Journal of Clinical Epidemiology 142 (2022) 230–241

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Mortality prediction in intensive care units including premorbid functional status improved performance and internal validity

André Moser[a,*], Matti Reinikainen[b], Stephan M. Jakob[c], Tuomas Selander[d], Ville Pettilä[e], Olli Kiiski[f], Tero Varpula[g], Rahul Raj[h,1], Jukka Takala[c,1]

[a] *CTU Bern, University of Bern, Mittelstrasse 43, 3012 Bern, Switzerland*
[b] *Department of Anaesthesiology and Intensive Care, Kuopio University Hospital and University of Eastern Finland, Kuopio, Finland*
[c] *Department of Intensive Care Medicine, Bern University Hospital, University of Bern, Bern, Switzerland*
[d] *Science Service Center, Kuopio University Hospital, Kuopio, Finland*
[e] *Division of Intensive Care, University of Helsinki and Helsinki University Hospital, Helsinki, Finland*
[f] *Health and Care, Benchmarking Services, TietoEvry, Helsinki, Finland*
[g] *Division of Intensive Care, University of Helsinki and Helsinki University Hospital, Helsinki, Finland*
[h] *Department of Neurosurgery, University of Helsinki and Helsinki University Hospital, Helsinki, Finland*

Accepted 17 November 2021; Available online 22 November 2021

Abstract

Objective: Prognostic models are key for benchmarking intensive care units (ICUs). They require up–to–date predictors and should report transportability properties for reliable predictions. We developed and validated an in–hospital mortality risk prediction model to facilitate benchmarking, quality assurance, and health economics evaluation.

Study Design and Setting: We retrieved data from the database of an international (Finland, Estonia, Switzerland) multicenter ICU cohort study from 2015 to 2017. We used a hierarchical logistic regression model that included age, a modified Simplified Acute Physiology Score–II, admission type, premorbid functional status, and diagnosis as grouping variable. We used pooled and meta–analytic cross–validation approaches to assess temporal and geographical transportability.

Results: We included 61,224 patients treated in the ICU (hospital mortality 10.6%). The developed prediction model had an area under the receiver operating characteristic curve 0.886, 95% confidence interval (CI) 0.882–0.890; a calibration slope 1.01, 95% CI (0.99–1.03); a mean calibration –0.004, 95% CI (–0.035 to 0.027). Although the model showed very good internal validity and geographic discrimination transportability, we found substantial heterogeneity of performance measures between ICUs (*I*–squared: 53.4–84.7%).

Conclusion: A novel framework evaluating the performance of our prediction model provided key information to judge the validity of our model and its adaptation for future use.

Keywords: Case mix; In–hospital mortality; Intensive care; Prediction model; Transportability; Validation

Abbreviations: APACHE, Acute Physiology and Chronic Health Evaluation; AUC, area under the curve; CI, confidence interval; ECOG, Eastern Cooperative Oncology Group; FICC, Finnish Intensive Care Consortium; ICI, Integrated Calibration Index; ICU, intensive care unit; LASSO, least absolute shrinkage and selection operator; LOESS, locally estimated scatterplot smoother; OR, odds ratio; PI, prediction interval; ROC, receiver operating characteristic; SAPS, Simplified Acute Physiology Score; WHO, World Health Organization.

Conflict of Interest: Dr. Moser has no conflicts of interest. Dr. Reinikainen has no conflicts of interest. Dr. Jakob: The Department of Intensive Care Medicine, University Hospital Bern, has, or has had in the past, research & development/consulting contracts with Edwards Lifesciences Services GmbH, Phagenesis Limited and Nestlé. The money was paid into a departmental fund, and Dr. S. Jakob did not receive any financial gain. The Department of Intensive Care Medicine, University Hospital Bern, has received in the past unrestricted educational grants from the following organizations for organizing bi-annual postgraduate courses in the fields of critical care ultrasound, management of ECMO and mechanical ventilation: Pierre Fabre Pharma AG (formerly known as RobaPharm), Pfizer AG, Bard Medica S.A., Abbott AG, Anandic Medical Systems, PanGas AG Healthcare, Orion Pharma, Bracco, Edwards Lifesciences AG, Hamilton Medical AG, Fresenius Kabi (Schweiz) AG, Getinge Group Maquet AG, Dräger Schweiz AG, Teleflex Medical GmbH. Mr. Selander has no conflicts of interest. Dr. Pettilä has no conflicts of interest. Mr. Kiiski has no conflicts of interest. Dr. Varpula has no conflicts of interest. Dr. Raj has no conflicts of interest. Dr. Takala has no conflicts of interest.

[1] Senior authors contributed equally.

* Corresponding author.
E-mail address: andre.moser@ctu.unibe.ch (A. Moser).

https://doi.org/10.1016/j.jclinepi.2021.11.028

What is new?

Key findings

- Our mortality prediction model—which combined established clinically relevant predictors with premorbid functional status and diagnoses as modeling variables—showed very good internal validity, geographic discrimination and temporal transportability, with a substantial heterogeneity of performance measures between ICUs.

What this adds to what is known?

- Premorbid functional status and diagnosis are known predictors of ICU–relevant study outcomes, but are not regularly implemented in established scoring systems. The inclusion of this information showed increased predictive model performance compared to predictions from established risk scoring systems, while showing good internal validation and transportability properties.

What is the implication, what should change now?

- To the best of our knowledge, this is one of the first development and validation studies to investigate geographical and temporal transportability properties of an ICU mortality prediction model. Transportability properties are key in the reliable monitoring and benchmarking of ICUs and for their planning. They provide an important piece of information about the model validity in other study populations and settings, and should be quantified in future validation studies of ICU prediction models.

1. Introduction

Scoring systems for prediction of mortality risk are widely used to characterize severity of illness in intensive care unit (ICU) patients in clinical trials, benchmarking, quality assurance, and health economics evaluations [1]. The two most widely used scoring systems—APACHE (acute physiology and chronic health evaluation) [2] and SAPS (simplified acute physiology score) [3]—were introduced in the 1980s and repeatedly updated to preserve and improve their predictive value in response to advances in patient care, therapeutic options with prognostic relevance, and changes in demographics. Despite the updates (the most recent generations are APACHE–IV and SAPS–3 [4–8]) and recalibration, these models may lose their validity over time, and have poor external validity when applied in different health care systems. This may interfere with benchmarking of ICUs, and ultimately impact the decision–making of clinicians, health care providers and regulatory bodies. ICU prediction models developed in study populations from the United Kingdom (ICNARC), Australia and New Zealand (ANZIC) and the Netherlands have been updated to address poor validity and were extended with new clinical predictors like functional status which were strongly associated with the study outcome [9–12].

The performance of prediction models depends on their validity and transportability, and can be classified into different frameworks [13–15]. Geographical and temporal transportability indicate performance outside the study population used for the development of the prediction model, for example, in other hospitals or time periods [14,15]. Lack of transportability, case mix differences, changes in mortality between ICUs and over time, drive the need to recalibrate existing prediction models or to develop new ones [14,16]. Reporting their validation and transportability is important to avoid biased outcome predictions and to support the planning of ICU benchmarking programs where new ICUs might be included or a comparison with ICUs outside an existing benchmark system might evolve.

The aim of this study is to develop and validate an in-hospital mortality risk prediction model by adding simple indicators of premorbid functional status to established outcome predictors (age, severity of illness, diagnosis, admission type) to quantify the validity and transportability properties of the prediction model and to interpret their impact using a proposed framework for validation [14,15].

2. Methods

The manuscript has been written in accordance with the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement [17]. The TRIPOD checklist is provided in Supplemental Table 1.

2.1. Study design and population

We conducted a retrospective multicenter study, with prospectively collected data from the international Finnish Intensive Care Consortium benchmarking database (FICC) that includes ICUs from Finland, Estonia and Switzerland. We used all intensive care admissions of the years 2015 to 2017. We excluded readmissions, admissions after cardiac surgery, and admissions with care restricted to evaluation for potential organ donation. Cardiac surgery admissions were excluded due to their specific pre– and perioperative risk profiles.

2.2. Data source

We used data from 2015 to 2017 in the FICC database. The FICC consists of 22 ICUs in Finland, and the ICUs

of one university hospital in Estonia (100% of university ICU admissions and 16% all ICU admissions in Estonia) and one in Switzerland (33% of all university hospital ICU admissions and 11% of all ICU admissions in Switzerland). All except two ICUs were multidisciplinary.

Data related to clinical diagnosis, severity of illness scoring systems, care interventions, physiologic, administrative, and in–hospital mortality are prospectively collected during the ICU and hospital stay into local patient health care records (electronic in Finland and Switzerland, paper in Estonia) [18]. Each admission is validated by trained ICU nurse data managers or intensivists using logical rules, median filtering and graphic displays to ensure data quality, before transfer to the FICC database.

2.3. Ethical approval

The model was developed in conjunction with a health economic analysis, whose study protocol, database contents and data management process were approved by the National Institute of Health and Welfare, Finland (Decision numbers THL/1524/5.05.00/2017 and THL/1173/05.00/2018). According to the regulations in Finland, Estonia, and Switzerland, no ethics committee approval was needed for retrospective use of anonymized data.

2.4. Study outcomes

Prediction of in–hospital mortality. Data on in–hospital mortality was prospectively collected for each admission from the hospital record.

2.5. Predictors

Based on the literature and discussions among clinical experts we used the following predictors: age, a modified SAPS–II score which excludes age and admission type (see Supplemental Text 1), surgical vs nonsurgical admission, emergency vs elective admission, diagnosis (APACHE–III diagnoses [19,20], Supplemental Table 2) and functional status based on the WHO ECOG classification [21], with the two best categories combined into one (category 0: Normal functional status or able to perform light work, 1: Light limitations: Unable to work but capable of all self–care, 2: Moderate limitations: Need for some help in self–care, 3: Severe limitations: Fully dependent on help).

2.6. Statistical methods

We describe the study population using frequencies (n), percentages (%), median, and interquartile range (IQR). For the prediction model development we followed the recommendations by Steyerberg and Harrell [22,23]. In brief, Steyerberg and Harrell recommend not to use data–splitting approaches for validation and that heterogeneity of performance measures should be assessed. We used hierarchical multivariable logistic regression models accounting for nested admissions within APACHE–III diagnostic groups to develop the prediction model ($M_{\text{developed}}$). A second prediction model excluded premorbid functional status as predictor ($M_{\text{developed2}}$). We modeled the continuous variables age and modified SAPS–II score as restricted cubic splines with three knots chosen at the 10th, 50th, and 90th percentiles [13,24]. A priori we included an interaction effect between age and the modified SAPS-II score. We report predictor effects as odds ratios (OR) with 95% confidence intervals (CI). We assessed the overall association of variables and model fit using χ^2(k)–statistics with k degrees of freedom from a deviance test. For parsimony of the risk score, we set varying intercept estimates for APACHE–III diagnoses to zero if the 95% CI overlapped by an OR of 1. We calculated the probability of hospital death from the original SAPS–II score to compare predictions from derived models [4]. We assessed discrimination ability with the area under the receiver operating characteristics curve (AUC), mean calibration (calibration–in–the–large), weak calibration (calibration slope) and moderate calibration performance using a fitted calibration curve from locally estimated scatterplot smoother (LOESS), the Integrated Calibration Index (ICI), the maximum, median and 90th percentile of the absolute difference between the LOESS calibration and the diagonal line (Emax, E50, E90), and the Brier score [25–27]. We used a second-degree polynomial with a smoothing parameter set to 0.75 for LOESS. We used a modified large sample Hosmer–Lemeshow test, which was devised from a model attaining a P value of the traditional Hosmer–Lemeshow test of 0.05 in a sample of one million observations [28]. We used a bootstrap approach for internal model validation based on 100 replicates [13,24]. We assessed temporal and/or geographical transportability, that is, to what extent predictions perform as well in other study populations (in time and/or at different hospitals) and used a proposed framework for interpretation [13,14]. For that purpose, we defined two time periods: Time period one covers the years 2015 and 2016, whereas time period two covers the year 2017.

In brief, we first assessed to what extent the development and validation samples were related (ie, to what extent case mix differed between hospitals and time periods). For that purpose, we derived AUC from hierarchical logistic membership models (including the predictors used for the mortality prediction model and the outcome variable) and reported standardized standard deviations (SDs) of linear predictors (ie, SDs from validation samples are divided by the SD of the development sample) [14]. Development and validation samples were defined by a "leave–one–unit–out" or a "leave–one–period–out" approach (“internal–external” validation) [15]. Each model performance measure is estimated from the left-out validation samples. We reported pooled estimates and es-

timates from random effects meta-analyses of AUC and calibration performance measures with 95% CI and prediction intervals (PI) [15]. Results from the meta-analyses include DerSimonian–Laird estimators for between–study standard deviation, *I*–squared and Cochran's Q–statistic for heterogeneity. Because of the small number of time periods no meta-analytic approach was used for assessing temporal transportability.

We performed several sensitivity analyses. First, we derived a prediction model which used the same predictors as $M_{\text{developed}}$ but did not account for APACHE–III diagnosis groups ($M_{\text{sensitivity}}$), by using an ordinary logistic regression model. We compared $M_{\text{developed}}$ with $M_{\text{sensitivity}}$ using analysis of deviance. We further calculated the probability of hospital death from $M_{\text{developed}}$, $M_{\text{developed2}}$ (same as $M_{\text{developed}}$ but without predictor premorbid functional status), $M_{\text{sensitivity}}$, and the original SAPS–II score, and plotted receiver operating characteristic curves. Second, we used (non-–hierarchical) ordinary and penalized logistic regression and least absolute shrinkage and selection operator (LASSO) with APACHE–III diagnostic groups as fixed effects. Penalized estimates were optimized with a modified Akaike Information Criterion [24]. The minimum penalizing factor for LASSO was derived from 30 cross-validation samples. Third, we compared discrimination and calibration performance of the shrunken risk score with its non–shrunken version, that is, when the varying intercept estimates of APACHE–III diagnoses were not set to zero. All analyses were complete case analyses and excluded patients with missing predictor information and were performed in R version 4.0.2.

3. Results

3.1. Study population

The eligible study population included 61,385 patients. We excluded 161 patients (0.3%) because of missing values in age ($n = 14$), admission type ($n = 126$), SAPS-II score ($n = 15$), and APACHE–III diagnosis ($n = 6$). Thus, the final analysis included 61,224 patients with 6,463 (10.6%) in-hospital deaths (Table 1). The median age was 63 years (IQR 24), the median SAPS–II score was 31 points (IQR 22) and the median modified SAPS–II was 15 points (IQR 19) and 40% were female. Eighty percent of the admissions were emergency admissions and 40% were surgical. Around 30% of all patients had a premorbid functional limitation. The number of admissions and number of deaths per each unit and year is shown in Supplemental Table 3.

3.2. Prediction model

The prediction model ($M_{\text{developed}}$) developed for the risk score included five variables (age, modified SAPS–II score, admission type, premorbid functional status, and APACHE–III diagnostic group) with 15 parameters (one parameter for intercept, two parameters for nonlinear age, two parameters for nonlinear modified SAPS–II score, four parameters for interaction effect, two parameters for admission type (surgical, emergency), three parameters for functional status, one parameter for APACHE–III diagnoses grouping variable). Fig. 1 shows the joint predictor effects of the categorical variables from the hierarchical logistic regression model. Surgical patients had lower odds of dying than nonsurgical patients [OR 0.79, 95% CI (0.64–0.98)]; emergency admissions had higher odds of dying than elective admissions [OR 2.85, 95% CI (2.36–3.45)]. Patients with premorbid functional limitations showed higher odds of dying than patients with normal premorbid functional status. Premorbid functional status was strongly associated with mortality (deviance($M_{\text{developed}}$) = 27,891, deviance($M_{\text{developed2}}$) = 28,052; $\chi^2(3) = 161.0$, $P < 0.001$). Age and the modified SAPS–II score showed strong evidence for a nonlinear relationship (for age: $\chi^2(1) = 8.07$, $P = 0.004$; for modified SAPS–II score $\chi^2(1) = 86.18$, $P < 0.001$; Fig. 2). There was strong evidence for an interaction effect of the nonlinear modeled variables ($\chi^2(4) = 67.22$, $P < 0.001$). Fig. 3 shows the varying intercept estimates for the APACHE–III diagnoses. The estimated between-diagnoses standard deviation was 0.63; that is, 95% of the ORs of the APACHE–III diagnosis estimates lie between 0.29 (2.5% OR) and 3.43 (97.5% OR). Estimates with 95% CIs overlapping an OR of 1 are shown in orange. These estimates were set to zero in the final risk score for model parsimony. APACHE–III code = 0 covers "other postoperative" admissions ($n = 560$, 61% various emergencies). Due to the clinical heterogeneity, their estimate was also set to zero. The final prediction model includes 40 APACHE–III diagnoses (calculation formula in Supplemental Text 2). Supplemental Fig s . 1 and 2 show crude predictor effects. The effect of emergency admissions decreased from a crude model [OR = 5.97, 95% CI (5.01–7.10)] to OR = 2.85 (reported above) in an adjusted model.

3.3. Discrimination and calibration

Fig. 4 shows the internal discrimination and calibration properties of the developed model ($M_{\text{developed}}$): AUC 0.886, 95% CI (0.882–0.890), mean calibration –0.004, 95% CI (–0.035 to 0.027), calibration slope 1.01, 95% CI (0.99–1.03), ICI 0.134, E50 0.032, E90 0.463, Emax 1.000, and a Brier score 0.067. The Hosmer–Lemeshow test for large samples resulted in $P = 0.05$. The prediction model without the predictor premorbid functional ($M_{\text{developed2}}$) had an AUC 0.885, 95% CI (0.881–0.889), mean calibration 0.001 (–0.030 to 0.031), and calibration slope 1.01, 95% CI (0.99–1.03). Predictions from the original SAPS–II score revealed only slightly lower discrimination ability [AUC = 0.864, 95% CI (0.860–0.869)], but poor calibration (Supplemental Fig . 3). Supplemental Fig .

Table 1. Patient characteristics

Characteristic	2015	2016	2017	All years
	n/Median (%/IQR)	*n*/Median (%/IQR)	*n*/Median (%/IQR)	*n*/Median (%/IQR)
Age (years)	63 (24)	63 (24)	64 (23)	63 (24)
Gender (male)	11,882 (60)	12,335 (60)	12,757 (61)	36,974 (60)
SAPS–II score	31 (22)	31 (23)	31 (22)	31 (22)
Modified SAPS–II score	15 (20)	15 (20)	15 (19)	15 (19)
Surgical admission	7,931 (40)	7,880 (39)	8,460 (40)	24,271 (40)
Emergency admission	15,730 (80)	16,365 (80)	16,643 (79)	48,738 (80)
Functional status				
Normal	14,128 (72)	14,650 (72)	15,002 (71)	43,780 (72)
Light limitation	3,484 (18)	3,656 (18)	3,803 (18)	10,943 (18)
Moderate limitation	1,609 (8)	1,667 (8)	1,695 (8)	4,971 (8)
Severe limitation	521 (3)	502 (3)	507 (2)	1,530 (3)
Nonsurvivor	2,094 (11)	2,156 (11)	2,213 (11)	6,463 (11)

Abbreviation: IQR, interquartile range.

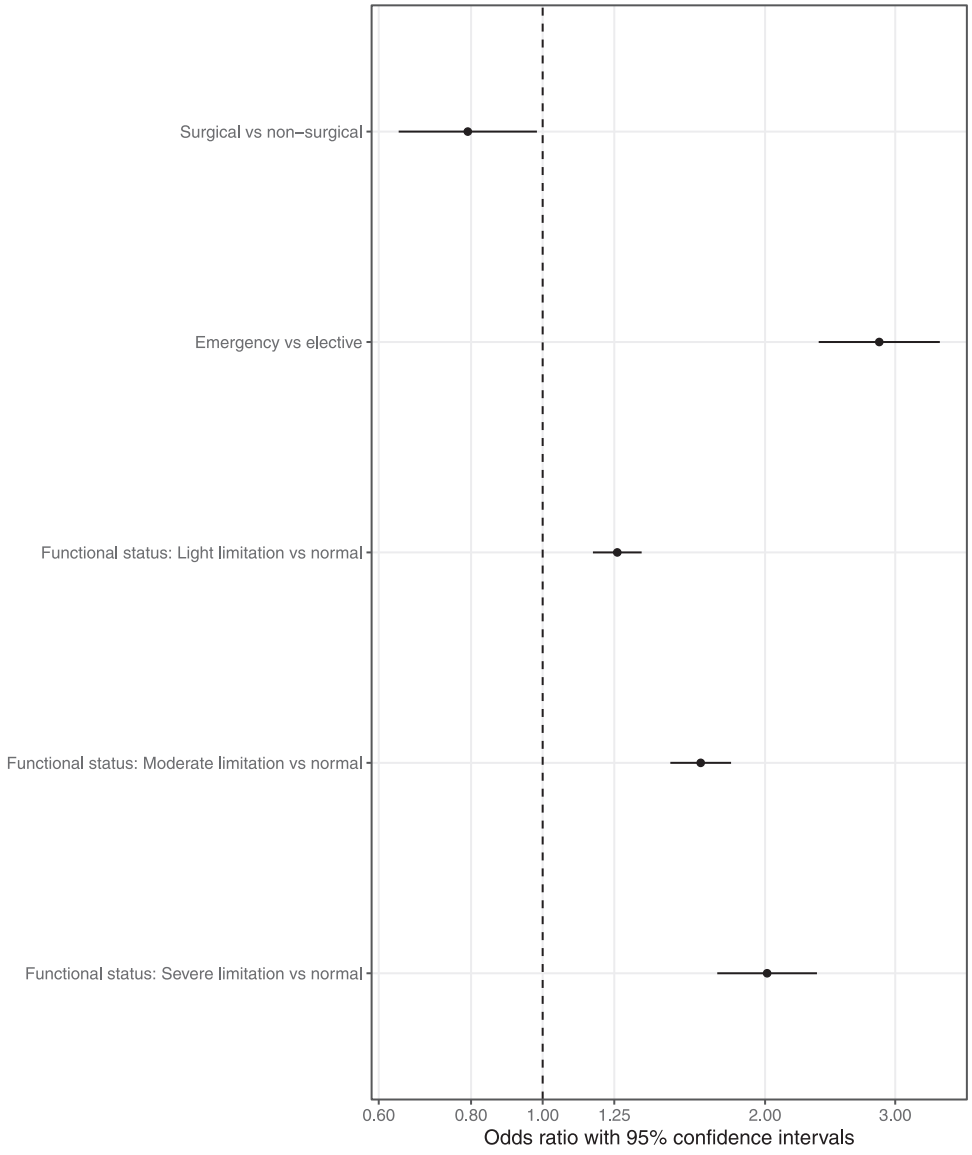

Fig. 1. Joint predictor effects of categorical variables.

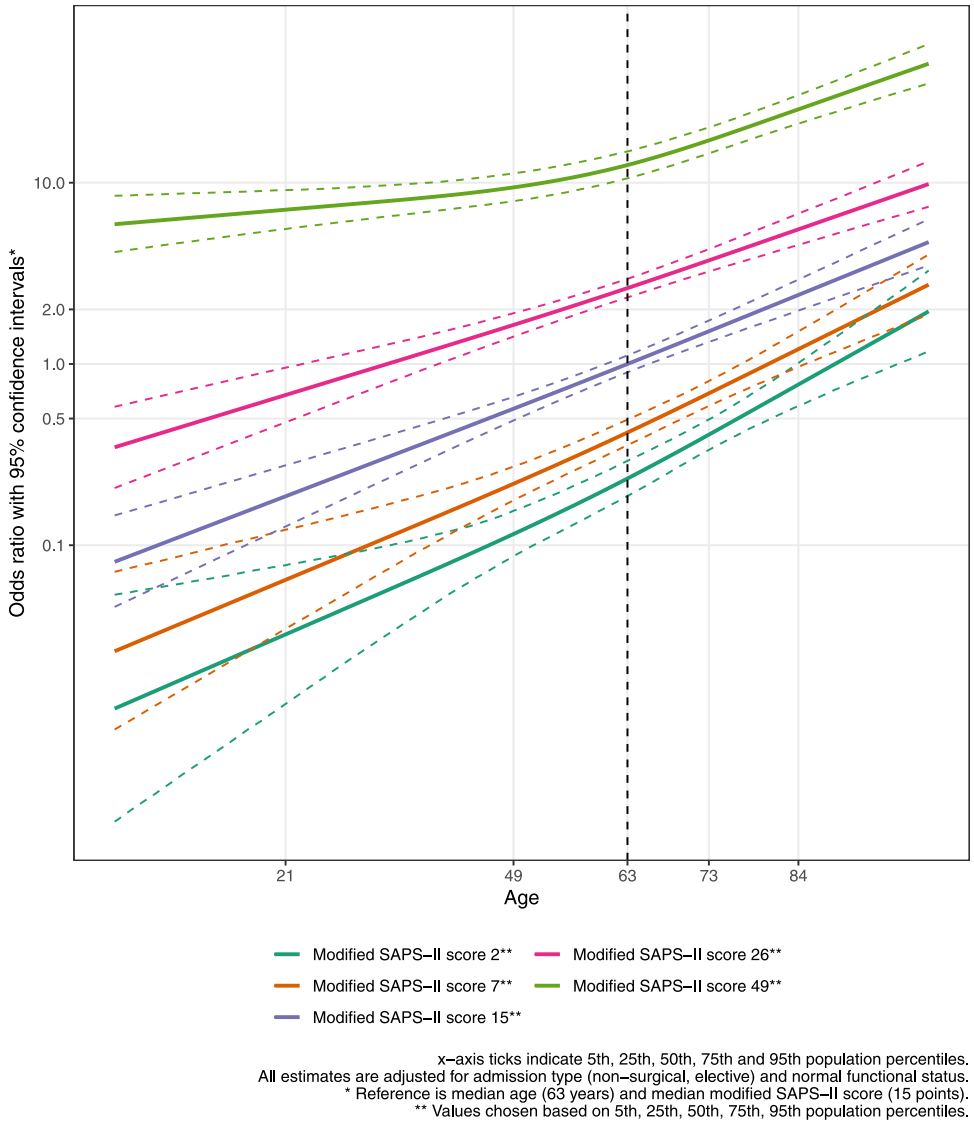

Fig. 2. Predictor effects of nonlinear modeled interaction effect between age and modified SAPS-II.

4 shows receiver operating characteristic curves from mortality predictions from $M_{\text{developed}}$, $M_{\text{developed2}}$, $M_{\text{sensitivity}}$ and from the original SAPS–II score. Discrimination ability was best for the developed prediction model.

3.4. Validation and interpretation

We assessed the relevance of case mix differences. Supplemental Fig . 5 shows AUC estimates from membership models. For geographical validation samples, AUC ranged from 0.65 to 0.96, indicating a substantial heterogeneity between the development and validation samples. Two units with AUC $>$ 0.9 are specialized ICUs. The development and validation samples between time periods were similar in terms of the used predictor variables and the outcome (AUC $=$ 0.54). Supplemental Fig . 6 shows results from standardized SDs of linear predictors (range 0.74 to 1.11), indicating moderate to large case-mix differences between hospitals. The interpretation of results from linear predictors is comparable to the ones from membership models. One hospital was left out of the membership modeling and the geographical transportability assessment because of a low number of admissions (n $=$ 14).

Table 2 shows results from internal and internal–external validation and transportability investigations. We observed very good internal model reproducibility (for example, the AUC from the bootstrap samples was 0.888 compared to 0.886 from the development sample). Although the overall internal–external calibration was very good, with 95% CIs and PIs overlapping a calibration slope of 1 or a mean calibration value of 0 for all validation approaches (except for the pooled geographical–temporal calibration slope, which was 0.978, 95% CI [0.963–0.993]), we found a moderate to high heterogeneity between ICUs for AUC, calibration slope and mean calibration (I–squared

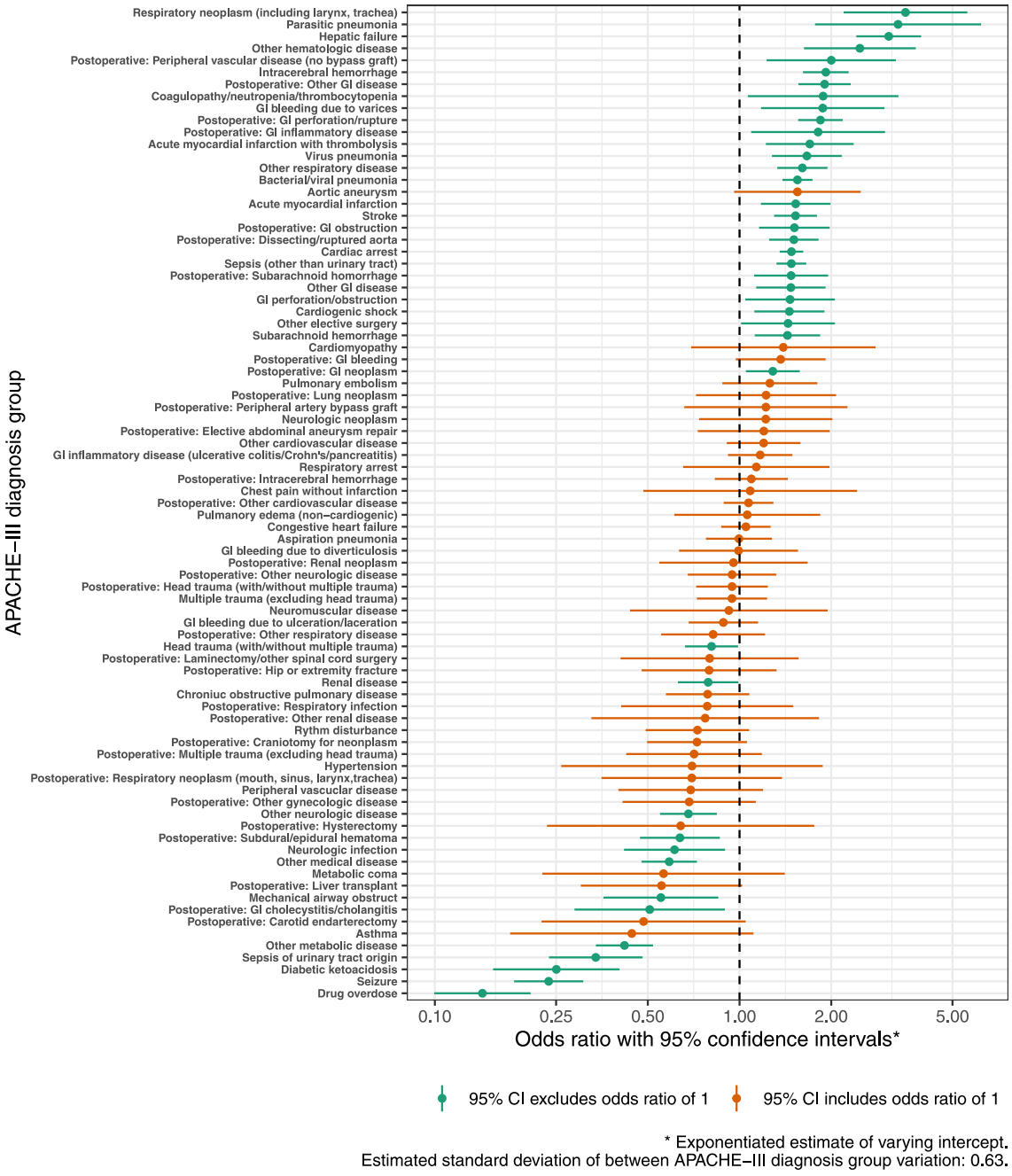

Fig. 3. Varying intercept estimates of APACHE-III diagnosis groups from a hierarchical logistic regression model.

ranging from 53.4–84.7%). Supplemental Figs. 7–15 show the in–depth results for transportability based on pooled estimates and random–effects meta–analyses.

3.5. Sensitivity analyses

In sensitivity analyses, we compared $M_{\text{developed}}$ with a prediction model ($M_{\text{sensitivity}}$) which did not account for APACHE–III diagnosis groups but used the same predictors as model $M_{\text{developed}}$. $M_{\text{sensitivity}}$ is an ordinary logistic regression model with 14 parameters. We found evidence for a better model fit of $M_{\text{developed}}$ compared to $M_{\text{sensitivity}}$ (deviance($M_{\text{developed}}$) = 27,891, deviance($M_{\text{sensitivity}}$) = 28,864; $\chi^2(1) = 973.3$, $P < 0.001$). Supplemental Figs. 16 and 17 show the sensitivity analysis results from different modeling strategies. In general, effect estimates from the investigated model approaches were similar. Supplemental Fig. 18 shows the discrimination and calibration properties of the nonshrunken model; that is, when the estimated varying intercepts of the APACHE–III diagnosis groups were not set to zero. The results were similar to the implemented risk score from $M_{\text{developed}}$.

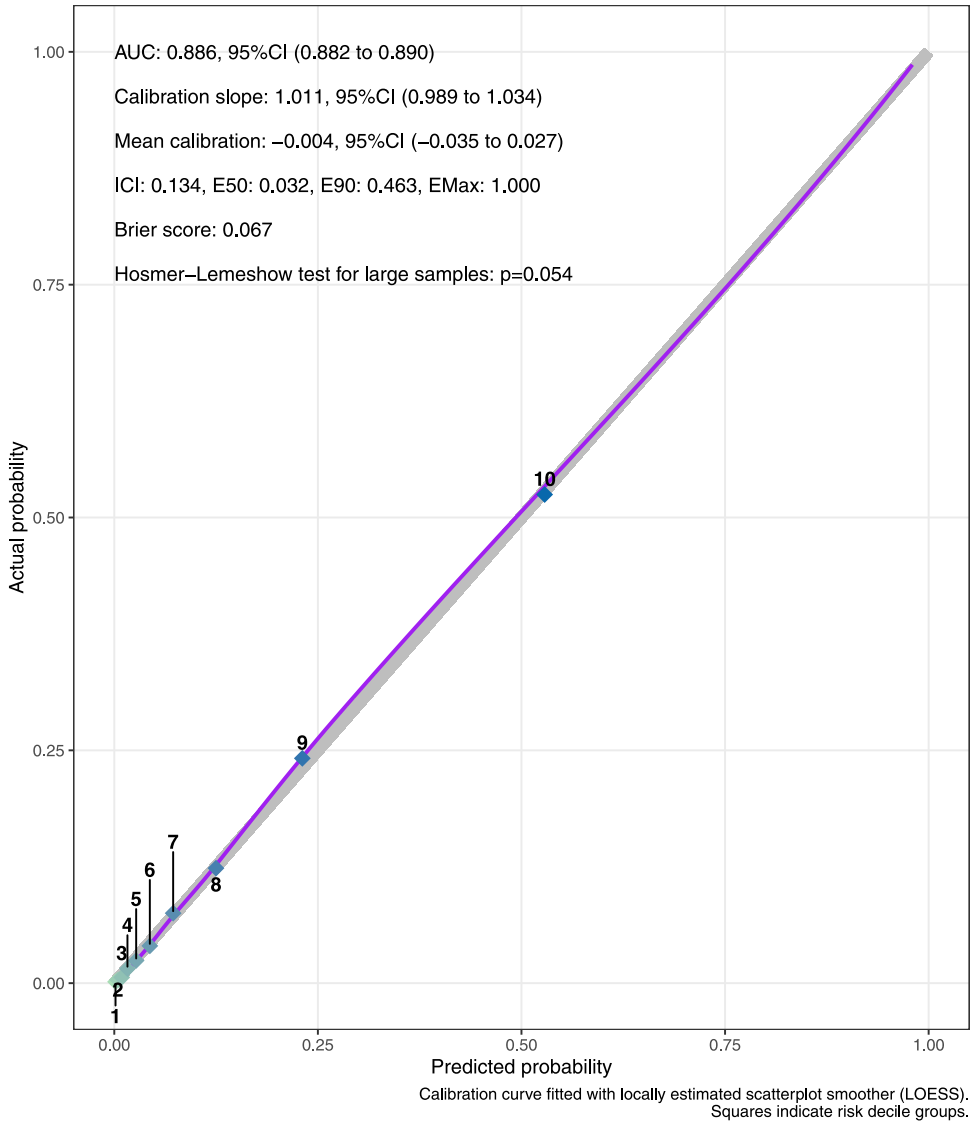

Fig. 4. Discrimination and calibration properties of developed prediction model.

4. Discussion

Based on data of 61,224 patients in a large international multicenter ICU database, we created a new in-hospital mortality prediction model by adding premorbid functional status to well–established predictors of in–hospital mortality (age, severity of acute illness, diagnosis, admission type). Premorbid functional status was strongly associated with mortality and increased the predictive performance of the model, while the inclusion of diagnosis accounted for the large heterogeneity between diagnosis groups. Overall, the prediction model showed excellent discrimination and calibration performance. We used a proposed framework for interpreting validation results which investigated case mix differences and transportability properties. We concluded that while a large heterogeneity between ICUs due to case mix differences exists, our prediction model might provide reliable predictions for future monitoring and benchmarking of key performance indicators.

Valid mortality predictions are essential for use of standardized mortality ratios (SMRs) to assess ICU performance, both for within–ICU evolution and for comparisons between ICUs in benchmarking programs [29,30]. However their use has been criticized for susceptibility to case mix differences [1]. By using different analyses approaches (membership model approach and internal–external validation) we found substantial differences in case mix among hospitals in our study population. Although our findings from internal validation and cross–validated pooled estimates showed very good discrimination and calibration properties, we found a moderate to high variation of prediction model performance measures between ICUs using meta–analytic approaches. Predictions intervals from meta–analytic approaches (that is, the uncertainty inter-

Table 2. Validation and transportability properties

Measure	Validation	Transportability	Overall estimate (95% CI/PI*)	*I*-squared (95% CI)
AUC	Internal: Bootstrap**	Reproducibility	0.888 (0.884–0.892)	
	Internal–external: Pooled	Geographical	0.884 (0.880–0.887)	
	Internal–external: Meta-analysis*	Geographical	0.881 (0.847–0.914)	70.8% (55.5–80.9%)
	Internal–external: Pooled	Temporal	0.886 (0.879–0.893)	
	Internal–external: Meta-analysis*	Temporal	Not reported	Not reported
	Internal–external: Pooled	Geographical–temporal	0.885 (0.880–0.890)	
	Internal–external: Meta-analysis*	Geographical–temporal	0.889 (0.850–0.928)	53.4% (24.3–71.3%)
Calibration slope	Internal: Bootstrap**	Reproducibility	1.00718 (1.00691–1.00745)	
	Internal–external: Pooled	Geographical	0.980 (0.958–1.002)	
	Internal–external: Meta-analysis*	Geographical	1.033 (0.839–1.227)	71.5% (56.6–81.2%)
	Internal–external: Pooled	Temporal	1.010 (0.972–1.049)	
	Internal–external: Meta-analysis	Temporal	Not reported	Not reported
	Internal–external: Pooled	Geographical–temporal	1.002 (0.975–1.029)	
	Internal–external: Meta-analysis*	Geographical–temporal	1.066 (0.813–1.319)	58.2% (33.0–73.9%)
Mean calibration	Internal: Bootstrap**	Reproducibility	0.00632 (0.000582–0.000680)	
	Internal–external: Pooled	Geographical	–0.008 (–0.039 to 0.022)	
	Internal–external: Meta-analysis*	Geographical	0.065 (–0.322 to 0.452)	84.7% (78.3–89.3%)
	Internal–external: Pooled	Temporal	0.011 (–0.042 to 0.063)	
	Internal–external: Meta-analysis	Temporal	Not reported	Not reported
	Internal–external: Pooled	Geographical–temporal	–0.006 (–0.043 to 0.031)	
	Internal–external: Meta-analysis*	Geographical–temporal	0.039 (–0.348 to 0.427)	65.8% (46.4–78.2%)

Abbreviations: CI, confidence interval; PI, prediction interval.
*PI reported.
**From 100 bootstrap replicates.

val in which a performance measure of a potentially new ICU will be expected to lie) indicated that the discriminative ability was very good (lowest lower 95% PI: 0.85), while the prediction uncertainties of the mean calibration and calibration slope indicated a potentially not optimal performing prediction model for new ICUs. Heterogeneity between ICUs was mostly expected, given that our data includes admissions from different hospital categories and from different health care systems, and can be explained by difference in case mix, in resource use or in quality of care. Our joint geographical–temporal transportability results require a careful interpretation because they are prone to potential model overfitting and biased performance measures because of the small event size of certain hospitals using only admissions from the year 2017 for validation.

The "customization," modernization, and development of new prediction models should offer clear benefits [16,31–34]. The inclusion of diagnoses should help cluster patients into homogenous groups, with similar treatment procedures, resource utilization and patient outcomes [9,35,36]. Nevertheless, it is conceivable that the premorbid functional status may have a relevant impact on risk of in–hospital death and consequently on the performance of mortality prediction models [12,34,37]. Approximately 30% of our study patients had premorbid functional limitation, which was strongly associated with mortality. Ferrando–Vivas et al. found that including the level of assistance needed in daily activities (none, some, total) improved the predictive performance of the ICNARC model [12]. The more detailed WHO ECOG classification we used also improved the performance, emphasizing the need to include premorbid functional status in ICU scoring systems. Demographic changes with an increasing proportion of older people will likely increase the ICU admissions of elderly. Muscedere et al. showed in a systematic review that frailty is an important factor of mortality in elderly treated in the ICU [38]. Dólera–Moreno et al. developed and validated a risk prediction model which uses frailty measures as predictors for all–cause mortality [34]. The implementation of geriatric assessments in ICUs might foster the use of specific predictors for older persons in future prediction models [39–41].

The statistical approach used for the development and validation of a prediction model should be critically addressed [13]. Due to the a priori decision to include diagnoses, we used a hierarchical regression model to account for grouping of intensive care admissions and discussed its advantages and disadvantages with clinical experts and statisticians [42]. The multicenter study design allowed in–depth validation of the developed risk model and we used the framework proposed by Debray et al. as a guide in interpreting our validation results [14,15]. The

internal, geographical, and temporal transportability validation methods provide important key measures for validity and for generalizability for broader study populations. Austin et al. used different admission eligibility criteria for different time periods to assess temporal and methodological portability [15]. Changes in the structural composition of ICUs in a benchmarking program (for example by the inclusion of specialized ICUs) inherently serve as a new set of admissions of an ICU which allows for the investigation of transportability properties. The FICC, for example, was extended by a neurosurgical ICU in 2017 and was included in the geographical transportability investigation. Our findings from the membership model and the used internal–external validation approaches will be included in regular reports of the FICC benchmarking program and might support the future planning of this program. Further, with the interpretation of measures of heterogeneity from the membership model or the used internal–external validation approaches we have a tool which supports the decision whether a recalibration of the prediction model might be required [14].

4.1. Strengths

First, the large sample size from different hospitals, years, and health care systems, and use of advanced statistical approaches, allows an in–depth investigation of the heterogeneity of performance measures and should enhance the generalizability of the model. For the validation of prediction models an event size of at least 100 events have been recommended [22,43]. Most hospitals in our data have a larger event size when data is combined for all years (ie, for the assessment of geographical transportability). Second, data validation by trained data managers, use of logical rules, median filtering, and graphic displays should enhance data quality. Third, the FICC cohort study collects relevant patient and clinical information which allows almost unbiased predictions, by using appropriate predictor and grouping information.

4.2. Limitations

All prediction models tend to deteriorate over time as medical care evolves, demographics change, and new diseases appear. Although our model was stable over the recent three years, its predictive ability is likely to change over time. Therefore, regular evaluation of the predictive value of the model should be foreseen and recalibration performed as necessary. The methodology we used will facilitate such a recalibration. Further, future prediction models for intensive care could be improved by using quality of life measures and other study endpoints or hospital information (like, hospital category, staffing information, or hospital volume) to possibly improve the performance and transportability of the prediction model for benchmarking programs [16,31]. Such information was not used for the prediction model development.

4.3. Implications

Our study has important implications for clinicians, health care providers, and health system evaluations. First, the inclusion of premorbid functional status in addition to established clinical predictors in our prediction model improved the performance of the prediction model. Second, the quantification of validation and transportability properties provides important information for future benchmarking programs. As proposed by Debray et al. [14], we encourage in the development of future ICU prediction models to report the transportability properties due to their relevance to changes in case mix, advances in medical technologies or changes in mortality. Third, this modeling strategy might help clinicians, health care providers, guideline developers and regulatory bodies to enhance evaluation of ICU care and finally to improve population health.

5. Conclusions

Premorbid functional status is an important predictor of hospital outcome and improved the predictive performance of our prediction model. Our model showed very good internal validity but a substantial heterogeneity of performance measures between ICUs, providing key information to judge the validity of our model and its adaptation for future use. We used a proposed framework for interpreting model validation findings, which proved helpful in the process of validating our prediction model. We encourage clinicians, health care providers, guideline developers and health service researchers to carefully address the multidimensional aspects of developing ICU prediction models. This includes discussions about the selection of relevant predictors, the interpretation of statistical findings and approaches used, as well as the reporting of key indicators to interpret the model performance and the potential consequences, like transportability to other study settings. We believe that a structured development process improves nonexperts' trust in the methodology, supports targeted communication of key performance indicators, and helps assess the reliability of a prediction model in selected settings.

Authors' contributions

A.M.: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – Original Draft, Visualization; M.R.: Conceptualization, Methodology, Writing – Review & Editing; S.J.: Conceptualization, Methodology, Resources, Writing - Review & Editing; T.S.: Conceptualization, Methodology, Writing – Review & Editing; V.P.: Conceptualization, Methodology, Writing – Review

& Editing; O.K.: Conceptualization, Methodology, Writing – Review & Editing; T.V.: Conceptualization, Methodology, Writing – Review & Editing; R.R.: Investigation, Resources, Data Curation, Writing – Review & Editing, Supervision, Project, administration; J.T.: Conceptualization, Methodology, Investigation, Resources, Data Curation, Writing - Review & Editing, Visualization, Supervision, Project administration.

Funding

Dr. Raj has received research grants from Finska Läkaresällskapet, Svenska Kulturfonden (grant number: 168580), and Medicinska Understödsföreningen Liv & Hälsa.

Availability of data statement

The study authors had the permission from FINDATA (Social and Health Data Permit Authority) to analyze the data. A secondary use of the data for other researchers can be obtained through FINDATA (https://findata.fi/en/) according to the Finish Secondary Data Act.

Acknowledgments

We thank Dr. Andreas Limacher for his helpful comments on the statistical analysis and methodology.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jclinepi.2021.11.028.

References

[1] Salluh JIF, Soares M, Keegan MT. Understanding intensive care unit benchmarking. Intensive Care Med 2017;43:1703–7. doi:10.1007/s00134-017-4760-x.

[2] Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE. APACHE—acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 1981;9:591–7. doi:10.1097/00003246-198108000-00008.

[3] Le Gall JR, Loirat P, Alperovitch A, Glaser P, Granthil C, Mathieu D, et al. A simplified acute physiology score for ICU patients. Crit Care Med 1984;12:975–7. doi:10.1097/00003246-198411000-00012.

[4] Le Gall JR. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA J Am Med Assoc 1993;270:2957–63. doi:10.1001/jama.270.24.2957.

[5] Metnitz PGH, Moreno RP, Almeida E, Jordan B, Bauer P, Campos RA, et al. SAPS 3—from evaluation of the patient to evaluation of the intensive care unit. Part 1: objectives, methods and cohort description. Intensive Care Med 2005;31:1336–44. doi:10.1007/s00134-005-2762-6.

[6] Moreno RP, Metnitz PGH, Almeida E, Jordan B, Bauer P, Campos RA, et al. SAPS 3—from evaluation of the patient to evaluation of the intensive care unit. Part 2: development of a prognostic model for hospital mortality at ICU admission. Intensive Care Med 2005;31:1345–55. doi:10.1007/s00134-005-2763-5.

[7] Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med 2006;34:1297–310. doi:10.1097/01.CCM.0000215112.84523.F0.

[8] Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med 1985;13:818–29.

[9] Harrison DA, Parry GJ, Carpenter JR, Short A, Rowan K. A new risk prediction model for critical care: the Intensive Care National Audit & Research Centre (ICNARC) model. Crit Care Med 2007;35:1091–8. doi:10.1097/01.CCM.0000259468.24532.44.

[10] Paul E, Bailey M, Pilcher D. Risk prediction of hospital mortality for adult patients admitted to Australian and New Zealand intensive care units: development and validation of the Australian and New Zealand Risk of Death model. J Crit Care 2013;28:935–41. doi:10.1016/j.jcrc.2013.07.058.

[11] Brinkman S, Bakhshi-Raiez F, Abu-Hanna A, de Jonge E, Bosman RJ, Peelen L, et al. External validation of Acute Physiology and Chronic Health Evaluation IV in Dutch intensive care units and comparison with Acute Physiology and Chronic Health Evaluation II and Simplified Acute Physiology Score II. J Crit Care 2011;26(105):e11–105 e18. doi:10.1016/j.jcrc.2010.07.007.

[12] Ferrando-Vivas P, Jones A, Rowan KM, Harrison DA. Development and validation of the new ICNARC model for prediction of acute hospital mortality in adult critical care. J Crit Care 2017;38:335–9. doi:10.1016/j.jcrc.2016.11.031.

[13] Steyerberg EW. Clinical prediction models. New York, NY: Springer New York; 2009. doi:101007/978-0-387-77244-8.

[14] Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol 2015;68:279–89. doi:10.1016/j.jclinepi.2014.06.018.

[15] Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. J Clin Epidemiol 2016;79:76–85. doi:10.1016/j.jclinepi.2016.05.007.

[16] Teres D, Lemeshow S. When to customize a severity model. Intensive Care Med 1999;25:140–2. doi:10.1007/s001340050806.

[17] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015;350 g7594–g7594. doi:10.1136/bmj.g7594.

[18] Reinikainen M, Mussalo P, Hovilehto S, Uusaro A, Varpula T, Kari A, et al. Association of automated data collection and data completeness with outcomes of intensive care. A new customised model for outcome prediction. Acta Anaesthesiol Scand 2012;56:1114–22. doi:10.1111/j.1399-6576.2012.02669.x.

[19] Zimmerman JE, Wagner DP, Draper EA, Wright L, Alzola C, Knaus WA. Evaluation of Acute Physiology and Chronic Health Evaluation III predictions of hospital mortality in an independent database. Crit Care Med 1998;26:1317–26. doi:10.1097/00003246-199808000-00012.

[20] Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, et al. The APACHE III prognostic system. Chest 1991;100:1619–36. doi:10.1378/chest.100.6.1619.

[21] Oken MM, Creech RH, Tormey DC, Horton J, Davis TE, McFadden ET, et al. Toxicity and response criteria of the Eastern Cooperative Oncology Group. Am J Clin Oncol 1982;5:649–55.

[22] Steyerberg EW. Validation in prediction research: the waste by data splitting. J Clin Epidemiol 2018;103:131–3. doi:10.1016/j.jclinepi.2018.07.010.

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

[24] Harrell FE. Regression modeling strategies. Cham: Springer International Publishing; 2015. doi:101007/978-3-319-19425-7.

[25] Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Stat Med 2019;38:4051–65. doi:10.1002/sim.8281.

[26] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models. Epidemiology 2010;21:128–38. doi:10.1097/EDE.0b013e3181c30fb2.

[27] Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 2016;74:167–76. doi:10.1016/j.jclinepi.2015.12.005.

[28] Nattino G, Pennell ML, Lemeshow S. Assessing the goodness of fit of logistic regression models in large samples: a modification of the Hosmer-Lemeshow test. Biometrics 2020;76:549–60. doi:10.1111/biom.13249.

[29] Tambeur W, Stijnen P, Vanden Boer G, Maertens P, Weltens C, Rademakers F, et al. Standardised mortality ratios as a user-friendly performance metric and trigger for quality improvement in a Flemish hospital network: multicentre retrospective study. BMJ Open 2019;9:e029857. doi:10.1136/bmjopen-2019-029857.

[30] Verburg IWM, de Jonge E, Peek N, de Keizer NF. The association between outcome-based quality indicators for intensive care units. PLoS One 2018;13:e0198522. doi:10.1371/journal.pone.0198522.

[31] Glance LG, Szalados JE. Benchmarking in critical care. Chest 2002;121:326–8. doi:10.1378/chest.121.2.326.

[32] Metnitz PGH, Lang T, Vesely H, Valentin A, Le Gall JR. Ratios of observed to expected mortality are affected by differences in case mix and quality of care. Intensive Care Med 2000;26:1466–72. doi:10.1007/s001340000638.

[33] Kramer AA, Zimmerman JE, Knaus WA. Severity of illness and predictive models in society of critical care medicine's first 50 years: a tale of concord and conflict. Crit Care Med 2021;49:728–40. doi:10.1097/CCM.0000000000004924.

[34] Dólera-Moreno C, Palazón-Bru A, Colomina-Climent F, Gil-Guillén VF. Construction and internal validation of a new mortality risk score for patients admitted to the intensive care unit. Int J Clin Pract 2016;70:916–22. doi:10.1111/ijcp.12851.

[35] de Keizer NF, Bonsel GJ, Goldfad C, Rowan KM. The added value that increasing levels of diagnostic information provide in prognostic models to estimate hospital mortality for adult intensive care patients. Intensive Care Med 2000;26:577–84. doi:10.1007/s001340051207.

[36] Austin PC, van Walraven C, Wodchis WP, Newman A, Anderson GM. Using the Johns Hopkins Aggregated Diagnosis Groups (ADGs) to predict mortality in a general adult population cohort in Ontario, Canada. Med Care 2011;49:932–9. doi:10.1097/MLR.0b013e318215d5e2.

[37] Krinsley JS, Wasser T, Kang G, Bagshaw SM. Pre-admission functional status impacts the performance of the APACHE IV model of mortality prediction in critically ill patients. Crit Care 2017;21:110. doi:10.1186/s13054-017-1688-z.

[38] Muscedere J, Waters B, Varambally A, Bagshaw SM, Boyd JG, Maslove D, et al. The impact of frailty on intensive care unit outcomes: a systematic review and meta-analysis. Intensive Care Med 2017;43:1105–22. doi:10.1007/s00134-017-4867-0.

[39] Kerminen H, Huhtala H, Jäntti P, Valvanne J, Jämsen E. Frailty index and functional level upon admission predict hospital outcomes: an interRAI-based cohort study of older patients in post-acute care hospitals. BMC Geriatr 2020;20:160. doi:10.1186/s12877-020-01550-7.

[40] Cesari M, Franchi C, Cortesi L, Nobili A, Ardoino I, Mannucci PM. Implementation of the Frailty index in hospitalized older patients: results from the REPOSI register. Eur J Intern Med 2018;56:11–18. doi:10.1016/j.ejim.2018.06.001.

[41] Evans SJ, Sayers M, Mitnitski A, Rockwood K. The risk of adverse outcomes in hospitalized older patients in relation to a Frailty index based on a comprehensive geriatric assessment. Age Ageing 2014;43:127–32. doi:10.1093/ageing/aft156.

[42] Bouwmeester W, Twisk JW, Kappen TH, van Klei WA, Moons KG, Vergouwe Y. Prediction models for clustered data: comparison of a random intercept and standard regression model. BMC Med Res Methodol 2013;13:19. doi:10.1186/1471-2288-13-19.

[43] Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med 2016;35:214–26. doi:10.1002/sim.6787.