



# External validation of a biomarker and clinical prediction model for hospital mortality in acute respiratory distress syndrome

Zhiguo Zhao<sup>1,2</sup>, Nancy Wickersham<sup>3</sup>, Kirsten N. Kangelaris<sup>4</sup>, Addison K. May<sup>5</sup>, Gordon R. Bernard<sup>3</sup>, Michael A. Matthay<sup>6,7</sup>, Carolyn S. Calfee<sup>6,7</sup>, Tatsuki Koyama<sup>1</sup> and Lorraine B. Ware<sup>3,8\*</sup>

© 2017 Springer-Verlag Berlin Heidelberg and ESICM

## Abstract

**Purpose:** Mortality prediction in ARDS is important for prognostication and risk stratification. However, no prediction models have been independently validated. A combination of two biomarkers with age and APACHE III was superior in predicting mortality in the NHLBI ARDSNet ALVEOLI trial. We validated this prediction tool in two clinical trials and an observational cohort.

**Methods:** The validation cohorts included 849 patients from the NHLBI ARDSNet Fluid and Catheter Treatment Trial (FACTT), 144 patients from a clinical trial of sivelestat for ARDS (STRIVE), and 545 ARDS patients from the VALID observational cohort study. To evaluate the performance of the prediction model, the area under the receiver operating characteristic curve (AUC), model discrimination, and calibration were assessed, and recalibration methods were applied.

**Results:** The biomarker/clinical prediction model performed well in all cohorts. Performance was better in the clinical trials with an AUC of 0.74 (95% CI 0.70–0.79) in FACTT, compared to 0.72 (95% CI 0.67–0.77) in VALID, a more heterogeneous observational cohort. The AUC was 0.73 (95% CI 0.70–0.76) when FACTT and VALID were combined.

**Conclusion:** We validated a mortality prediction model for ARDS that includes age, APACHE III, surfactant protein D, and interleukin-8 in a variety of clinical settings. Although the model performance as measured by AUC was lower than in the original model derivation cohort, the biomarker/clinical model still performed well and may be useful for risk assessment for clinical trial enrollment, an issue of increasing importance as ARDS mortality declines, and better methods are needed for selection of the most severely ill patients for inclusion.

**Keywords:** Validation, Prediction, Biomarker, Hospital mortality, ARDS

\*Correspondence: lorraine.ware@vanderbilt.edu

<sup>3</sup> Division of Allergy, Pulmonary and Critical Care Medicine, Department of Medicine, Vanderbilt University, T1218 MCN, 1161 21st Avenue S., Nashville, TN 37232-2650, USA

Full author information is available at the end of the article

Tatsuki Koyama and Lorraine B. Ware contributed equally to this manuscript.

**Take-home message:** The externally validated biomarker/clinical prediction model may provide prognostic and predictive enrichment in clinical trials enrollment, and improve bedside prognostication.

## Introduction

The acute respiratory distress syndrome (ARDS) is responsible for more than 2 million critical care days and 75,000 deaths in the USA yearly [1]. There is a pressing need for development and clinical testing of new therapies that might improve clinical outcomes in ARDS. However, the design of investigational trials for this complex and heterogeneous syndrome is not straightforward

[2]. The success of clinical trials in ARDS is highly dependent on the power, which is primarily determined by the mortality rate for enrolled patients [3–5]. Methods to better predict hospital mortality may provide a basis for prognostic enrichment [6], optimizing the power of clinical trials to detect a treatment effect, and improving bedside prognostication [7].

To date, much effort has been spent on identifying predictors of mortality in patients with ARDS [8–18], and developing scoring systems to improve prognostication [19–23]. However, the two most widely used scoring systems, APACHE III [21] and SAPS 3 [22], were developed and validated in general ICU patients; these scores were not focused on patients with ARDS. Other simpler scoring systems have been developed in the target population of ARDS patients [19, 20, 23, 24]. However, these scores were either outperformed by APACHE III [19, 20] or could not be validated in independently collected data [20, 23, 24]. Recently, Ware et al. demonstrated that a combination of plasma biomarkers of inflammation and lung epithelial injury [interleukin-8 (IL-8), surfactant protein D (SP-D)] and clinical predictors (age, APACHE III) was superior to either biomarkers or clinical factors alone in predicting ARDS mortality in patients enrolled in the NHLBI ARDSNet ALVEOLI trial [25]. However, this biomarker/clinical prediction model (biomarker/clinical model) has not yet been externally validated across multiple independent patient groups.

In the current study, the primary goal was to validate the previously published biomarker/clinical model in three independent ARDS patient cohorts, including both clinical trial cohorts and a heterogeneous group of patients enrolled in an observational cohort study. A second goal was to confirm that inclusion of the two biomarkers added value for predicting ARDS mortality in these independent cohorts. A portion of this work was presented in abstract form at the American Thoracic Society International Conference in 2015 [26].

## Materials and methods

### Original prediction model

The previously reported hospital mortality model was developed with 528 patients from the NHLBI ARDS Clinical Trials Network multicenter randomized controlled trial of two PEEP titration strategies (the ALVEOLI study) [27]. Study details and prediction model development have been published [25]. Briefly, the model includes patient age, APACHE III score, and plasma IL-8 and SP-D as predictors. The formula for the model is shown in Supplemental Fig. 1 (e-Fig. 1), and a Web-based module is available (<https://cqs.mc.vanderbilt.edu/shiny/ChestModel/>).

### Study population and measurements

Detailed methods are provided in the Supplemental methods (e-methods). Briefly, the current study included a total of 1538 mechanically ventilated patients with ARDS who participated either in the NIH ARDS Network Fluid and Catheter Treatment Trial (FACTT) [28], the Sivelestat Trial in ALI Patients Requiring Mechanical Ventilation (STRIVE) [29], or the Validating Acute Lung Injury biomarkers for Diagnosis (VALID) study [30]. All eligible patients in FACTT and STRIVE were mechanically ventilated and had acute lung injury (ALI) or ARDS by American European Consensus Conference (AECC) definitions [31], thus meeting the current Berlin definition of ARDS [32]. All VALID patients were eligible for inclusion if they were mechanically ventilated on at least 1 day and met Berlin ARDS criteria on two consecutive days of the first four ICU days. IRB approval was obtained in all studies; informed consent was obtained from all subjects except in VALID, where a subset of subjects was enrolled under a waiver of informed consent. The current study includes 849, 144, and 545 participants from FACTT, STRIVE, and VALID, respectively, depending on the availability of the clinical data and plasma samples for biomarker measurements. For some analyses the FACTT and VALID patients were combined into one cohort.

Plasma samples were obtained at enrollment (prior to randomization) for patients in the FACTT and STRIVE trials, and on the morning of ICU day 2 in VALID. SP-D and IL-8 were measured in stored plasma samples from each study for this validation. Age and APACHE score were extracted from each study database. APACHE II scores were converted to estimated APACHE III for patients in VALID and STRIVE using a translation equation ( $\text{APACHE III} = 5.57 + 3.08 \times \text{APACHE II}$ ) that was developed in a cohort of 634,428 patients [33].

### Statistical analysis

Detailed analytical approaches are reported in e-methods. Briefly, demographics, clinical variables, and biomarker values were summarized and compared by individual study cohort and combined. For the primary goal, to evaluate the performance of the prediction model in the independent validation sets, model discrimination and calibration were assessed. Discrimination was quantified using the area under the receiver operator characteristic (ROC) curve (AUC), also known as the C-statistic. The 95% confidence intervals (CI) calculated from 300-iteration bootstrap were reported. The benchmark AUC, which is the best possible AUC by refitting the model on each validation dataset, was also reported to provide readers an estimate of optimal discrimination on each validation cohort as a reference. Calibration

was assessed graphically with a calibration plot. A simple recalibration method (logistic calibration) to recalibrate the model as suggested by Harrell et al. [34] and Janssen et al. [35] was also used. For the second goal, to evaluate the added value of the two biomarkers in predicting hospital mortality in the validation datasets, the likelihood ratio test, the net reclassification improvement (NRI), and integrated discrimination improvement (IDI) were used. Finally, to demonstrate a potential application of the prediction model, we stratified the participants in FACTT into low and high mortality risk groups and then evaluated the effects of the treatments separately in each subgroup. The differences in the treatment effects between these two subgroups were evaluated by testing the interaction term of the treatment by risk group in models. For ventilator-free days (VFD), zero inflated negative binomial models were used because of the high frequency of patients who had zero VFDs. This demonstration was not attempted in STRIVE because of the limited sample size. Statistical significance was considered at a two-sided 5% level. All statistical analyses were performed using R software version 3.3.1.

## Results

### Patient characteristics

The participants in the model development cohort (ALVEOLI,  $n = 528$ ) and validation cohorts were similar in age, but different in hospital mortality rate, APACHE scores, cause of ARDS, and distribution of biomarker values (Table 1). The overall hospital mortality rates in FACTT ( $n = 849$ ) or FACTT and VALID combined ( $n = 1394$ ) were lower than in ALVEOLI (19% or 21 vs 27%,  $P < 0.01$ , Table 1), while the hospital mortality rates were similar in VALID ( $n = 545$ , 24 vs. 27%,  $P = 0.25$ , Table 1) and STRIVE ( $n = 144$ , 32 vs. 27%,  $P = 0.27$ , e-Table 1) to ALVEOLI.

### Discrimination and calibration

Despite the differences in patient characteristics, when we applied (with fixed model coefficients) the original biomarker/clinical model to the validation sets, the discrimination for hospital mortality was good. The model achieved AUCs of 0.74 (95% CI 0.70–0.79), 0.72 (95% CI 0.67–0.77), and 0.73 (95% CI 0.70–0.76) in FACTT, VALID, and the combined dataset, respectively (Table 2), which are similar to the benchmark AUCs in the independent study cohorts (0.75, 0.74, and 0.73, respectively, Table 2). In STRIVE, the model achieved AUC of 0.78 (95% CI 0.70–0.87, e-Table 2), comparing with the benchmark AUC of 0.87.

In the FACTT and VALID cohorts, the original biomarker/clinical model tended to predict somewhat higher hospital mortality risk than the actual observed mortality among those at highest risk, indicated by the right tail of the calibration curve falling below the 45° line. Overall, the slopes for the calibration curves were 0.53, 0.56, and 0.52 in FACTT, VALID or combined, respectively (Fig. 1a, c, e). By contrast, the calibration curve in STRIVE was shifted up, but almost parallel to the ideal line with an estimated slope of 0.92 (e-Fig. 2A). After recalibration, however, the model calibrated well in all validation cohorts (Fig. 1b, d, f; e-Fig. 2B). The tails of the calibration curves in the FACTT and VALID remain slightly off the 45° line, which may primarily be due to the limited number of patients with an extremely high predicted hospital mortality.

### Two biomarkers added predictive value

In the original published model, inclusion of the two biomarkers along with patient age and APACHE score significantly improved the model fit and the predictive ability [25]. To confirm that the two biomarkers (SP-D and IL-8) were also of value in the validation cohorts, we analyzed the added value of these two biomarkers compared to clinical factors alone. The model AUCs increased from 0.72 to 0.75, 0.67 to 0.74, and 0.70 to 0.73 in FACTT, VALID, and combined, respectively (Table 3), with addition of the two biomarkers to the clinical variables. The NRI ranged from 0.41 to 0.45 and the IDI ranged from 0.04 to 0.09. All of these improvements reached statistical significance ( $P < 0.001$ , Table 3).

### Prognostic enrichment: an illustration

To illustrate how the mortality prediction model might have value for prognostic enrichment in a clinical trial, we applied the original biomarker/clinical model to patients in the FACTT cohort, classifying patients into two prognostic groups: a low-risk group (predicted mortality at most 20%) or a high-risk group (predicted mortality greater than 20%). We then assessed the treatment effect of randomization to conservative versus liberal fluid therapy separately in each prognostic group. In the low-risk group, no significant treatment effect for conservative versus liberal fluid therapy was observed with regard to mortality or ventilator-free days (VFDs) (Table 4). In the high-risk group, however, there was a significant treatment effect for conservative fluid therapy; those randomized to conservative fluid therapy had 20% more VFDs compared to those randomized to liberal fluid therapy (RR = 1.2, 95% CI 1.09–1.33). These

**Table 1 Patients characteristics in the derivation cohort (ALVEOLI), and the external validation cohorts (FACTT and VALID)**

Characteristics	Derivation cohort		External validation cohorts				
	ALVEOLI (N = 528)	FACTT (N = 849)		VALID (N = 545)		FACTT + VALID (N = 1394)	
	Summary <sup>b</sup>	Summary <sup>b</sup>	P <sup>c</sup>	Summary <sup>b</sup>	P <sup>c</sup>	Summary <sup>b</sup>	P <sup>c</sup>
Age (years)	50 (39, 65)	49 (39, 61)	0.22	53 (39, 64)	0.41	50 (39, 62)	0.68
Male gender	290 (55%)	449 (53%)	0.46	313 (57%)	0.41	762 (55%)	0.92
Caucasian race	398 (75%)	555 (65%)	<0.01	466 (86%)	<0.01	1021 (73%)	0.34
APACHE III score <sup>a</sup>	92 (71, 144)	91 (70, 116)	1.00	95 (79, 107)	0.63	92 (73, 112)	0.82
Plasma SP-D (ng/ml)	99 (50, 212)	136 (63, 283)	<0.01	60 (34, 112)	<0.01	96 (46, 216)	0.27
Plasma IL-8 (pg/ml)	40 (16, 98)	32 (16, 78)	0.01	22 (6, 78)	<0.01	28 (13, 78)	<0.01
Cause of ARDS			<0.01		<0.01		<0.01
Sepsis	117 (22.2%)	207 (24.4%)		147 (27.0%)		354 (25.4%)	
Pneumonia	209 (39.6%)	397 (46.8%)		104 (19.1%)		501 (36.0%)	
Trauma	43 (8.1%)	62 (7.3%)		180 (33.1%)		242 (17.4%)	
Multiple transfusion	26 (4.9%)	8 (0.9%)		18 (3.3%)		26 (1.9%)	
Aspiration	81 (15.3%)	121 (14.3%)		71 (13.1%)		192 (13.8%)	
Other	52 (9.8%)	54 (6.4%)		24 (4.4%)		78 (5.6%)	
Number of nonpulmonary organ failures <sup>d</sup>					<0.01		
0	209 (39.6%)	–		111 (20.4%)		–	
1	208 (39.4%)	–		245 (45.0%)		–	
2	82 (15.5%)	–		140 (25.7%)		–	
3	24 (4.5%)	–		39 (7.2%)		–	
4	5 (0.9%)	–		10 (1.8%)		–	
Hospital mortality			<0.01		0.25		<0.01
Alive	384 (73%)	684 (81%)		413 (76%)		1097 (79%)	
Dead	144 (27%)	165 (19%)		132 (24%)		297 (21%)	

<sup>a</sup> APACHE III scores were recorded in ALVEOLI and FACTT. APACHE II scores were recorded in VALID and translated to APACHE III using the formula APACHE III = 5.57 + 3.08 × APACHE II

<sup>b</sup> In FACTT, the number of nonpulmonary organ failures was not available to the current study

<sup>c</sup> Compared with ALVEOLI cohort; Pearson  $\chi^2$  test for categorical characteristics and Wilcoxon rank sum test for continuous characteristics

<sup>d</sup> Median (IQR) for continuous characteristics

**Table 2 Discrimination of the original biomarker/clinical model in the derivation (ALVEOLI) and external validation cohorts as measured by the area under the receiver operator characteristic curve (AUC)**

Measurements	Derivation cohort		External validation cohorts	
	ALVEOLI (N = 528)	FACTT (N = 849)	VALID (N = 545)	FACTT + VALID (N = 1394)
	AUC	AUC (95% CI)	AUC (95% CI)	AUC (95% CI)
AUC <sup>a</sup> <sub>Original biomarker/clinical model</sub>	0.83	0.74 (0.70, 0.79)	0.72 (0.67, 0.77)	0.73 (0.70, 0.76)
AUC <sup>b</sup> <sub>Benchmark</sub>	–	0.75 (0.72, 0.80)	0.74 (0.70, 0.80)	0.73 (0.70, 0.77)

AUC area under the ROC curve

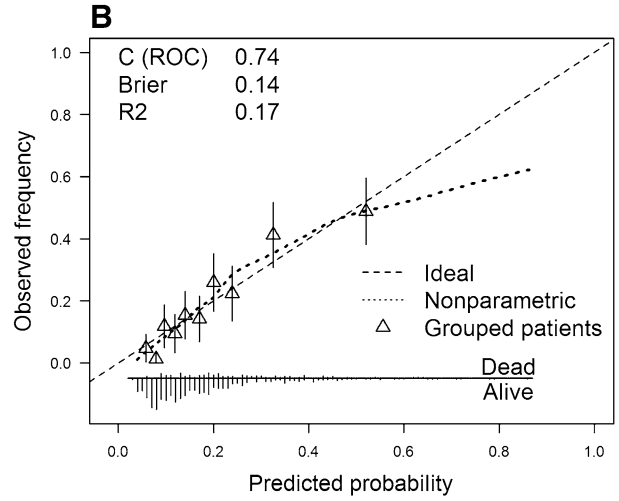
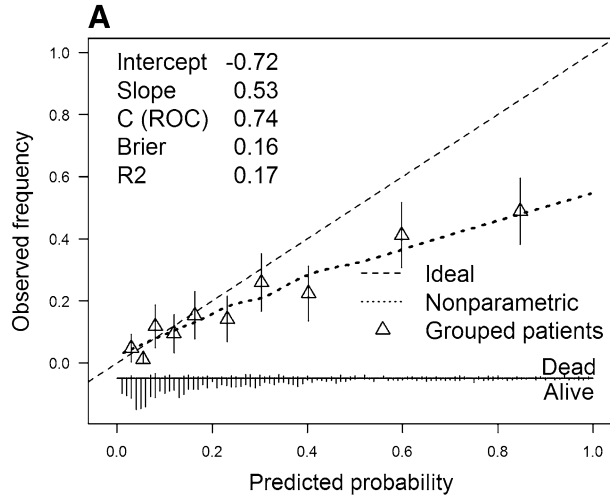
<sup>a</sup> From applying the original biomarker/clinical model on new datasets without refitting (re-estimating the coefficients)

<sup>b</sup> Benchmark AUC by refitting the original biomarker/clinical model that includes age, APACHE, SP-D, and IL-8 on the dataset

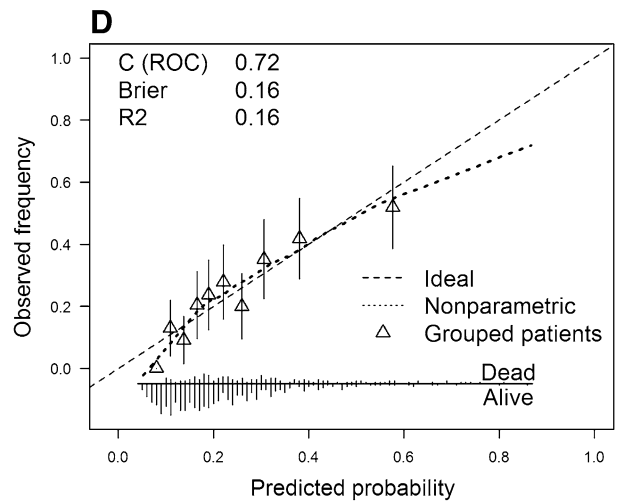
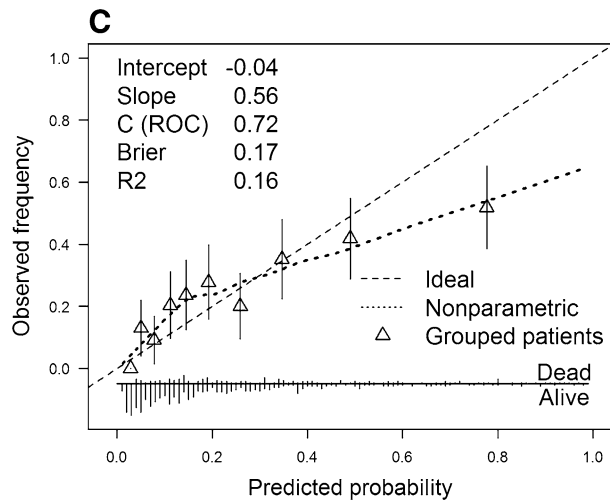
findings illustrate how prognostic enrichment using the mortality prediction model could be used to target clinical trial enrollment to a subset of patients with ARDS

at a higher risk of a clinical outcome of interest, thereby improving the power of the study to detect a treatment effect.

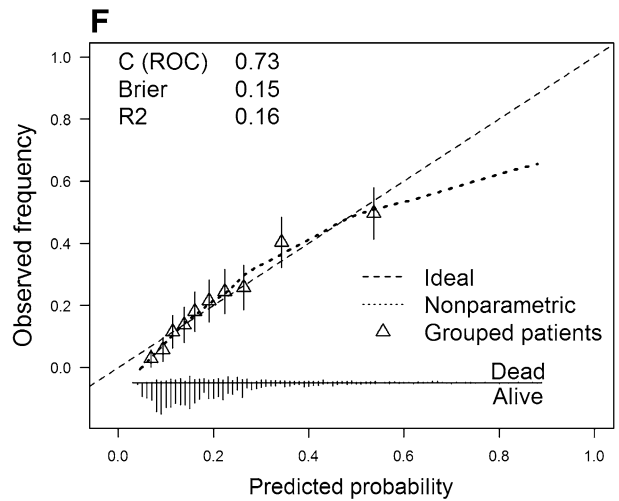
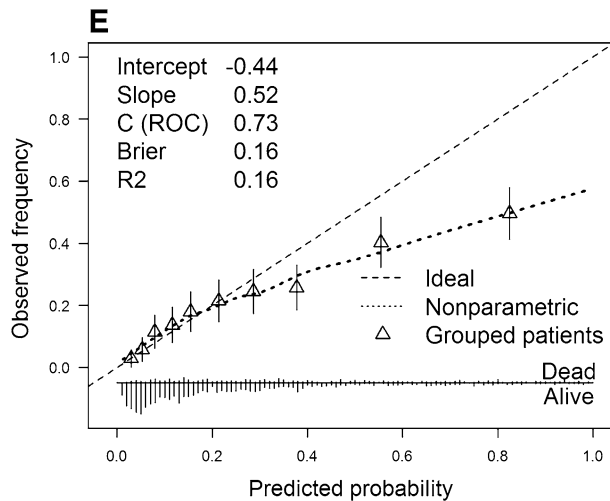
**The FACTT Study (N=849)**



**The VALID Study (N=545)**



**FACTT and VALID Combined (N=1394)**



(See figure on previous page.)

**Fig. 1 a, c, e** Calibration plots of the original biomarker/clinical model and **b, d, f** recalibrated model in the validation cohorts. The recalibration was done by (1) estimating the calibration intercept and calibration slope; (2) multiplying all the regression coefficients of the original biomarker/clinical model by the calibration slope; (3) updating the intercept of the original biomarker/clinical model with the calibration intercept. This method does not involve re-estimating the coefficient of any individual predictor

## Discussion

Despite decades of experimental and clinical investigation, and improvements over time in ICU survival rates [36], effective pharmacotherapy for ARDS remains extremely limited [37, 38]. Inadequately powered trials [3–5] and failure to identify appropriate subsets of patients for enrollment may have contributed to the persistent lack of effective pharmacologic interventions. A recent application of latent class analysis methods to several NHLBI ARDS Network trials has consistently identified two subphenotypes within enrolled ARDS patients, a finding that may be useful to reduce heterogeneity in ARDS clinical trials and potentially provide a basis for predictive enrichment in clinical trials [2, 39]. Prognostic

enrichment is an approach that could be used to identify patients with a higher risk of death for enrollment in clinical trials. This approach has recently been recommended by the US Food and Drug Administration (FDA) to improve efficiency of drug development. However, development of methods that can predict patient clinical outcomes, such as hospital mortality in ARDS, remains challenging.

In the current study, utilizing patients from three independent, heterogeneous cohorts of patients with ARDS, we externally validated a previously published biomarker/clinical model for hospital mortality in ARDS [25]. In the validation cohorts, the AUCs of the original biomarker/clinical model were close to benchmark AUCs, indicating

**Table 3 Evaluation of the added value of the two biomarkers for mortality prediction compared to age and APACHE alone in the derivation and validation cohorts**

Measurements	Derivation cohort		External validation cohorts					
	ALVEOLI (N = 528)		FACTT (N = 849)		VALID (N = 545)		FACTT + VALID (N = 1394)	
	Estimate	P <sup>a</sup>	Estimate	P <sup>a</sup>	Estimate	P <sup>a</sup>	Estimate	P <sup>a</sup>
AUC <sup>b</sup> <sub>Clinical-only model</sub>	0.80		0.72		0.67		0.70	
AUC <sup>c</sup> <sub>Biomarker/clinical model</sub>	0.83	<0.001	0.75	<0.001	0.74	<0.001	0.73	<0.001
NRI <sup>d</sup>	0.59	<0.001	0.44	<0.001	0.45	<0.001	0.41	<0.001
IDI <sup>d</sup>	0.07	<0.001	0.04	<0.001	0.09	<0.001	0.04	<0.001

AUC area under the receive operator characteristic curve, NRI net reclassification improvement, IDI integrated discrimination improvement

<sup>a</sup> P values are from the likelihood ratio tests comparing the refitted biomarker/clinical model to the clinical-only model on the dataset

<sup>b</sup> From refitting the clinical-only model that includes age and APACHE on the dataset

<sup>c</sup> From refitting the biomarker/clinical model that includes age, APACHE, SP-D, and IL-8 on the dataset

<sup>d</sup> Improvement from the clinical model to the biomarker/clinical model

**Table 4 Treatment effects in FACTT stratified by the predicted mortality categories**

Predicted mortality	Treatment effects on VFDs (conservative vs liberal fluid therapy)	Summary of VFDs by treatment <sup>d</sup>	
	RR <sup>a</sup> (95% CI)	Conservative	Liberal fluid
Low <sup>b</sup> (N = 437)	1.05 (0.98, 1.12)	22 (12, 25)	19 (9, 24)
High <sup>b</sup> (N = 412)	1.20 (1.09, 1.33)	9 (0, 22)	1 (0, 19)
P for interaction <sup>c</sup>	0.026		

VFD ventilator-free days

<sup>a</sup> Risk ratio was estimated from zero inflated negative binominal models. RR of 1.2 can be interpreted as compared with the patients who were randomized to liberal fluid therapy, those randomized to conservative fluid therapy had 20% more VFDs

<sup>b</sup> Low, predicted mortality  $\leq 20\%$ ; high, predicted mortality  $> 20\%$ ; median predicted mortality was 19%

<sup>c</sup> P value for the product term of the treatment and predicted risk groups (low vs high)

<sup>d</sup> Median (IQR) were reported

that the original biomarker/clinical model achieved discrimination that was close to optimal in these cohorts. Based on calibration plots, the prediction accuracy of the original biomarker/clinical model was moderate in the validation cohorts. However, after recalibration, the updated model performed well on all three validation cohorts and combined. We also confirmed the original finding that two plasma biomarkers, SP-D and IL-8, added value to clinical predictors in predicting ARDS mortality in the validation cohorts. Finally, despite the overall significant treatment effect of conservative fluid therapy on VFDs originally reported in one of the validation clinical trials (FACTT), we observed a significant treatment effect on VFDs only in the high-risk subgroup as classified by the predicted hospital mortality using the biomarker/clinical model, but not in the low-risk subgroup, indicating that the originally observed treatment effect was confined to a subgroup of the participants. These results illustrate how the prediction model might be used for both prognostic and predictive enrichment for clinical trial enrollment.

Attempts have been made to develop simpler scoring systems for ARDS that are easier to use in clinical practice [19, 20, 23]. However, none have succeeded to date. For example, Brown et al. developed a classification tree for hospital mortality including age, BUN, shock, respiratory rate, and minute ventilation [19]. The tree model is simpler than the widely used APACHE III score, but does not outperform it, and has not been externally validated. Cooke et al. developed a clinical predictive index for mortality including hematocrit, bilirubin, fluid balance, and age [20]. This predictive index, however, failed in the external validation, and the performance was worse than the APACHE III score (AUC 0.68 vs 0.75 respectively,  $P = 0.03$ ). In a Spanish study, Villar et al. developed another scoring system, including age, PaO<sub>2</sub>/FiO<sub>2</sub> ratio, and plateau pressure, termed the APPS [23]. The APPS showed good discrimination ability in both the derivation and internal validation cohort, with an AUC of 0.76 and 0.80, respectively. However, Bos et al. reported that it is likely that the APPS was overfitted to the derivation cohort, since it could not be validated using data collected from two hospitals in the Netherlands [24]. These findings demonstrate that a prediction tool that performs well in derivation and internal validation datasets is not guaranteed to perform well in another population. Thus, an external validation is required before a prediction tool can be generalized to, and applied in, other populations.

In the current study, despite the strong performance of the original published mortality prediction model in its derivation cohort (AUC of 0.83), performance was not as strong in the current study with AUCs of 0.74, 0.72, and 0.73 in FACTT, VALID, or the combined dataset, respectively. The observed drop in discriminative ability from

the model derivation cohort to the validation cohort is a common phenomenon. Several explanations may apply. First, the model may have been overfitted in the derivation cohort. However, on the basis of the calibration plots, this is not likely the case in our study. In STRIVE, the calibration curve is almost parallel to the 45° line, but shifted upwards. This was mainly caused by a higher mortality rate in STRIVE compared to ALVEOLI (32% vs 27%, respectively) and was confirmed by the recalibration curve (e-Fig. 2B). In FACTT and VALID, the majority of the patients had a predicted mortality of less than 40% (x-axis in Fig. 1), among whom, the predicted mortality is close to the observed mortality. The recalibration further improved the prediction accuracy (Fig. 1b, d, f). A second explanation for the drop in AUCs from derivation to validation is the differences in the case mix. As shown in Table 1, the patients included in the validation cohorts are different from those in the derivation cohort with regard to race, APACHE scores, and biomarker values. Despite the heterogeneity of the patients, the AUCs for validation of the original published model in FACTT and VALID (0.74 and 0.72, respectively) are very close to the benchmark AUC, which is the best possible AUC derived by refitting the model on each validation dataset (0.75 and 0.74, respectively). It is also worth noting that, in all three validation cohorts, the benchmark AUC lies within the bootstrap 95% CIs of the AUC from strict validations, further indicating that the biomarker/clinical model may have discrimination power when applied to future datasets. These findings reveal the potential value of the biomarker/clinical model as a prognostic enrichment tool for future clinical trial enrollment.

To our knowledge, the current study is the first to successfully externally validate a prediction model for hospital mortality in ARDS patients across multiple, diverse patient groups. However, our study has some limitations. First, in two of the validation sets, the APACHE II score was recorded, and we used a published formula to estimate the APACHE III score. Although the translation equation was developed in a large study [33], it is possible that it may not accurately reflect the true APACHE III score. Second, some of the biomarker values in the validation sets were not within the range of those in the model derivation cohort. This may cause inaccurate predicted mortality for those with extreme values. However, excluding those participants will decrease the precision of study performance estimation and the usefulness of the developed models. Thus, we decided to include all of the eligible patients. Third, FACTT and STRIVE represent a highly selected subgroup of all patients with ARDS enrolled over 10 years ago that may not be reflective of the general population of current patients with ARDS. This concern is mitigated to some extent by the inclusion

of the VALID cohort, a more recently enrolled and much more heterogeneous and inclusive group of critically ill patients with ARDS compared to clinical trial cohorts. Fourth, the STRIVE study is relatively small, and we were only able to study a subset of the STRIVE patients because of limited plasma availability. However, inclusion of this study does provide additional evidence that the model validates and with more generalizability [40], though the results for this particular study may not be as precise as those from the other two larger cohorts.

## Conclusion

Using three independent patient groups, we found that a published mortality prediction model that combines two clinical variables and plasma biomarkers of two aspects of ARDS pathogenesis (inflammation and lung epithelial injury) could serve as a simple tool for the prediction and stratification of mortality among patient with ARDS. Although the model performance as measured by AUC was lower than in the original model derivation cohort, the biomarker/clinical model still performed well and may be useful for prognostic enrichment for enrollment in clinical trials, an increasingly important issue as mortality in ARDS declines, and better methods are needed for selection of the most severely ill patients for inclusion in clinical trials.

## Electronic supplementary material

The online version of this article (doi:[10.1007/s00134-017-4854-5](https://doi.org/10.1007/s00134-017-4854-5)) contains supplementary material, which is available to authorized users.

## Abbreviations

AEC: American European Consensus Conference; ALI: Acute lung injury; APACHE: Acute physiology and chronic health evaluation; ARDS: Acute respiratory distress syndrome; AUC: Area under receiver operating characteristic curve; CI: Confidence interval; FACTT: Fluid and catheter treatment trial; ICU: Intensive care unit; IDI: Integrated discrimination improvement; NHLBI: National Heart, Lung, and Blood Institute; NRI: Net reclassification improvement; PEEP: Positive end-expiratory pressure; ROC: Receiver operating characteristic curve; RR: Risk ratio; STRIVE: Sivelestat trial in ALI patients requiring mechanical ventilation; VALID: Validating acute lung injury biomarkers for diagnosis; VFD: Ventilator-free days.

## Author details

<sup>1</sup> Department of Biostatistics, Vanderbilt University, Nashville, TN, USA. <sup>2</sup> The Institution for Medicine and Public Health, Vanderbilt University, Nashville, TN, USA. <sup>3</sup> Division of Allergy, Pulmonary and Critical Care Medicine, Department of Medicine, Vanderbilt University, T1218 MCN, 1161 21st Avenue S, Nashville, TN 37232-2650, USA. <sup>4</sup> Division of Hospital Medicine, Department of Medicine, University of California San Francisco, San Francisco, CA, USA. <sup>5</sup> Division of Trauma and Surgical Critical Care, Vanderbilt University, Nashville, TN, USA. <sup>6</sup> Department of Medicine, University of California San Francisco, San Francisco, CA, USA. <sup>7</sup> Department of Anesthesia and Cardiovascular Research Institute, University of California San Francisco, San Francisco, CA, USA. <sup>8</sup> Department of Pathology, Microbiology and Immunology, Vanderbilt University, Nashville, TN, USA.

## Acknowledgements

This manuscript was prepared using FACTT Research Materials obtained from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the FACTT investigators or the NHLBI. We would like to thank the NHLBI BioLINCC/

Biorepository for providing clinical samples and clinical data from the FACTT clinical trial. We also thank Eli Lilly and Company for providing clinical data and plasma samples from the STRIVE study.

## Compliance with ethical standards

### Funding

This study was supported by NIH HL112656 (LBW), 1K23HL116800-01 (KK), HL51856 (MAM), HL110969 and HL131621 (CSC), and the NHLBI BioLINCC/ Biorepository.

### Conflicts of interest

The authors declare that they have no conflict of interest.

### Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

### Informed consent

Informed consent was obtained from all individual participants included in the study with the exception of some patients in the VALID study who were enrolled under an IRB-approved waiver of informed consent.

Received: 13 March 2017 Accepted: 25 May 2017

Published online: 7 June 2017

## References

- Rubinfeld GD, Caldwell E, Peabody E, Weaver J, Martin DP, Neff M, Stern EJ, Hudson LD (2005) Incidence and outcomes of acute lung injury. *N Engl J Med* 353:1685–1693
- Calfee CS, Delucchi K, Parsons PE, Thompson BT, Ware LB, Matthay MA, NHLBI ARDS Network (2014) Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. *Lancet Respir Med* 2:611–620
- Brower RG, Matthay MA, Morris A, Schoenfeld D, Thompson BT, Wheeler A, Acute Respiratory Distress Syndrome Network (2000) Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *N Engl J Med* 342:1301–1308
- Takeda S, Ishizaka A, Fujino Y, Fukuoka T, Nagano O, Yamada Y, Takezawa J, Multicenter Clinical Trial Committee, Japan Society of Respiratory Care Medicine (2005) Time to change diagnostic criteria of ARDS: towards the disease entity-based subgrouping. *Pulm Pharmacol Ther* 18:115–119
- Ospina-Tascon GA, Buchele GL, Vincent JL (2008) Multicenter, randomized, controlled trials evaluating mortality in intensive care: doomed to fail? *Crit Care Med* 36:1311–1322
- Prescott HC, Calfee CS, Thompson BT, Angus DC, Liu VX (2016) Toward smarter lumping and smarter splitting: rethinking strategies for sepsis and acute respiratory distress syndrome clinical trial design. *Am J Respir Crit Care Med* 194:147–155
- Ware LB (2005) Prognostic determinants of acute respiratory distress syndrome in adults: impact on clinical trial design. *Crit Care Med* 33:S217–222
- Brun-Buisson C, Minelli C, Bertolini G, Brazzi L, Pimentel J, Lewandowski K, Bion J, Romand JA, Villar J, Thorsteinsson A, Damas P, Armaganidis A, Lemaire F, Alive Study Group (2004) Epidemiology and outcome of acute lung injury in European intensive care units. Results from the ALIVE study. *Intensive Care Med* 30:51–61
- Doyle RL, Szaflarski N, Modin GW, Wiener-Kronish JP, Matthay MA (1995) Identification of patients with acute lung injury. Predictors of mortality. *Am J Respir Crit Care Med* 152:1818–1824
- Monchi M, Bellenfant F, Cariou A, Joly LM, Thebert D, Laurent I, Dhainaut JF, Brunet F (1998) Early predictive factors of survival in the acute respiratory distress syndrome. A multivariate analysis. *Am J Respir Crit Care Med* 158:1076–1081



11. Nuckton TJ, Alonso JA, Kallet RH, Daniel BM, Pittet JF, Eisner MD, Matthay MA (2002) Pulmonary dead-space fraction as a risk factor for death in the acute respiratory distress syndrome. *N Engl J Med* 346:1281–1286
12. Venet C, Guyomarc'h S, Pingat J, Michard C, Laporte S, Bertrand M, Gery P, Page D, Vermesch R, Bertrand JC, Zeni F (2003) Prognostic factors in acute respiratory distress syndrome: a retrospective multivariate analysis including prone positioning in management strategy. *Intensive Care Med* 29:1435–1441
13. Ware JH (2006) The limitations of risk factors as prognostic tools. *N Engl J Med* 355:2615–2617
14. Ware LB, Matthay MA (2000) The acute respiratory distress syndrome. *N Engl J Med* 342:1334–1349
15. Sapru A, Calfee CS, Liu KD, Kangelaris K, Hansen H, Pawlikowska L, Ware LB, Alkhouli MF, Abbott J, Matthay MA, NHLBI ARDS Network (2015) Plasma soluble thrombomodulin levels are associated with mortality in the acute respiratory distress syndrome. *Intensive Care Med* 41:470–478
16. Geboers DG, de Beer FM, Tuip-de Boer AM, van der Poll T, Horn J, Cremer OL, Bonten MJ, Ong DS, Schultz MJ, Bos LD (2015) Plasma sUPAR as a prognostic biological marker for ICU mortality in ARDS patients. *Intensive Care Med* 41:1281–1290
17. Laffey JG, Bellani G, Pham T, Fan E, Madotto F, Bajwa EK, Brochard L, Clarkson K, Esteban A, Gattinoni L, van Haren F, Heunks LM, Kurahashi K, Laake JH, Larsson A, McAuley DF, McNamee L, Nin N, Qiu H, Ranieri M, Rubenfeld GD, Thompson BT, Wrigge H, Slutsky AS, Pesenti A, LUNG SAFE Investigators and the ESICM Trials Group (2016) Potentially modifiable factors contributing to outcome from acute respiratory distress syndrome: the LUNG SAFE study. *Intensive Care Med* 42:1865–1876
18. Cooke CR, Kahn JM, Caldwell E, Okamoto VN, Heckbert SR, Hudson LD, Rubenfeld GD (2008) Predictors of hospital mortality in a population-based cohort of patients with acute lung injury. *Crit Care Med* 36:1412–1420
19. Brown LM, Calfee CS, Matthay MA, Brower RG, Thompson BT, Checkley W, The National Institutes of Health Acute Respiratory Distress Syndrome Network (2011) A simple classification model for hospital mortality in patients with acute lung injury managed with lung protective ventilation. *Crit Care Med* 39:2645–2651
20. Cooke CR, Shah CV, Gallop R, Bellamy S, Ancukiewicz M, Eisner MD, Lanken PN, Localio AR, Christie JD, National Heart Lung Blood Institute Acute Respiratory Distress Syndrome Network (2009) A simple clinical predictive index for objective estimates of mortality in acute lung injury. *Crit Care Med* 37:1913–1920
21. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, Sirio CA, Murphy DJ, Lotring T, Damiano A et al (1991) The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* 100:1619–1636
22. Moreno RP, Metnitz PG, Metnitz B, Bauer P, Afonso de Carvalho S, Hoechst A, SAPS 3 Investigators (2008) Modeling in-hospital patient survival during the first 28 days after intensive care unit admission: a prognostic model for clinical trials in general critically ill patients. *J Crit Care* 23:339–348
23. Villar J, Ambros A, Soler JA, Martinez D, Ferrando C, Solano R, Mosteiro F, Blanco J, Martin-Rodriguez C, Fernandez MM, Lopez J, Diaz-Dominguez FJ, Andaluz-Ojeda D, Merayo E, Perez-Mendez L, Fernandez RL, Kacmarek RM, Stratification Outcome of Acute Respiratory Distress Syndrome Network (2016) Age, PaO<sub>2</sub>/FIO<sub>2</sub>, and plateau pressure score: a proposal for a simple outcome score in patients with the acute respiratory distress syndrome. *Crit Care Med* 44:1361–1369
24. Bos LD, Schouten LR, Cremer OL, Ong DS, Schultz MJ, MARS Consortium (2016) External validation of the APPS, a new and simple outcome prediction score in patients with the acute respiratory distress syndrome. *Ann Intensive Care* 6:89
25. Ware LB, Koyama T, Billheimer DD, Wu W, Bernard GR, Thompson BT, Brower RG, Standiford TJ, Martin TR, Matthay MA, NHLBI ARDS Clinical Trials Network (2010) Prognostic and pathogenetic value of combining clinical and biochemical indices in patients with acute lung injury. *Chest* 137:288–296
26. Ware LB, Zhao Z, Koyama T, Wickersham N, Kangelaris KN, Bernard GR, Matthay MA, Calfee CS (2015) Validation of a biomarker and clinical prediction model for mortality in ARDS in three patient cohorts. *Am J Respir Crit Care Med* 191:A6249
27. Brower RG, Lanken PN, MacIntyre N, Matthay MA, Morris A, Ancukiewicz M, Schoenfeld D, Thompson BT, The National Heart Lung Blood Institute Ards Clinical Trials Network (2004) Higher versus lower positive end-expiratory pressures in patients with the acute respiratory distress syndrome. *N Engl J Med* 351:327–336
28. Wiedemann HP, Wheeler AP, Bernard GR, Thompson BT, Hayden D, deBoisblanc B, Connors AF Jr, Hite RD, Harabin AL, The National Heart Lung and Blood Institute Acute Respiratory Distress Syndrome Clinical Trials Network (2006) Comparison of two fluid-management strategies in acute lung injury. *N Engl J Med* 354:2564–2575
29. Zeiher BG, Artigas A, Vincent JL, Dmitrienko A, Jackson K, Thompson BT, Bernard G, Strive Study Group (2004) Neutrophil elastase inhibition in acute lung injury: results of the STRIVE study. *Crit Care Med* 32:1695–1702
30. O'Neal HR Jr, Koyama T, Koehler EA, Siew E, Curtis BR, Fremont RD, May AK, Bernard GR, Ware LB (2011) Prehospital statin and aspirin use and the prevalence of severe sepsis and acute lung injury/acute respiratory distress syndrome. *Crit Care Med* 39:1343–1350
31. Bernard GR, Artigas A, Brigham KL, Carlet J, Falke K, Hudson L, Lamy M, Legall JR, Spragg R (1994) The American-European Consensus Conference on ARDS. Definitions, mechanisms, relevant outcomes, and clinical trial coordination. *Am J Respir Crit Care Med* 149:818–824
32. ARDS Definition Task Force, Ranieri VM, Rubenfeld GD, Thompson BT, Ferguson ND, Caldwell E, Fan E, Camporota L, Slutsky AS (2012) Acute respiratory distress syndrome: the Berlin definition. *JAMA* 307:2526–2533
33. Schneider AG, Lipcsey M, Bailey M, Pilcher DV, Bellomo R (2013) Simple translational equations to compare illness severity scores in intensive care trials. *J Crit Care* 28(885):e881–888
34. Harrell FE Jr, Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 15:361–387
35. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y (2008) Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol* 61:76–86
36. Sigurdsson MI, Sigvaldason K, Gunnarsson TS, Moller A, Sigurdsson GH (2013) Acute respiratory distress syndrome: nationwide changes in incidence, treatment and mortality over 23 years. *Acta Anaesthesiol Scand* 57:37–45
37. Frank AJ, Thompson BT (2010) Pharmacological treatments for acute respiratory distress syndrome. *Curr Opin Crit Care* 16:62–68
38. Duggal A, Ganapathy A, Ratnapalan M, Adhikari NK (2015) Pharmacological treatments for acute respiratory distress syndrome: systematic review. *Miner Anesthesiol* 81:567–588
39. Famous KR, Delucchi K, Ware LB, Kangelaris KN, Liu KD, Thompson BT, Calfee CS, ARDS Network (2017) Acute respiratory distress syndrome subphenotypes respond differently to randomized fluid management strategy. *Am J Respir Crit Care Med* 195:331–338
40. Taylor JM, Ankerst DP, Andridge RR (2008) Validation of biomarker-based risk prediction models. *Clin Cancer Res* 14:5977–5983