Journal of Intensive Care Medicine
2020, Vol. 35(6) 595-605

DOI: 10.1177/0885066618771001
journals.sagepub.com/home/jic


Early Prediction of Intensive Care Unit–Acquired Weakness: A Multicenter External Validation Study

Esther Witteveen, MD, PhD[1,2,3], Luuk Wieske, MD, PhD[1,2,3], Juultje Sommers, MSc[4], Jan-Jaap Spijkstra, MD, PhD[5], Monique C. de Waard, PhD[5], Henrik Endeman, MD, PhD[6], Saskia Rijkenberg, RN, MSc[6], Wouter de Ruijter, MD, PhD[7], Mengalvio Sleeswijk, MD[8], Camiel Verhamme, MD, PhD[2], Marcus J. Schultz, MD, PhD[1,3], Ivo N. van Schaik, MD, PhD[2], and Janneke Horn, MD, PhD[1,3]

Abstract

Objectives: An early diagnosis of intensive care unit–acquired weakness (ICU-AW) is often not possible due to impaired consciousness. To avoid a diagnostic delay, we previously developed a prediction model, based on single-center data from 212 patients (development cohort), to predict ICU-AW at 2 days after ICU admission. The objective of this study was to investigate the external validity of the original prediction model in a new, multicenter cohort and, if necessary, to update the model. **Methods:** Newly admitted ICU patients who were mechanically ventilated at 48 hours after ICU admission were included. Predictors were prospectively recorded, and the outcome ICU-AW was defined by an average Medical Research Council score <4. In the validation cohort, consisting of 349 patients, we analyzed performance of the original prediction model by assessment of calibration and discrimination. Additionally, we updated the model in this validation cohort. Finally, we evaluated a new prediction model based on all patients of the development and validation cohort. **Results:** Of 349 analyzed patients in the validation cohort, 190 (54%) developed ICU-AW. Both model calibration and discrimination of the original model were poor in the validation cohort. The area under the receiver operating characteristics curve (AUC-ROC) was 0.60 (95% confidence interval [CI]: 0.54-0.66). Model updating methods improved calibration but not discrimination. The new prediction model, based on all patients of the development and validation cohort (total of 536 patients) had a fair discrimination, AUC-ROC: 0.70 (95% CI: 0.66-0.75). **Conclusions:** The previously developed prediction model for ICU-AW showed poor performance in a new independent multicenter validation cohort. Model updating methods improved calibration but not discrimination. The newly derived prediction model showed fair discrimination. This indicates that early prediction of ICU-AW is still challenging and needs further attention.

Keywords

ICU–acquired weakness, prediction, prediction model, predictors, model validation, external validation

[1] Department of Intensive Care Medicine, Academic Medical Center (AMC), Amsterdam, the Netherlands
[2] Department of Neurology, Academic Medical Center (AMC), Amsterdam, the Netherlands
[3] Laboratory of Experimental Intensive Care and Anesthesiology (LEICA), Academic Medical Center (AMC), Amsterdam, the Netherlands
[4] Department of Rehabilitation, Academic Medical Center (AMC), Amsterdam, the Netherlands
[5] Department of Intensive Care Medicine, VU medical center (VUmc), Amsterdam, the Netherlands
[6] Department of Intensive Care Medicine, OLVG, Amsterdam, the Netherlands
[7] Department of Intensive Care Medicine, Noordwest Ziekenhuisgroep, Alkmaar, the Netherlands
[8] Department of Intensive Care Medicine, Flevoziekenhuis, Almere, the Netherlands

Received January 24, 2018. Received revised March 21, 2018. Accepted March 26, 2018.

Corresponding Author:
Esther Witteveen, Department of Neurology, Academic Medical Center, Meibergdreef 9, 1105 AZ Amsterdam, the Netherlands.
Email: e.witteveen@amc.nl

Introduction

Intensive care unit–acquired weakness (ICU-AW) is a frequent complication of critical illness and is associated with prolonged stay in the ICU and increased short- and long-term morbidity and mortality.[1-3] Before structural muscle and nerve damage is detectable, muscle and nerve dysfunction occurs, which may be fully reversible.[4-6] Electrophysiological signs of critical illness, polyneuropathy or myopathy, can often be detected within the first week after ICU admission and may resolve before ICU discharge.[6] Therefore, future treatments may be most beneficial early in the disease. Furthermore, the benefits of being able to predict the development of ICU-AW in a given patient may allow for the more timely initiation of supportive interventions, such as early mobilization.[7,8]

Intensive care unit–acquired weakness is currently diagnosed by assessment of manual muscle strength.[9] This is often not possible in the first couple of days after ICU admission because of impaired consciousness or attentiveness, for example, due to delirium, coma, or sedation.[10-12] To avoid this diagnostic delay, we previously developed a prediction model for ICU-AW, including 3 early available predictors obtained 2 days after ICU admission.[13] The model showed fair discriminative performance after internal validation but was built on data collected in only 1 hospital.

Before prediction models can be applied in practice, the external validity should be studied in a new independent population.[14] The aim of this study was to externally validate and, if necessary, update the previously developed prediction model for ICU-AW. External validation included both temporal (patients from a later time period) and geographical (patients from other institutions) validation to assess generalizability of the model.

Methods

Design and Ethical Approval

We performed a multicenter prospective observational cohort validation study. This study was reported according to the recently published TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) guidelines.[15,16]

The institutional review board of the Academic Medical Center, Amsterdam, the Netherlands, decided that the Medical Research Involving Human Subjects Act does not apply to this study (decision notice W13_193#13.17.0239), and therefore, written informed consent was not needed. Verbal consent to use patient data was obtained from all included patients. The study was registered in the Netherlands Trial Register (#NTR4331).

Study Setting

The study was conducted in medical-surgical ICUs of 5 hospitals in the Netherlands: 2 university hospitals, 2 university affiliated teaching hospitals, and 1 regional hospital.

Inclusion and Exclusion Criteria

Consecutive, newly admitted ICU patients, $\geq$18 years old, mechanically ventilated at 48 hours, after ICU admission, were included (irrespective of the duration of mechanical ventilation). This was different from the development study where patients who were mechanically ventilated for >2 days after admission were included. As in the development study, we excluded patients with an admission diagnosis of cardiac arrest, neuromuscular disease, or central nervous system (CNS) disease (stroke, traumatic brain or spinal cord injury, CNS infection, or CNS tumor). Furthermore, patients with preexisting spinal injury, a poor pre-ICU functional status (modified Rankin scale $\geq$4),[17] and patients who were expected to die within 48 hours were excluded.

Predictor Assessment

All 20 candidate predictors of the development study[13] were assessed. These predictors were defined, collected, and interpreted as in the model development study, except for lowest Pao_2/Fio_2 (P/F) ratio which was defined by the lowest Pao_2 in the first 48 hours divided by the Fio_2 on the concurrent time point (instead of the lowest of all P/F ratios in the development cohort).

Additionally, 3 new candidate predictors, based on newly described risk factors, were collected: erythrocyte transfusion,[18] hypercalcemia,[19] and hypophosphatemia (own data, not published). These were defined as any erythrocyte transfusion within 24 hours before ICU admission or in the first 48 hours after ICU admission, highest ionized calcium (mmol/L), and lowest phosphate (mmol/L) in the first 48 hours after ICU admission, respectively. All predictors were prospectively assessed and recorded in an online case report form by local investigators, blinded for the strength assessment results.

Strength Assessment (Reference Standard)

As in the development study, trained physiotherapists assessed muscle strength as soon as patients were alert (Richmond Agitation and Sedation Scale [RASS] between -1 and 1) and attentive (able to follow verbal commands using facial expressions).[12,20,21] Muscle strength was assessed using the Medical Research Council (MRC) score in 6 prespecified muscle groups, as in the development study.[13,22] The average MRC score of these muscle groups was used for the analysis (values were not imputed when a muscle group could not be assessed). Intensive care unit–acquired weakness was defined by an average MRC score <4, in accordance with international consensus statements.[1,9] Physiotherapists were blinded for the predictors (except age, gender, and admission reason).

Additional Data Collected

We additionally collected the following clinical characteristics: the Acute Physiology and Chronic Health Evaluation IV (APACHE IV) score, the maximal Sequential Organ

Failure Assessment (SOFA) score of the first 2 days after ICU admission, day of MRC assessment, number of days on mechanical ventilation, length of stay in the ICU, and ICU mortality.

Data Analysis

We applied the original model, with its predictors and assigned weights as estimated in the development study,[13] to our new data. The original model was:

$$P_{ICUAW} = \frac{e^{-2.7763 + 0.0212 \times \text{Age} + 0.7324 \times \text{Highest Lactate} + 0.9506 \times \text{Treatment any aminoglycoside (=yes)}}}{1 + e^{-2.7763 + 0.0212 \times \text{Age} + 0.7324 \times \text{Highest Lactate} + 0.9506 \times \text{Treatment any aminoglycoside (=yes)}}}.$$

We assessed the performance by calibration and discrimination. Calibration reflects the agreement between the predicted ICU-AW risk by the model and the observed ICU-AW frequency in the validation cohort. This was assessed for each decile of predicted risk, ensuring 10 equally sized groups, by calculating the ratio of predicted ICU-AW risk to observed ICU-AW frequency. Calibration was analyzed graphically and using goodness of fit (Hosmer-Lemeshow test). Discrimination, the ability of the test to correctly classify those with and without the disease, was assessed by the area under the receiver operating characteristic curve (AUC-ROC). We defined AUC-ROC values between 0.90 and 1 as excellent, 0.80 and 0.90 as good, 0.70 and 0.80 as fair, 0.60 and 0.70 as poor, and <0.60 as failed.

Next, to improve the performance of the original model, we used updating methods, which combine the information that is captured in the original model with the information of the new patients, instead of making a whole new prediction model.

The previously described updating methods[23] vary from simple recalibration (reestimation of the intercept or slope of the linear predictor) to more extensive revisions, like reestimation of some or all regression coefficients and model extension with new predictors. Before stepwise addition to the model, distributions of the 3 new candidate predictors were checked for normality. The AUC-ROCs of the updated models were calculated. The change in Akaike information criterion (AIC) between the updated models and the recalibrated model was compared.[24] A model in which the AIC was at least 2 points lower than the AIC of the recalibrated model was considered an improved model. In this improved model, the reestimated predictors were shrinked toward the recalibrated model, any new predictors were shrinked toward zero, and the intercept was again determined.

To further assess improved discrimination, we evaluated the degree of correct reclassification using the continuous net reclassification improvement (cNRI), which is more sensitive to change than the AUC-ROC.[25] The cNRI of the updated model was compared to the recalibrated model. We also assessed the cNRI of the APACHE IV score and maximal SOFA score in the first 2 ICU days.

As a sensitivity analysis to assess the influence of missing data, we examined calibration and discrimination in data sets in which missing data were imputed, using multivariate imputation by chained equations (10 iterations of 10 imputations).[26] All predictors and the outcome (ICU-AW) were used for the imputation model. We checked validity of imputed data. The AUC-ROC and the corresponding confidence intervals (CIs) of the 10 imputed data sets were averaged using Rubin's rules, a method to take into account variation within and between multiple imputation data sets.[27]

A second sensitivity analysis assessed the influence of the difference in inclusion criteria between the development study and the validation study. We repeated the analysis for patients who were 2 days mechanically ventilated at time of inclusion. As an additional sensitivity analysis, we assessed the performance of the model in only the patients of the hospital in which the model was developed.

Furthermore, we used the combined data from the development and validation cohort to make a new prediction model. Predictor selection was done as comprehensively described in the development study.[13] In short, we used bootstrapped backward selection and selected those predictors who were selected in >50% of the bootstrap samples (n = 1000; $P < .05$). Calibration and discrimination were assessed.

Proportions are presented with percentages and total numbers, mean values with standard deviation, and median values with interquartile range. Differences between proportions were assessed using χ^2 test, differences between normally distributed variables using Welch's t test, and differences between nonnormally distributed continuous variables using Wilcoxon rank sum test. Test results are presented with corresponding 95% CI. Analyses were done using R (version: 3.3.1).

Power Calculation

Empirical evidence suggests a minimum of 100 events and 100 nonevents for external validation studies.[28] With an incidence of ICU-AW of about 50%, at least 200 patients were needed for validation (and updating). To further validate an updated model, 200 additional patients would be needed. We aimed to include at least 500 patients to account for people in whom MRC measurements could not be performed (ie, because they died before it could be measured, had an ongoing delirium, etc).

Results

Screened and Included Patients

Figure 1 displays the flow chart. Consecutive ICU patients were screened for inclusion from February 2014 to December 2015. A total of 538 patients fulfilled the inclusion criteria and did not meet exclusion criteria. In 349 (65%) patients, muscle strength

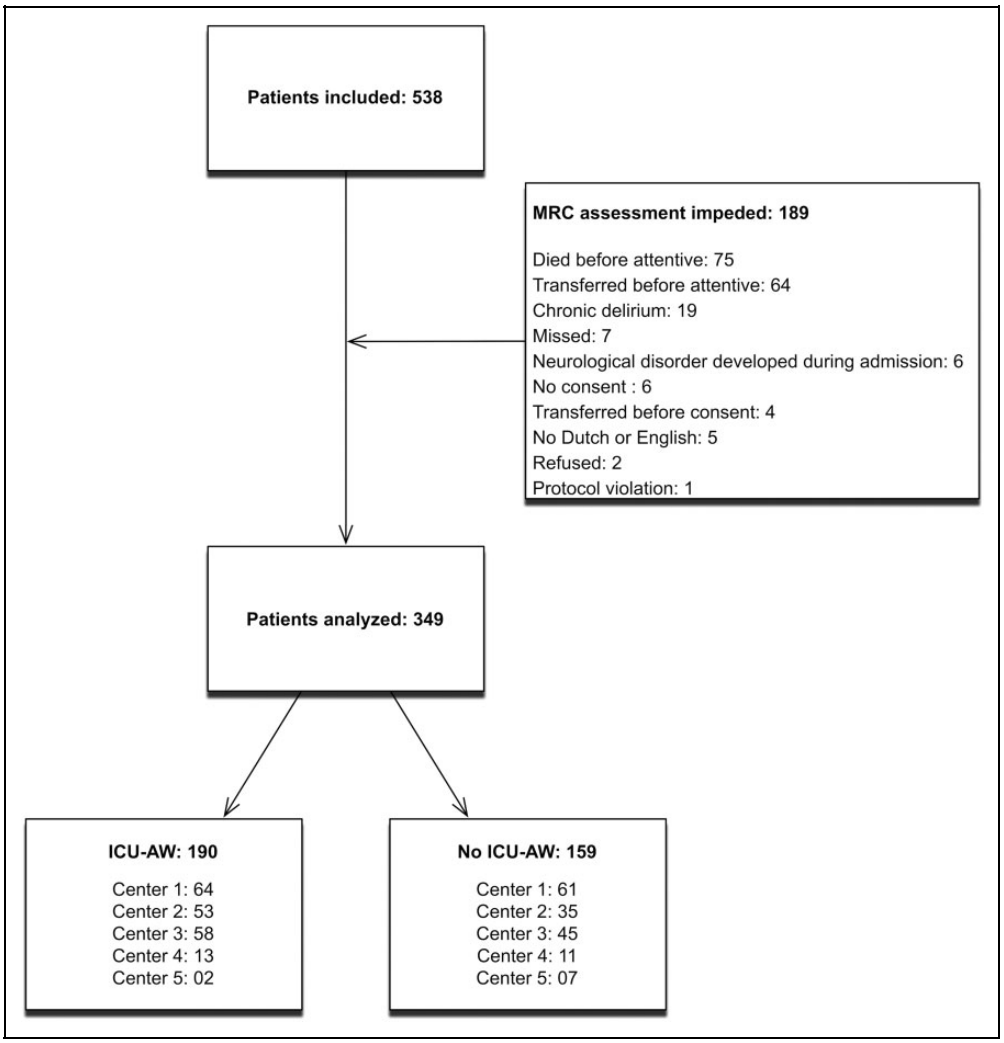

Figure 1. Flowchart of screened and included patients. Center 1 is the center in which the original model was developed. ICU-AW indicates intensive care unit–acquired weakness; MRC, Medical Research Council.

could be assessed; 190 (54%) patients were classified as having ICU-AW. Unfortunately, loss to follow-up was larger than expected. We decided to deviate from the analysis plan and chose to only validate and update the model. This meant that no separate cohort of patients was left for validation of the updated model.

Relatedness Between the Development and External Validation Cohorts

Table 1 shows the study and patient characteristics of the development and validation study. Table 2 shows the distribution of the assessed predictors of the development and external validation cohort.

Performance of the Original Model in the Validation Cohort

The original model was applied to our validation cohort. Calibration was poor with evidence for lack of fit (Figure 2). The predictions were too extreme: for low predicted probabilities by the model, the true fraction with ICU-AW was higher; and for high predicted probabilities, the true fraction was lower. The AUC-ROC was 0.60 (95% CI: 0.54-0.66), which is interpreted as poor discrimination.

Model Updating

We tried several methods to update our model (Table 3 and Figure 3) using recalibration, reestimation, and extension with new candidate predictors. Model updating, using method 6 in which the new candidate predictors were added one-by-one to the recalibrated model (method 3), improved discrimination when lowest phosphate was added to the model. With all updating methods, calibration improved, but the AUC-ROC remained 0.60 (95% CI: 0.54-0.66). The cNRI of the updated model was as good as the cNRI of the recalibrated model.

Table 1. Study and Patient Characteristics.

Characteristic	Development Cohort, N = 212	External Validation Cohort, N = 349	*P* Value
Data collection period	January 2011 to December 2012	February 2014 to December 2015	
Study design	Prospective observational cohort	Prospective observational cohort	
Setting	Mixed medical-surgical ICU of 1 academic medical center in the Netherlands	Mixed medical-surgical ICUs of 5 hospitals in the Netherlands	
Inclusion criteria	Consecutive, newly admitted ICU patients mechanically ventilated for ≥2 days	Consecutive, newly admitted ICU patients mechanically ventilated at 48 hours after ICU admission	
Outcome	Presence of ICU-AW	Presence of ICU-AW	
Reference standard	Average MRC score<4	Average MRC score < 4	
Incidence of ICU-AW, n (%)	103 (49)	190 (55)	.208
Age, mean (SD)	61 (16)	63 (14)	.050
Females, n (%)	92 (43)	136 (39)	.347
Reason for admission			
Planned surgical, n (%)	44 (21)	72 (21)	.994
Emergency surgical, n (%)	49 (23)	85 (24)	
Medical, n (%)	119 (56)	192 (55)	
APACHE IV score, mean (SD)	81 (28), 3 missing	79 (27), 16 missing	.272
Maximal SOFA score in first 2 days, mean (SD)	10 (3)	9 (3), 12 missing	.013
Average MRC score, median (IQR)	4.0 (2.6-4.8)	3.8 (3.2-4.5)	.834
Day of MRC assessment after ICU admission, median (IQR)	8 (6-12)	6 (4-10)	<.001
Days with MV, median days (IQR)	8 (4-16)	6 (4-11)	.001
LOS ICU, median days (IQR)	10 (7-8)	9 (6-17)	.107
ICU mortality, n (%)	21 (10)	25 (7)	.269

Abbreviations: APACHE IV: Acute Physiology and Chronic Health Evaluation IV; ICU-AW, intensive care unit–acquired weakness; IQR, interquartile range; LOS ICU, length of stay in the intensive care unit; MRC, Medical Research Council; MV, mechanical ventilation; SD, standard deviation; SOFA, Sequential Organ Failure Assessment.

Comparison With SOFA and APACHE IV Scores

The AUC-ROC of the maximal SOFA score in the first 2 days after admission for prediction of ICU-AW in the validation cohort was 0.63 (95% CI: 0.58-0.69), and the AUC-ROC of the APACHE IV score was 0.63 (95% CI: 0.57-0.69). Compared to using the SOFA score, the updated model reduced classification with 31% (cNRI; 95% CI: 9-52), whereas it performed as good as the APACHE IV score (21% [95% CI: 1-43]).

Sensitivity Analysis

Of the predictors used in external validation and updating analyses, highest lactate levels were missing in 2 patients and lowest phosphate in 8 patients; the other predictors included in the original model did not have missing values. The combined AUC-ROC of the imputed data sets was 0.59 (95% CI: 0.53-0.65).

When the original model was only applied to patients who were mechanically ventilated for 2 days at the time of inclusion (n = 291), the AUC-ROC was 0.58 (95% CI: 0.51-0.64) and when it was applied to the patients in the hospital of the development study (center 1, n = 123), the AUC-ROC was 0.59 (95% CI: 0.49-0.69).

New Prediction Model

The following predictors were included in >50% of the bootstrap samples: RASS score, gender, highest lactate, lowest P/F ratio, highest glucose, and ICU treatment with corticosteroids. In the final model (based on 536 patients due to missing values of RASS score [n = 6] and lactate [n = 19]), RASS, gender, highest lactate, and treatment with corticosteroids were included (selected by a drop in AIC >2). A universal shrinkage factor (0.94) was applied to adjust for overfitting. The new prediction model is described with the following formula:

$$P_{ICUAW} = \frac{e^{-1.5724 - 0.2233 \times \text{RASS SCORE} + 0.5699 \times \text{gender (female = 1)} + 0.5107 \times \text{Highest Lactate} + 0.4631 \times \text{Treatment corticosteroids (=yes)}}}{1 + e^{-1.5724 - 0.2233 \times \text{RASS SCORE} + 0.5699 \times \text{gender (female = 1)} + 0.5107 \times \text{Highest Lactate} \times 0.4631 \times \text{Treatment corticosteroids (=yes)}}}.$$

Table 2. Distributions of Candidate Predictors.[a]

Predictors	Development Cohort, (n = 212)	External Validation Cohort, (n = 349)	*P* Value
Patient characteristics			
Females, n (%)	92 (43)	136 (39)	.344
Age, mean (SD)	*61 (16)*	*63 (14)*	*.053*
Risk factor for a polyneuropathy in medical history, n (%)	75 (35)	147 (43), 9 missing	.082
Preexisting polyneuropathy prior to ICU admission, n (%)	4 (2)	11 (3), 18 missing	.467
Systemic corticosteroid use prior to ICU admission, n (%)	16 (8)	25 (7), 12 missing	1.000
Clinical parameters			
Suspected sepsis, n (%)	148 (70)	199 (57)	.003
Unplanned admission, n (%)	168 (79)	277 (79)	1.000
Presence of shock, n (%)	142 (67)	222 (64)	.472
RASS score, median (IQR)	−3 (−4 to 0)	−2 (−4 to −1), 6 missing	.388
Laboratory parameters			
Average urine production, median, mL/h (IQR)	87 (40 to 128)	64 (41 to 98)	.002
Highest glucose, mean (SD), mg/dL	231.8 (73.7)	219.3 (63.9)	.034
Lowest glucose, mean (SD), mg/dL	87.8 (24.2)	103.5 (24.4)	<.001
Lowest pH, mean (SD)	7.23 (0.10)	7.23 (0.11)	.790
Lowest P/F ratio, median (IQR)	180 (129 to 246)	144 (96 to 200), 1 missing	<.001
Lowest platelet count, median, $\times 10^9$/L (IQR)	118 (66 to 173)	150 (83 to 221), 5 missing	<.001
Highest lactate, median, mmol/L (IQR)	*3.7 (2.2 to 6.0), 17 missing*	*3.3 (2.1 to 5.2), 2 missing*	*.087*
Lowest ionized Ca^{2+}, mean (SD), mmol/L	0.98 (0.12)	1.03 (0.14)	<.001
Highest ionized Ca^{2+}, mean (SD), mmol/L		1.22 (0.12)	
Highest phosphate, mean (SD), mmol/L		0.89 (0.37), 8 missing	
Treatment			
Treatment with any corticosteroid, n (%)	144 (68)	244 (69.9)	.689
Repeated treatment with any neuromuscular blocker, n (%)	35 (17)	33 (9.5)	.019
Treatment with any aminoglycoside, n (%)	*81 (38)*	*48 (13.8)*	*<.001*
Transfusion of erythrocytes, n (%)		132 (37.8)	

Abbreviations: Ca, calcium; ICU, intensive care unit; IQR, interquartile range; P/F, Pao_2/Fio_2; RASS, Richmond Agitation and Sedation Scale; SD, standard deviation.
[a]The predictors in italic are the predictors included in the original prediction model.

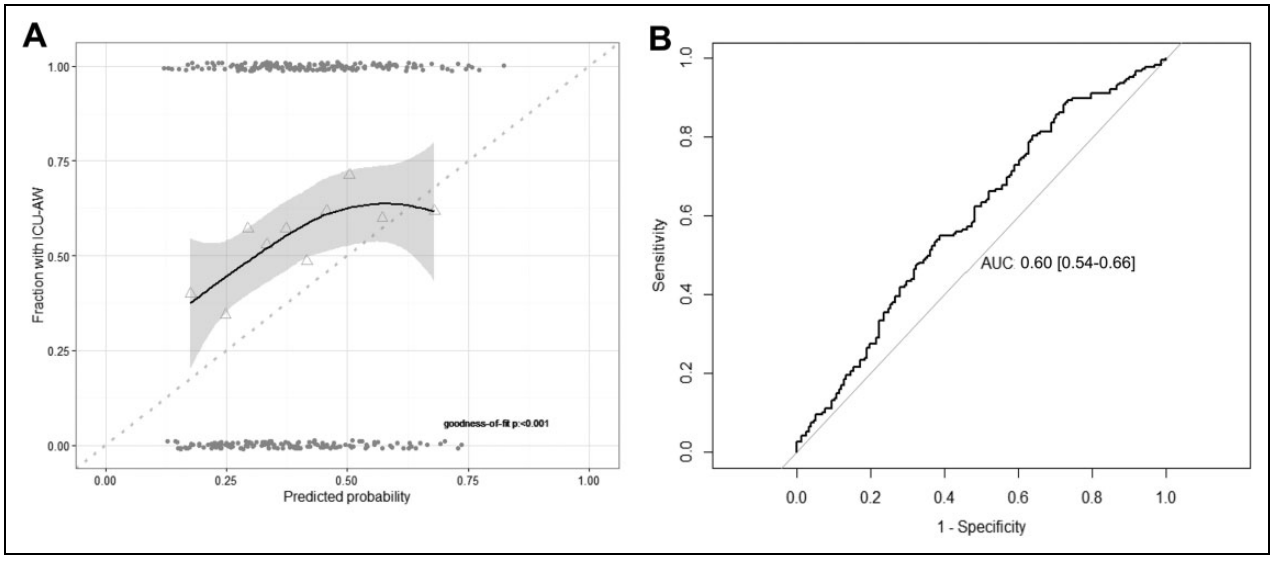

Figure 2. Model performance: calibration and discrimination of original model. A, The model calibration assessed with a fitted curve based on Loess regression with 95% confidence interval. Perfect calibration is illustrated by the dotted line. Triangles represent deciles of predicted probability and grey points represent predicted probabilities of individual patients. Goodness of fit was assessed with the Hosmer-Lemeshow test. B, Model discrimination assessed with the receiver operating characteristic curve. AUC, area under the curve; ICU-AW indicates intensive care unit–acquired weakness.

Table 3. Model Updating Results.[a,b]

	No Updating	Recalibration		Model Revision		Model Extension			
	Method 1: Original Model	Method 2: Update Intercept	Method 3: Recalibration of Intercept and Slope	Method 4: Recalibration and Selective Reestimation	Method 5: Reestimation	Method 6: Recalibration (and Selective Reestimation) and Selective Reextension	Method 7: Reestimation and Selective Extension	Method 8: Re-estimation and Extension	Shrinkage Model 6
Intercept	−2.776	−2.154	−1.049	−1.049	−0.849	−1.361	−1.119	−1.027	−1.115
Age	0.021[c]	0.021	0.011	0.011	0.006	0.011	0.004	0.004	0.011[d]
Highest lactate[e]	0.732[c]	0.732	0.384	0.384	0.508	0.384	0.537	0.515	0.384[d]
Aminoglycoside	0.951[c]	0.951	0.498	0.498	0.275	0.498	0.172	0.175	0.498[d]
Lowest phosphate						0.347	0.403	0.384	0.070[f]
Erythrocyte transfusion								0.083	
Highest calcium								−0.06	
Hosmer-Lemeshow test	<0.001	0.038	0.550	0.550	0.739	0.265	0.837	0.549	0.208
AUC-ROC	0.60 (0.54-0.66)	0.60 (0.54-0.66)	0.60 (0.54-0.66)	0.60 (0.54-0.66)	0.60 (0.54-0.66)	0.60 (0.54-0.66)	0.61 (0.55-0.67)	0.61 (0.55-0.67)	0.60 (0.54-0.66)

Abbreviation: AUC-ROC, area under the receiver operating characteristic curve.

[a]Method 1 is the original model. The model was recalibrated by adjusting only the intercept (method 2) or both the intercept and slope (method 3). With method 4, we investigated whether predictors were having a clearly different effect in the validation cohort, by selective reestimation of one or more of the included predictors. None of the models with reestimations improved the model; therefore, no selective reestimations were done. In method 5, the model was fitted in the validation data by reestimation of the intercept and regression coefficients for all predictors. In method 6, the 3 new predictors were one-by-one added to the recalibrated model. Only adding lowest phosphate improved the model. In method 7, model 5 was extended with new predictors. In method 8, a model with all old and new predictors was assessed.

[b]Shrinkage was applied to the improved model (method 6), and the intercept was recalculated.

[c]Uniform shrinkage factor applied.

[d]Shrinkage toward recalibrated values.

[e]Transformed using the natural logarithm.

[f]Shrinkage toward zero.

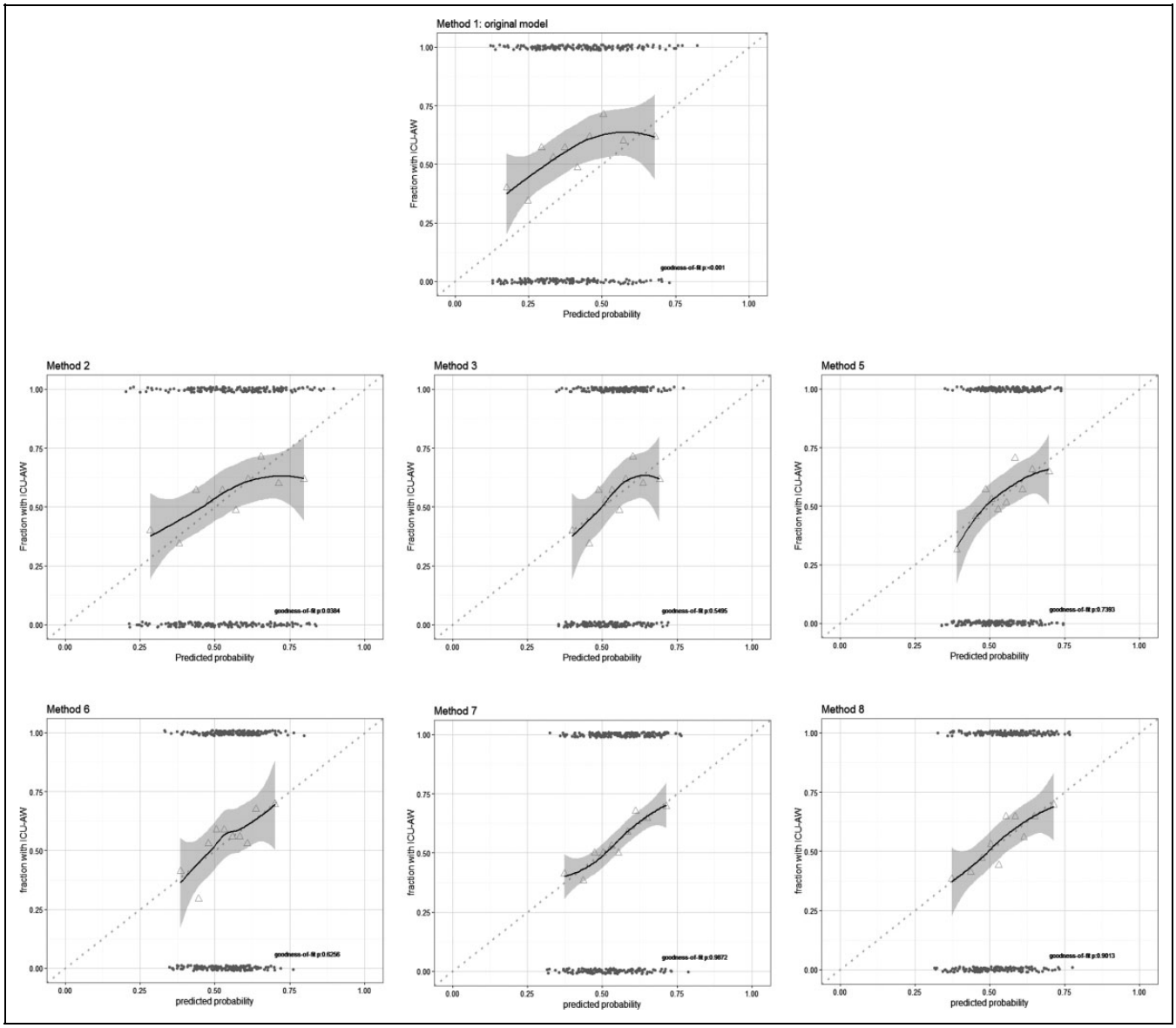

Figure 3. Calibration plots of updated models. Model calibration of the updated models from Table 3 were assessed with a fitted curve based on Loess regression with 95% confidence interval. Perfect calibration is illustrated by the dotted line. Triangles represent deciles of predicted probability and grey points represent predicted probabilities of individual patients. Goodness of fit was assessed with the Hosmer-Lemeshow test.

Calibration was excellent (Figure 4). The AUC-ROC after internal validation was 0.70 (95% CI: 0.66-0.75). Discrimination improved when using the new prediction model compared to the SOFA or APACHE IV score (cNRI: 38% [95% CI: 21-55] and 30% [95% CI: 13-47], respectively).

Discussion

In this study, we assessed the performance of a previously developed prediction model for ICU-AW. The model showed both poor calibration and discrimination in our new patient cohort. Updating methods improved calibration but not discrimination. A new prediction model based on combined data from the development and validation cohort had an excellent calibration. The AUC-ROC of this new model was 0.70 after internal validation. The new prediction model classified patients better than the SOFA and APACHE IV scores.

Reasons for Poor Performance of the Model

Poor performance in new data sets is often seen and can have several reasons.[16] First of all, this can be caused by differences in case mix. The distribution of baseline characteristics and predictors showed differences between the development and the validation cohort. Patients in the validation cohort seemed to be less severely ill indicated by lower SOFA scores, less often sepsis, less days of mechanical ventilation, less repeated administration of neuromuscular blockers, and earlier MRC

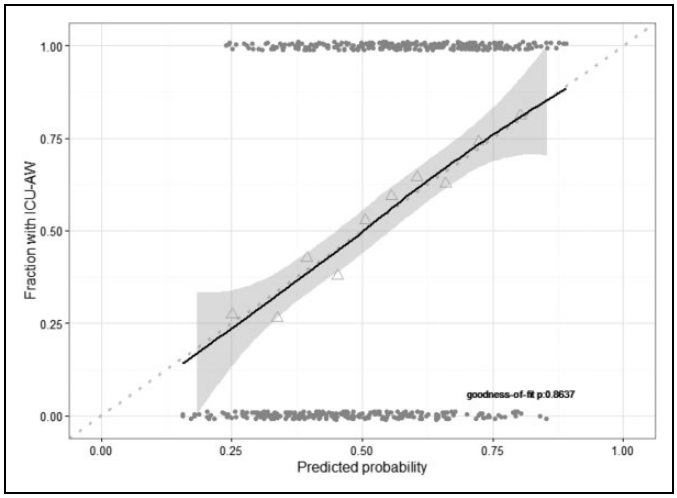

Figure 4. Calibration plot of new model. Calibration plot of new model based on combined data of the development and validation cohort. Model calibration was assessed with a fitted curve based on Loess regression with 95% confidence interval. Perfect calibration is illustrated by the dotted line. Triangles represent deciles of predicted probability and grey points represent predicted probabilities of individual patients. Goodness of fit was assessed with the Hosmer-Lemeshow test. ICU-AW indicates intensive care unit–acquired weakness.

assessment; whereas, on the other hand, these patients had less urine production and lower P/F ratios. A major difference in use of aminoglycosides was seen, because it was regularly used in the center in which the model was developed but less frequent in the other centers. The differences in case mix cannot solely be explained by the multicenter design since these differences were also seen when the development cohort was compared with validation cohort patients only from the hospital in which the model was developed. Although no major changes in standard of care were noted, unrecognized changes in care over time may cause differences in case-mix, which could be a reason for failed temporal validation.

Besides differences in subject-level characteristics as described previously, differences in study-level characteristics (such as inclusion criteria) can also lead to a worse performance. Our inclusion criteria in the validation cohort differed slightly from the development cohort, possibly selecting less severe patients because some patients had a lower duration of mechanical ventilation in the first 48 hours after admission. We chose this inclusion criterion to make inclusion more easy for the investigators and to increase the amount of eligible patients. We assumed that this population would be comparable to the population in the development study. Actually, sensitivity analyses showed that when the original model was applied to only those patients who were mechanically ventilated for a duration of 2 days at inclusion (83% of the patients in the validation cohort), as in the development cohort, calibration and discrimination remained poor. Thus, differences in inclusion criteria do not explain the poor performance.

At last, the fact that the performance of the original model could not be reproduced in the validation cohort may be attributable to the small sample size of the development study causing unstable predictions and possibly incorrect predictor selection. In fact, in the newly developed model, including the cohorts of the development and validation cohort, other predictors (except for lactate) were selected.

Importance of External Validation

The performance of any prediction model tends to be lower than expected when it is applied to new patients.[29] Therefore, every developed prediction model should be validated in new individuals before the model is applied in practice or implemented in guidelines. This step is, erroneously, often skipped.[16]

This study underlines the importance of external validation. It showed that generalizability and transportability of the previously developed model was poor and that the original model could thus not be used in clinical practice, also after extensive updating. Even the maximal SOFA score in the first 2 days after admission could predict ICU-AW better than the updated models.

Limitations of the Study

This study has some limitations not previously declared. Of all patients in which strength could not be measured (n = 189), 64 patients were transferred before they were attentive. As the clinical condition of these patients allowed a transfer from the ICU to the ward, this group may be less severely ill and may contain less patients with ICU-AW, masking the true incidence of ICU-AW in the validation cohort.

Furthermore, because strength measurements were available in less patients than beforehand accounted for, we did not have enough data to validate the model, update the model, and again externally validate an updated or new model. Therefore, we used all available data to validate and update the model. Future studies should account for more loss of patients (due to dead, transfer, delirium, etc).

Development of a New Prediction Model

Model updating did not result in a useful model with sufficient discrimination and therefore a new model was developed using the development and validation cohort together. This new model included RASS score, gender, highest lactate, and treatment with corticosteroids as predictors. The new model was based on a much larger cohort than the original development cohort, resulting in more stable estimates. The AUC-ROC was fair (0.70 [95% CI: 0.66-0.75]) and comparable with that of the original model. External validation is needed to prove performance and clinical usefulness in a new validation cohort.

Recently another prediction model for ICU-AW was proposed,[30] including the following predictors: steroid therapy, intensive insulin therapy, number of days on mechanical ventilation, sepsis, renal failure, and hematologic failure. This

model, which was based on data of 4157 patients, at least 12 hours mechanically ventilated, in whom only 3% had ICU-AW, showed good discrimination (AUC-ROC: 0.81 [95% CI: 0.78-0.84]). Calibration was, however, not reported, and external validation was not performed. In this study, the definition of ICU-AW was based on an operational definition and not on the MRC score. Therefore, it is very likely that patients with mild-to-moderate ICU-AW have been missed, explaining the very low incidence rate of ICU-AW (3%) in their cohort. These differences make comparison of the study results difficult. No other studies investigating the early prediction of ICU-AW, using clinical parameters, have been published.

Conclusions

External validation of a previously developed prediction model for ICU-AW showed poor calibration and discrimination. Updating methods improved calibration but not discrimination. A new prediction model using data from the development and validation cohort showed fair discrimination and classified patients better than the APACHE IV and the SOFA scores. However, early prediction of ICU-AW, using clinical parameters, with good discrimination seems to be challenging.

Acknowledgments

The authors thank Vanessa Sai-A-Tjin and all physiotherapists in the participating centers for their help with the data collection.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Professor I. N. van Schaik received departmental honoraria for serving on scientific advisory boards and a steering committee for CSL-Behring.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Dr L. Wieske is supported by a personal grant (ZonMw-AGIKO grant [project number 40-00703-98-11636]) from the Netherlands Organization for Health Research and Development.

ORCID iD

Esther Witteveen, MD, PhD https://orcid.org/0000-0001-7378-799X

References

1. Stevens RD, Marshall SA, Cornblath DR, et al. A framework for diagnosing and classifying intensive care unit-acquired weakness. *Crit Care Med*. 2009;37(10 suppl):S299-S308. doi:10.1097/CCM.0b013e3181b6ef67.
2. Stevens RD, Dowdy DW, Michaels RK, Mendez-Tellez PA, Pronovost PJ, Needham DM. Neuromuscular dysfunction acquired in critical illness: a systematic review. *Inten Care Med*. 2007;33(11):1876-1891.
3. Hermans G, Van Mechelen H, Clerckx B, et al. Acute outcomes and 1-year mortality of intensive care unit-acquired weakness. A cohort study and propensity-matched analysis. *Am J Respir Crit Care Med*. 2014;190(4):410-420. doi:10.1164/rccm.201312-2257OC.
4. Tennilä A, Salmi T, Pettilä V, Roine RO, Varpula T, Takkunen O. Early signs of critical illness polyneuropathy in ICU patients with systemic inflammatory response syndrome or sepsis. *Inten Care Med*. 2000;26(9):1360-1363. doi:10.1007/s001340000586.
5. Novak KR, Nardelli P, Cope TC, et al. Inactivation of sodium channels underlies reversible neuropathy during critical illness in rats. *J Clin Invest*. 2009;119(5):1150-1158. doi:10.1172/JCI36570.1150.
6. Latronico N, Bertolini G, Guarneri B, et al. Simplified electrophysiological evaluation of peripheral nerves in critically ill patients: the Italian multi-centre CRIMYNE study. *Crit Care*. 2007;11(1). doi:10.1186/cc5671.
7. Connolly B, O'Neill B, Salisbury L, McDowell K, Blackwood B. Physical rehabilitation interventions for adult patients with critical illness across the continuum of recovery: an overview of systematic reviews protocol. *Syst Rev*. 2015;4:130. doi:10.1186/s13643-015-0119-y.
8. Schaller SJ, Anstey M, Blobner M, et al. Early, goal-directed mobilisation in the surgical intensive care unit: a randomised controlled trial. *Lancet*. 2016;388(10052):1377-1388. doi:10.1016/S0140-6736(16)31637-3.
9. Fan E, Cheek F, Chlan L, et al. An official american thoracic society clinical practice guideline: the diagnosis of intensive care unit-acquired weakness in adults. *Am J Respir Crit Care Med*. 2014;190(12):1437-1446. doi:10.1164/rccm.201411-2011ST.
10. Hough CL, Lieu BK, Caldwell ES. Manual muscle strength testing of critically ill patients: feasibility and interobserver agreement. *Crit Care*. 2011;15(1):R43. doi:10.1186/cc10005.
11. Connolly BA, Jones GD, Curtis AA, et al. Clinical predictive value of manual muscle strength testing during critical illness: an observational cohort study. *Crit Care*. 2013;17(5):R229. doi:10.1186/cc13052.
12. Hermans G, Clerckx B, Vanhullebusch T, et al. Interobserver agreement of Medical Research Council sum-score and handgrip strength in the intensive care unit. *Muscle Nerve*. 2012;45(1):18-25. doi:10.1002/mus.22219.
13. Wieske L, Witteveen E, Verhamme C, et al. Early prediction of intensive care unit-acquired weakness using easily available parameters: a prospective observational study. *PLoS One*. 2014;9(10):e111259. doi:10.1371/journal.pone.0111259.
14. Toll DB, Janssen KJM, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol*. 2008;61(11):1085-1094. doi:10.1016/j.jclinepi.2008.04.008.
15. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med*. 2015;13(1):1-10. doi:10.1016/j.eururo.2014.11.025.
16. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): explanation and Elaboration. *Ann Intern Med*. 2015;162(1):W1-W73. doi:10.7326/M14-0698.

17. Van Swieten J, Koudstaal P, Visser M, Schouten H, van Gijn J. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke*. 1988;19:604-607.
18. Parsons EC, Kross EK, Ali NA, et al. Red blood cell transfusion is associated with decreased in-hospital muscle strength among critically ill patients requiring mechanical ventilation. *J Crit Care*. 2013;28(6):1079-1085. doi:10.1016/j.jcrc.2013.06.020.
19. Anastasopoulos D, Kefaliakos A, Michalopoulos A. Is plasma calcium concentration implicated in the development of critical illness polyneuropathy and myopathy? *Crit Care*. 2011;15(5): R247. doi:10.1186/cc10505.
20. De Jonghe B, Sharshar T, Lefaucheur JP, et al. Paresis acquired in the intensive care unit: a prospective multicenter study. *JAMA*. 2002;288(22):2859-2867. doi:10.1001/jama.288.22.2859.
21. Gosselink R, Clerckx B, Robbeets C, Vanhullebusch T, Vanpee G, Segers J. Physiotherapy in the intensive care unit. *Phys Ther Rev*. 2006;11(1):49-56. doi:10.1179/108331906X98921.
22. Sommers J, Engelbert RH, Dettling-Ihnenfeldt D, et al. Physiotherapy in the intensive care unit: an evidence-based, expert driven, practical statement and rehabilitation recommendations. *Clin Rehabil*. 2015;29(11):1051-1063. doi:http://dx.doi.org/10.1177/0269215514567156.
23. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med*. 2004;23(16):2567-2586. doi:10.1002/sim.1844.
24. Akaike H. A new look at the statistical model identification. *Autom Control IEEE Trans*. 1974;19:716-723.
25. Pencina MJ, Steyerberg EW, DÁgostino RB. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med*. 2011;30(1):11-21. doi:10.1002/sim.4085.Extensions.
26. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*. 2011;30(4):377-399. doi:10.1002/sim.4067.
27. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley & Sons; 1987.
28. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol*. 2005;58(5):475-483. doi:10.1016/j.jclinepi.2004.06.017.
29. Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, ... Woodward M. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*. 2012;98(9):691-698. doi:10.1136/heartjnl-2011-301247.
30. Penuelas O, Muriel A, Frutos-Vivar F, et al. Prediction and outcome of intensive care unit-acquired paresis. *J Intensive Care Med*. 2018;33(1):16-28. doi:10.1177/0885066616643529.