NLR	PPV	NPV	Overall Performance: Overall Correctness
Sanchez et al (2023)[41]	ΔPv-aCO2 to ΔCa-vO2 ratio	Ratio	NR	0.69 (0.60-0.77)	NR	Yes	NR	NR	85.2	49.2	3.32	0.60	NR	NR	NR
	SAPS II	Scoring system	NR	0.70 (0.60-0.78)	NR	Yes	NR	NR	NR	NR	NR	NR	NR	NR	NR
Villar and González-Martín (2023)[35,a]	SPIRES	Scoring system	*V*: RF intercept, 0.93 (0.49-1.38); RF slope, 0.38 (0.30-0.46); XGB intercept, 0.10 (−0.20 to 0.40); XGB slope, 1.04 (0.80-1.28); LR intercept, 0.09 (−0.21 to 0.39); LR slope, 1.08 (0.84-1.33)	*D*: 0.86; *V*: RF, 0.85 (0.79-0.90); XGB, 0.87 (0.83-0.91); LR, 0.88 (0.83-0.91)	NR	NR	NR	*V*: RF, 0.85 (0.79-0.90) XGB, 0.87 (0.83-0.91) LR, 0.88 (0.83-0.91)	*V*: RF, 0.83; XGB, 0.84; LR, 0.84	*V*: RF, 0.78; XGB, 0.79; LR, 0.79	NR	NR	*V*: RF, 0.73; XGB, 0.74; LR, 0.74	*V*: RF, 0.86; XGB, 0.87; LR, 0.87	NR

This table provides an overview of the reported performance measures of the models in the external validation studies. Certain studies validated multiple models, resulting in performance measures reported in rows underneath each other. APACHE = Acute Physiology and Chronic Health Evaluation; APPS = age, plateau, and Pao_2 to Fio_2 ratio score; AUC = area under the receiving operating characteristic curve; CAD4COVID = Computer-Aided Detection for Coronavirus Disease; *D* = derivation; ΔCa-vO2 = change in arteriovenous oxygen content difference ratio; ΔPv-aCO2 = change in venous to arterial CO_2 gradient; df = degrees of freedom; LR = logistic regression; NLR = negative likelihood ratio; NPV = negative predictive value; NR = not reported; PLR = positive likelihood ratio; PPV = positive predictive value; RF = random forest; RIFLE-D1 = risk, injury, failure, loss, end-stage kidney disease day 1; SAPS = simplified acute physiology score; SOFA = Sequential Organ Failure Assessment; SPIRES = Stratification for Identification of Prognostic Categories in the ARDS Score; *V* = validation; XGB = extreme gradient boosting.

[a]In this study, a model was developed using 3 distinct methods: logistic regression analysis following variable selection by a genetic algorithm, random forest, and extreme gradient boosting machine learning techniques.

TABLE 4] Overview of Model Feasibility

Study	Model and Predictors	Timing	Resources	Additional Costs Beyond Standard of Care	Personnel
Villar et al (2011)[36]	Age, Pplat, Pa_{O_2} to F_{IO_2} ratio	Immediate	MV, blood sample	No	Nurse or intensivist
Villar et al (2019)[30]	Age, Pplat, Pa_{O_2} to F_{IO_2} ratio, OF	24 h (Pa_{O_2} to F_{IO_2} ratio and Pplat)	MV, blood sample	No	Nurse or intensivist
Villar et al (2016)[28]	Age, Pplat, Pa_{O_2} to F_{IO_2} ratio	24 h (Pa_{O_2} to F_{IO_2} ratio and Pplat)	MV, blood sample	No	Nurse or intensivist
Türe et al (2005)[31]	Age, fT3	24 h (fT3)	Blood sample	Yes, fT3 measurement	Nurse or intensivist
	APACHE III score and sex	24 h (APACHE)	Standard care	No	—
	SOFA score and age	Immediate	Standard care	No	—
Swaroopa et al (2016)[32]	APACHE II score, IL-6, IL-8	24 h (APACHE, IL-6 or IL-8 within 24 h)	Low	Yes, IL-6, and IL-8 measurement	Nurse or intensivist
Sharma et al (2016)[29]	OF, SAPS II score, Ppeak	24 h (SAPS II, OF)	MV, blood sample	No	Nurse or intensivist
Rocco et al (2001)[33]	Age, MODS, LIS	Onset of ARDS	MV, blood sample, chest radiography	No	Nurse or intensivist
Puhr et al (2021)[37]	CAD4COVID	After CT scan	CT scan, CAD4COVID CT scan software tool	Yes, CT scan, and CAD4COVID CT scan software[a]	Radiologist to measure PA to AA ratio
	SOFA score	Immediate	Standard care	No	Nurse or intensivist
Monchi et al (1998)[34]	ARDS score	After right heart catheterization	MV, blood sample, right heart catheterization	Yes, right heart catheterization	Intensivist
Lin et al (2010)[38]	APACHE II, III, and IV scores; RIFLE-D1	After 24 h	Standard care	No	—
	SOFA score	Immediate	Standard care	No	—
Hwang et al (2020)[39]	APPS	At 24 h (maximal airway pressure, Pa_{O_2} to F_{IO_2} ratio)	MV, blood sample	No	Nurse or intensivist

(Continued)

TABLE 4] *(Continued)*

Study	Model and Predictors	Timing	Resources	Additional Costs Beyond Standard of Care	Personnel
	APACHE II score	At 24 h	Standard care	No	—
Bos et al (2016)[40]	APPS	At 24 h (maximal airway pressure, Pao_2 to Fio_2 ratio)	MV, blood sample	No	Nurse or intensivist
Sanchez et al (2023)[41]	ΔPv-aCO2 to ΔCa-vO2 ratio	Within 30 min after intubation	MV, blood sample	No	Nurse or intensivist
	SAPS II score	At 24 h	MV, blood sample	No	—
Villar and González-Martín (2023)[35]	POSTCARDS model and SPIRES	At 24 h (Pao_2 to Fio_2 ratio, Pplat, OF)	Low	No	Nurse or intensivist

This table provides an overview of the feasibility of the models included, based on the factors timing, resources, additional costs, and personnel. Additional costs were recorded as either yes or no, whereas further specifying the types of expenses incurred. Most of the included prediction models did not generate costs beyond standard care. When additional costs were incurred, they were associated with biomarker measurements, extra CT scans, software, and right heart catheterization. AA = ascending aorta; APACHE = Acute Physiology and Chronic Health Evaluation; APPS = age, plateau, and Pao_2 to Fio_2 ratio score; AUC = area under the receiving operating characteristic curve; CAD4COVID = Computer-Aided Detection for COVID; *D* = derivation; fT3 = free triiodothyronine; ΔCa-vO2 = change in arteriovenous oxygen content difference ratio; ΔPv-aCO2 = change in venous to arterial CO_2 gradient; LIS = lung injury score; MODS = Multiple Organ Dysfunction Score; MV = mechanical ventilation; OF = extrapulmonary organ failure; PA = pulmonary artery; POSTCARDS = Predicting Outcome and StratifiCation of Severity in ARDS; Pplat = plateau pressure; RIFLE-D1 = risk, injury, failure, loss, end-stage kidney disease day 1; SAPS = Simplified Acute Physiology Score; SOFA = Sequential Organ Failure Assessment; SPIRES = Stratification for Identification of Prognostic Categories in the ARDS Score; *V* = validation; XGB = extreme gradient boosting.
[a]CAD4COVID CT scan software was free of charge during the pandemic.

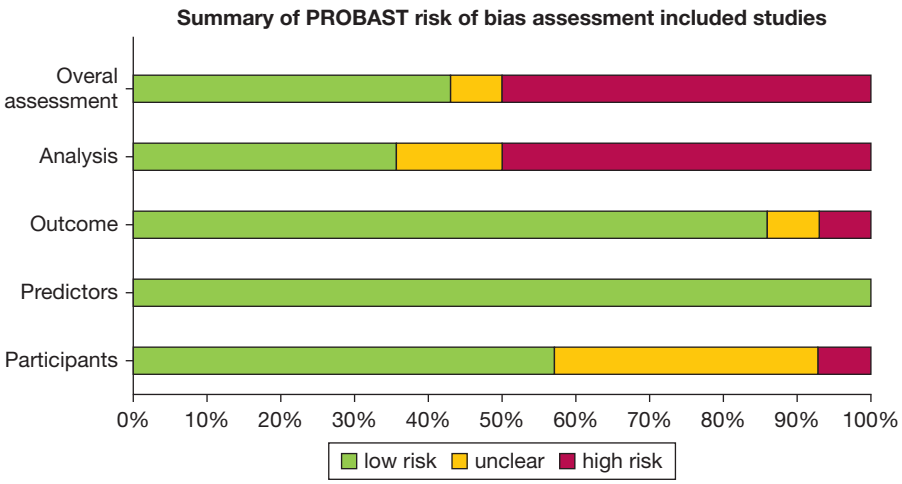

Figure 3 – *Summary of Prediction Model Risk of Bias Assessment Tool risk of bias assessment in included studies. PROBAST = prediction model risk of bias assessment tool.*

Discussion

Our study presents a comprehensive overview of multivariable prediction models for mortality in moderate to severe ARDS. Nine development studies and 5 validation studies evaluated a total of 20 unique prediction models. Models specifically developed for moderate to severe ARDS often show good discriminative power, but frequently lack calibration measures and rarely are validated externally.[32,34,35] Most models were considered feasible because the inclusion criteria that required variables collected within 24 hours was met, and included variables primarily were obtained through routine standard care. General ICU severity models, such as APACHE II and SOFA, were the most frequently validated externally. However, based on the ranges of AUCs and calibration measures, their performance generally was less accurate than models specifically developed for patients with moderate to severe ARDS. Pooled performances were calculated for the APPS, APACHE II score, and SOFA score, with the SOFA score having the highest pooled AUC.

Models specifically developed for moderate to severe ARDS show the best performance and often prioritize respiratory parameters derived from ventilator data and blood sample analysis as predictors. From these models, only the ARDS and POSTCARDS models reported both discrimination and calibration measures and were validated externally. The ARDS score, with good discriminatory power, excellent calibration, and low bias risk, may predict mortality in these patients accurately. However, it has not undergone external validation by another study. Villar et al[28,30,36] and Villar and González-Martín[35] contributed 4 distinct studies that

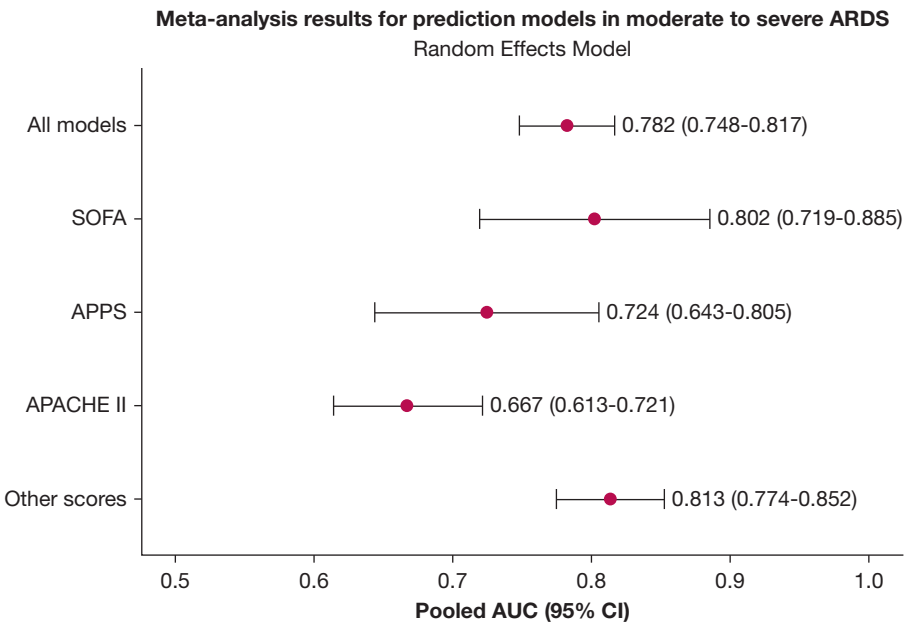

Figure 4 – *Graph showing random effects model of the pooled AUC of all models in moderate to severe acute respiratory distress. APACHE = Acute Physiology and Chronic Health Evaluation; APPS = Age, Plateau, and Pa_{O_2} to Fi_{O_2} Ratio Score; AUC = area under the receiving operating characteristic curve; SOFA = Sequential Organ Failure Assessment.*

both developed and validated models in patients with ARDS, using various combinations of variables, with the APPS score being the only one externally validated by 2 studies. The APPS score had AUCs of 0.76 in the derivation cohort and 0.80 in its validation cohort, but lower AUCs of 0.71 and 0.66, respectively, in external validation studies.[28,39,40] Their recently published POSTCARDS model based on 7 variables is promising, but needs further external validation.[35]

Our findings revealed that the SOFA and APACHE scores, already integrated into general ICU practice, were externally validated most frequently in patients with moderate to severe ARDS. These scores initially were developed to evaluate and predict disease severity and organ failure across a wide range of patients who were critically ill. Although typically demonstrating good performance within general ICU populations (AUC, 0.84-0.85),[42,43] our review showed substantially lower performance in moderate to severe ARDS cohorts. None of the models reached a pooled AUC of > 0.72, with substantial heterogeneity between studies; therefore, caution is warranted to avoid overinterpretation of the results. Our analysis revealed that no individual or group of studies significantly influenced the pooled AUC, thus suggesting minimal ROB from outliers. However, we did find that smaller studies had lower AUCs than expected, indicating small-study bias. Subgroup analyses by score type revealed significant differences in pooled AUCs, likely contributing to the heterogeneity. Additionally, we hypothesize that because ARDS is a syndrome with diverse causes,[44] variations in ARDS causes, patient characteristics, and the evolution of management practices over time[45,46]—such as positive end-expiratory pressure titration[47,48] and corticosteroid therapy[49]—also may play a role. Unfortunately, the inability to perform meta-regression because of the small numbers of studies precludes us from commenting further on the specific sources of heterogeneity.

One-half of the included studies demonstrated a high ROB, which was driven mostly by the analysis domain (eg, not handling missing data, lacking calibration measures, and failing to account for overfitting and optimism).[29,32,33,37,41] For instance, Swaroopa et al[32] presented an AUC of 0.94 in a cohort of 30 patients. The risk of overfitting would have required other techniques to be used, such as penalized regression or bootstrapped CIs. Clinicians need to be cognizant of these shortcomings when deciding to use these scores in practice, whereas future research should strive to minimize bias and use reporting guidance (eg, Standards for Reporting of Diagnostic Accuracy Studies)[50] when presenting results.

Accurate prediction of mortality provides clinicians with valuable information to decide who needs intensified monitoring. This may include more frequent assessments of potential ARDS-related complications or coinfections, increased diagnostic testing (eg, blood cultures or bronchoalveolar lavage), earlier consideration of advanced therapies, and a more proactive multidisciplinary approach. Despite this potential, the feasibility of such models remains difficult to quantify because no universal feasibility scoring system exists. Our systematic review includes a feasibility assessment based on time point of measurements, resources, costs, and personnel—an area that has been relatively ignored in previous literature. Feasibility becomes increasingly important during periods of capacity constraints, aiding in resource and personnel prioritization. Among the included models, feasibility varied with many models posing immediate feasibility challenges by relying on variables collected up to 24 hours after ICU admission, rather than immediately or within the first hour.[28-32,35,38,39] Some models require specific resources, reducing feasibility and complicating implementation, particularly in low-income settings.[34,37] One example is the ARDS score, which relies on right heart catheterization, an invasive, time-consuming, and cost-intensive procedure that is potentially harmful.[34] Although these prediction models provide a structured, quantitative tool for assessing mortality risks, caution is essential in interpreting their results, because they cannot fully capture the clinical nuances and complexity of patients treated in the ICU. Therefore, the use of these scores always should be considered in the specific clinical context, complementary to the expertise of the ICU team. Moving forward, we advocate for the inclusion of feasibility as a criterion in the development of prediction models. An ideal model should be based on parameters accessible on ICU admission, should incorporate ventilator parameters not subject to physician interpretation, and should include a standard care biomarker of inflammation.[51-55]

Our systematic review has several strengths. To our knowledge, this is the first comprehensive overview and meta-analysis within this patient population. Second, we included a feasibility evaluation and an ROB assessment, enhancing directions for future research by identifying potential weaknesses and strengths. However, we also acknowledge several limitations. Some well-conducted

studies were not included because of our specific inclusion criteria, such as those involving patients with mild to severe ARDS with no stratification by ARDS severity. Although the strict selection of patients with moderate to severe ARDS results in a more homogenous patient population, it limits generalizability. Additionally, the predominance of European studies in our review may affect generalizability to health care settings in other regions.

Looking ahead, recent advancements in artificial intelligence have catalyzed the development of novel prediction models. In a review on machine learning-driven models in ARDS, several have been identified.[56] One notable example is the model by Zhang[57] that includes neural networks of predictors, which outperformed the APACHE III score in predicting mortality. This study was excluded from our review because of the inclusion of patients with mild ARDS. Another promising direction for future research lies in shifting the focus of outcome prediction models from mortality to predicting treatment efficacy, such as extracorporeal membrane oxygenation or corticosteroid therapy. Developing models that can predict the potential benefits of these interventions can lead to more personalized and effective treatment strategies for patients with ARDS.

Interpretation

We present a comprehensive overview of mortality prediction models in moderate to severe ARDS. Although models developed specifically for this patient population demonstrate the best performance, general disease severity models like APACHE and SOFA are validated more extensively. Currently, no well-validated model with good discrimination and calibration for moderate to severe ARDS exists. Promising models tailored to ARDS, especially the POSTCARDS models, require further external validation. Furthermore, we emphasize the importance of assessing feasibility before clinical implementation of prediction models.

Funding/Support

The authors have reported to *CHEST Critical Care* that no funding was received for this study.

Financial/Nonfinancial Disclosures

None declared.

Acknowledgements

Author contributions: K. D., J. A. H., H.E., and S. C. M. S. conceptualized the study. K. D., H. v. W., and S. C. M. S. screened and included relevant articles. K. D. and A. B. performed the meta-analysis. K. D., S. C. M. S., H. E., A. B., and J. A. H. were major contributors in drafting the manuscript. All authors gave critical revisions to the manuscript and approved the final version. K. D. is the guarantor of the study.

Availability of data and materials: All data and statistical code for running the data analysis can be found on the website: https://github.com/boyd0094/SELECT_ards_metaanalysis. All other data generated or analyzed during this study are included in this article and e-Appendix 1.

Other contributions: The authors thank Maarten Engel, PhD, for helping with the systematic literature search.

Additional information: The e-Appendix, e-Figures, and e-Table are available online under "Supplementary Data."

References

1. Falcao ALE, Barros AGA, Bezerra AAM, et al. The prognostic accuracy evaluation of SAPS 3, SOFA and APACHE II scores for mortality prediction in the surgical ICU: an external validation study and decision-making analysis. *Ann Intensive Care*. 2019;9(1):18.
2. Gursel G, Demirtas S. Value of APACHE II, SOFA and CPIS scores in predicting prognosis in patients with ventilator-associated pneumonia. *Respiration*. 2006;73(4):503-508.
3. Miu T, Joffe AM, Yanez ND, et al. Predictors of reintubation in critically ill patients. *Respir Care*. 2014;59(2):178-185.
4. Wong HR, Cvijanovich NZ, Anas N, et al. A multibiomarker-based model for estimating the risk of septic acute kidney injury. *Crit Care Med*. 2015;43(8):1646-1653.
5. Kent DM, Paulus JK, van Klaveren D, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) statement. *Ann Intern Med*. 2020;172(1):35-45.
6. Monteiro F, Meloni F, Baranauskas JA, Macedo AA. Prediction of mortality in intensive care units: a multivariate feature selection. *J Biomed Inform*. 2020;107:103456.
7. Bellani G, Laffey JG, Pham T, et al. Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. *JAMA*. 2016;315(8):788-800.
8. Parhar KKS, Zjadewicz K, Soo A, et al. Epidemiology, mechanical power, and 3-year outcomes in acute respiratory distress syndrome patients using standardized screening. An observational cohort study. *Ann Am Thorac Soc*. 2019;16(10):1263-1272.
9. El-Solh AA, Meduri UG, Lawson Y, Carter M, Mergenhagen KA. Clinical course and outcome of COVID-19 acute respiratory distress syndrome: data from a national repository. *J Intensive Care Med*. 2021;36(6):664-672.
10. Virk S, Quazi MA, Nasrullah A, et al. Comparing clinical outcomes of COVID-19 and influenza-induced acute respiratory distress syndrome: a propensity-matched analysis. *Viruses*. 2023;15(4):922.
11. Force ADT, Ranieri VM, Rubenfeld GD, et al. Acute respiratory distress syndrome: the Berlin definition. *JAMA*. 2012;307(23):2526-2533.
12. Busico M, Fuentes NA, Gallardo A, et al. The Predictive validity of the Berlin definition of acute respiratory distress syndrome for patients with COVID-19-related respiratory failure treated with high-flow nasal oxygen: a multicenter, prospective cohort study. *Crit Care Med*. 2024;52(1):92-101.
13. Calfee CS, Delucchi K, Parsons PE, et al. Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled

trials. *Lancet Respir Med*. 2014;2(8):611-620.

14. Calfee CS, Delucchi KL, Sinha P, et al. Acute respiratory distress syndrome subphenotypes and differential response to simvastatin: secondary analysis of a randomised controlled trial. *Lancet Respir Med*. 2018;6(9):691-698.
15. Bramer WM, de Jonge GB, Rethlefsen ML, Mast F, Kleijnen J. A systematic approach to searching: an efficient and complete method to develop literature searches. *J Med Libr Assoc*. 2018;106(4):531-541.
16. Bernard GR, Artigas A, Brigham KL, et al. The American-European Consensus Conference on ARDS. Definitions, mechanisms, relevant outcomes, and clinical trial coordination. *Am J Respir Crit Care Med*. 1994;149(3 pt 1):818-824.
17. Pelosi P, D'Onofrio D, Chiumello D, et al. Pulmonary and extrapulmonary acute respiratory distress syndrome are different. *Eur Respir J Suppl*. 2003;42:48s-56s.
18. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016;69:245-247.
19. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Vol. 19. Springer Science & Business Media; 2009.
20. 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.
21. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-138.
22. Binuya MAE, Engelhardt EG, Schats W, Schmidt MK, Steyerberg EW. Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. *BMC Med Res Methodol*. 2022;22(1):316.
23. 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.
24. Olkin I, Dahabreh IJ, Trikalinos TA. GOSH - a graphical display of study heterogeneity. *Res Synth Methods*. 2012;3(3):214-223.
25. Viechtbauer W, Cheung MW. Outlier and influence diagnostics for meta-analysis. *Res Synth Methods*. 2010;1(2):112-125.
26. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997;315(7109):629-634.
27. Villar J, Pérez-Méndez L, Basaldúa S, et al. A risk tertiles model for predicting mortality in patients with acute respiratory distress syndrome: age, plateau pressure, and PaO2/FiO2 at ARDS onset can predict mortality. *Respiratory*. 2011;56(4):420-428.
28. Villar J, Ambrós A, Soler JA, et al. Age, Pao2/Fio2, and plateau pressure score: A proposal for a simple outcome score in patients with the acute respiratory distress syndrome. *Crit Care Med*. 2016;44(7):1361-1369.
29. Sharma SK, Gupta A, Biswas A, et al. Aetiology, outcomes & predictors of mortality in acute respiratory distress syndrome from a tertiary care centre in North India. *Indian J Med Res*. 2016;143(JUNE):782-792.
30. Villar J, Ambros A, Mosteiro F, et al. A Prognostic Enrichment Strategy for Selection of Patients With Acute Respiratory Distress Syndrome in Clinical Trials. *Critical care medicine*. 2019;47(3):377-385.
31. Türe M, Memiş D, Kurt I, Pamukçu Z. Predictive value of thyroid hormones on the first day in adult respiratory distress syndrome patients admitted to ICU: Comparison with SOFA and APACHE II scores. *Ann Saudi Med*. 2005;25(6):466-472.
32. Swaroopa D, Bhaskar K, Mahathia T, et al. Association of serum interleukin-6, interleukin-8, and Acute Physiology and Chronic Health Evaluation II score with clinical outcome in patients with acute respiratory distress syndrome. *Indian J Crit Care Med*. 2016;20(9):518-525.
33. Rocco TR Jr, Reinert SE, Cioffi W, Harrington D, Buczko G, Simms HH. A 9-year, single-institution, retrospective review of death rate and prognostic factors in adult respiratory distress syndrome. *Ann Surg*. 2001;233(3):414-422.
34. Monchi M, Bellenfant F, Cariou A, et al. Early predictive factors of survival in the acute respiratory distress syndrome: A multivariate analysis. *Am J Respir Crit Care Med*. 1998;158(4):1076-1081.
35. Villar J, González-Martín JM. Predicting ICU mortality in Acute Respiratory Distress Syndrome patients using machine learning: the Predicting Outcome and STratifiCation of severity in ARDS. *Crit Care Med*. 2023;51(12):1638-1649.
36. Villar J, Pérez-Méndez L, Basaldúa S, et al. A risk tertiles model for predicting mortality in patients with acute respiratory distress syndrome: age, plateau pressure, and PaO2/FIO2 at ARDS onset can predict mortality. *Respir Care*. 2011;56(4):420-428.
37. Puhr-Westerheide D, Reich J, Sabel BO, et al. Sequential Organ Failure Assessment outperforms quantitative chest CT imaging parameters for mortality prediction in COVID-19 ARDS. *Diagnostics (Basel)*. 2021;12(1):10.
38. Lin CY, Kao KC, Tian YC, et al. Outcome scoring systems for acute respiratory distress syndrome. *Shock*. 2010;34(4):352-357.
39. Hwang H, Choi SM, Lee J, et al. Validation of age, PaO2/FlO2 and plateau pressure score in Korean patients with acute respiratory distress syndrome: a retrospective cohort study. *Respir Res*. 2020;21(1):94.
40. Bos LD, Schouten LR, Cremer OL, et al. External validation of the APPS, a new and simple outcome prediction score in patients with the acute respiratory distress syndrome. *Ann Intensive Care*. 2016;6(1):89.
41. Sanchez Diaz JS, Peniche Moguel KG, Reyes-Ruiz JM, et al. The Pv-aCO2/Ca-vO2 ratio as a predictor of mortality in patients with severe acute respiratory distress syndrome related to COVID-19. *PloS One*. 2023;18(9):e0290272.
42. Bennett CE, Wright RS, Jentzer J, et al. Severity of illness assessment with application of the APACHE IV predicted mortality and outcome trends analysis in an academic cardiac intensive care unit. *J Crit Care*. 2019;50:242-246.
43. Sungono V, Hariyanto H, Soesilo TEB, et al. Cohort study of the APACHE II score and mortality for different types of intensive care unit patients. *Postgrad Med J*. 2022;98(1166):914-918.
44. Meyer NJ, Gattinoni L, Calfee CS. Acute respiratory distress syndrome. *Lancet*. 2021;398(10300):622-637.
45. Alqahtani JS, Mendes RG, Aldhahir A, et al. Global current practices of ventilatory support management in COVID-19 patients: an international survey. *J Multidiscip Healthc*. 2020;13:1635-1648.
46. Ashbaugh DG, Bigelow DB, Petty TL, Levine BE. Acute respiratory distress in adults. *Lancet*. 1967;2(7511):319-323.
47. Millington SJ, Cardinal P, Brochard L. Setting and titrating positive end-expiratory pressure. *Chest*. 2022;161(6):1566-1575.
48. Pelosi P, Ball L, Barbas CSV, et al. Personalized mechanical ventilation in acute respiratory distress syndrome. *Crit Care*. 2021;25(1):250.
49. Annane D, Pastores SM, Rochwerg B, et al. Guidelines for the diagnosis and management of critical illness-related corticosteroid insufficiency (CIRCI) in critically ill patients (part I): Society of Critical Care Medicine (SCCM) and European Society of Intensive Care Medicine (ESICM) 2017. *Intensive Care Med*. 2017;43(12):1751-1763.
50. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. 2016;6(11):e012799.
51. Ren Y, Zhang L, Xu F, et al. Risk factor analysis and nomogram for predicting in-hospital mortality in ICU patients with sepsis and lung infection. *BMC Pulm Med*. 2022;22(1):17.
52. Yan Y, Xie Y, Wang Y, et al. [Diagnostic value of mechanical power in patients with moderate to severe acute respiratory

distress syndrome: an analysis using the data from MIMIC-III]. *Zhonghua Wei Zhong Bing Ji Jiu Yi Xue*. 2022;34(1): 35-40.

53. Kao HC, Lai TY, Hung HL, et al. Sequential oxygenation index and organ dysfunction assessment within the first 3 days of mechanical ventilation predict the outcome of adult patients with severe acute respiratory failure. *ScientificWorldJournal*. 2013;2013: 413216.

54. Huber W, Findeisen M, Lahmer T, et al. Prediction of outcome in patients with ARDS: a prospective cohort study comparing ARDS-definitions and other ARDS-associated parameters, ratios and scores at intubation and over time. *PloS One*. 2020;15(5):e0232720.

55. Balzer F, Menk M, Ziegler J, et al. Predictors of survival in critically ill patients with acute respiratory distress syndrome (ARDS): an observational study. *BMC Anesthesiol*. 2016;16(1):108.

56. Bhattarai S, Gupta A, Ali E, et al. Can big data and machine learning improve our understanding of acute respiratory distress syndrome? *Cureus*. 2021;13(2): 13529.

57. Zhang Z. Prediction model for patients with acute respiratory distress syndrome: use of a genetic algorithm to develop a neural network model. *PeerJ*. 2019;7:7719.