THE GERONTOLOGICAL SOCIETY OF AMERICA®

Journals of Gerontology: Medical Sciences
cite as: *J Gerontol A Biol Sci Med Sci*, 2023, Vol. 78, No. 4, 718–726
https://doi.org/10.1093/gerona/glac107
Advance Access publication June 3, 2022

OXFORD

Research Article

Clinically Interpretable Machine Learning Models for Early Prediction of Mortality in Older Patients with Multiple Organ Dysfunction Syndrome: An International Multicenter Retrospective Study

Xiaoli Liu, BS,[1,2,3] Clark DuMontier, MD, MPH,[4,5] Pan Hu, MS,[6,7] Chao Liu, MD,[7] Wesley Yeung, MBBS,[8,2] Zhi Mao, MD,[7] Vanda Ho, MBBCHIR, MRCP,[9] Patrick J. Thoral, MD,[10] Po-Chih Kuo, PhD,[2,11] Jie Hu, MD,[7] Deyu Li, PhD,[1] Desen Cao, PhD,[12] Roger G. Mark, PhD, MD,[2] FeiHu Zhou, MD,[7,13] Zhengbo Zhang, PhD,[1,3,*] and Leo Anthony Celi, MD, MPH[2,14,15,†]

[1]Key Laboratory for Biomechanics and Mechanobiology of Ministry of Education, Beijing Advanced Innovation Center for Biomedical Engineering, School of Biological Science and Medical Engineering, Beihang University, Beijing, China. [2]Laboratory for Computational Physiology, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [3]Center for Artificial Intelligence in Medicine, The General Hospital of PLA, Beijing, China. [4]New England Geriatric Research Education and Clinical Center, VA Boston Healthcare System, Boston, Massachusetts, USA. [5]Division of Aging, Brigham and Women's Hospital, Boston, Massachusetts, USA. [6]Department of Anesthesiology, The 920 Hospital of Joint Logistic Support Force of Chinese PLA, Kunming, Yunnan, China. [7]Department of Critical Care Medicine, The First Medical Center, The General Hospital of PLA, Beijing, China. [8]Department of Medicine, National University Hospital, Singapore, Singapore. [9]Division of Geriatric Medicine, Department of Medicine, National University Hospital, Singapore, Singapore. [10]Department of Intensive Care Medicine, Amsterdam UMC, Amsterdam, The Netherlands. [11]Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan. [12]Department of Biomedical Engineering, The General Hospital of PLA, Beijing, China. [13]Elderly Center, The General Hospital of PLA, Beijing, China. [14]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA. [15]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA.

*Address correspondence to: Zhengbo Zhang, PhD, Center for Artificial Intelligence in Medicine, The General Hospital of PLA, No. 28 Fuxing Rd, Beijing 100853, China. E-mail: zhangzhengbo@301hospital.com.cn

[†]Senior author.

Received: January 28, 2022; Editorial Decision Date: April 24, 2022

Decision Editor: Lewis A. Lipsitz, MD, FGSA

Abstract

Background: Multiple organ dysfunction syndrome (MODS) is associated with a high risk of mortality among older patients. Current severity scores are limited in their ability to assist clinicians with triage and management decisions. We aim to develop mortality prediction models for older patients with MODS admitted to the ICU.
Methods: The study analyzed older patients from 197 hospitals in the United States and 1 hospital in the Netherlands. The cohort was divided into the young-old (65–80 years) and old-old (≥80 years), which were separately used to develop and evaluate models including internal, external, and temporal validation. Demographic characteristics, comorbidities, vital signs, laboratory measurements, and treatments were used as predictors. We used the XGBoost algorithm to train models, and the SHapley Additive exPlanations (SHAP) method to interpret predictions.
Results: Thirty-four thousand four hundred and ninety-seven young-old (11.3% mortality) and 21 330 old-old (15.7% mortality) patients were analyzed. Discrimination AUROC of internal validation models in 9 046 U.S. patients was as follows: 0.87 and 0.82, respectively;

© The Author(s) 2022. Published by Oxford University Press on behalf of The Gerontological Society of America.
All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

discrimination of external validation models in 1 905 EUR patients was as follows: 0.86 and 0.85, respectively; and discrimination of temporal validation models in 8 690 U.S. patients: 0.85 and 0.78, respectively. These models outperformed standard clinical scores like Sequential Organ Failure Assessment and Acute Physiology Score III. The Glasgow Coma Scale, Charlson Comorbidity Index, and Code Status emerged as top predictors of mortality.

Conclusions: Our models integrate data spanning physiologic and geriatric-relevant variables that outperform existing scores used in older adults with MODS, which represents a proof of concept of how machine learning can streamline data analysis for busy ICU clinicians to potentially optimize prognostication and decision making.

Keywords: International multicenter, Interpretable models, Machine learning, Mortality, Multiple organ dysfunction syndrome

Multiple organ dysfunction syndrome (MODS) is a continuous process with physiologic derangement in more than one organ (1), usually occurring after physiologic insults such as infection, burns, trauma, and shock (2). MODS is the leading cause of morbidity and mortality in patients who are admitted to intensive care unit (ICU) (1,3). Older patients (≥65 years old) with MODS have a significantly higher mortality risk compared with younger patients due to decreased physiologic reserve and pre-existing comorbidities (4,5). Accurate prognostication can help clinicians provide appropriate and individualized care.

A growing body of literature has demonstrated that clinical scoring systems—such as the Sequential Organ Failure Assessment (SOFA) score and the Acute Physiology and Chronic Health Evaluation-II (APACHE-II) score—fail to accurately assess and predict the risk of death (6) for the following reasons: the entailed prognostic factors had their weights assigned by experts, not fully reflecting the characteristics of larger populations (7); the fixed monotonic aggregation of each organ system state does not represent the complexity of the associations between the organ systems (7); and models were not adequately validated in multicenter and large sample cohorts. In recent years, the use of electronic health records (EHR) data has allowed researchers to develop machine learning (ML) algorithms for analysis of heterogeneous data yielding sophisticated prediction models like multitask Gaussian process model, Autoscore, recurrent neural network, and Federated Learning for dynamic risk prediction (8–12).

However, the application of these modern approaches to mortality prediction in older adults with MODS has had limited success. Studies to date have been marred by small patient cohorts (330–9 800 patients), single-center model training and validation, and the use of logistic regression (LR) models and univariate statistical methods that do not account for collinearity and complex interactions among predictors (13). Moreover, many ICU prediction models overemphasize acute physiologic and laboratory variables while ignoring prevalent geriatric syndromes—such as multimorbidity—that limit older adults' ability to withstand acute stressors (14,15). In the present study, we aimed to develop prediction models to assist clinicians in the early prognostication of older patients who were admitted to the ICU with MODS. Because there is heterogeneity in health status among adults aged over 65 years old (16), we developed and validated separate models for young-old (65–80 years) and old-old (≥80 years) patients using a large multicenter data set. We further analyzed the models to identify important predictors of mortality in each subgroup.

Method

We performed a multicenter retrospective cohort study using 4 open-access clinical databases including the Medical Information Mart for Intensive Care Database v1.4 (MIMIC-III) and MIMIC-IV v1.0 collected from the Beth Israel Deaconess Medical Center in Boston from 2001 to 2012 and 2014 to 2019, respectively (17,18); the eICU Collaborative Research Database v1.2 (eICU-CRD) collected from 208 hospitals in United States from 2014 to 2015 (19); and the AmsterdamUMCdb v1.0.2 collected from the Amsterdam University Medical Centers, The Netherlands from 2003 to 2016 (20). A detailed description of these databases is provided in Supplementary Material.

Study Population

We included all ICU patients ≥65 years old with MODS (21), defined as failure of 2 or more organs systems according to the SOFA score (22). We excluded patients with unknown outcomes, who stayed in the ICU for less than 24 hours, or who incurred repeat ICU admissions within the same hospital admission. We also excluded patients without any measurements of heart rate, respiratory rate, mean arterial pressure, Glasgow Coma Scale (GCS), temperature, and oxygen saturation in the first 24 hours of ICU admission. Data extracted from the MIMIC-III and eICU-CRD databases were combined into a single cohort for model development, whereas data from the AmsterdamUMCdb and MIMIC-IV were kept as separate cohorts for external and temporal validation, respectively. The young-old (65–80 years old) and old-old (≥80 years old) were studied separately (23). The inclusion criteria of all study cohorts was displayed in Figure 1.

Data Collection and Feature Construction

Five types of information were collected for model development: patient characteristics of age, gender, body mass index (BMI), Charlson Comorbidity Index (CCI), and Code Status (CS); vital signs such as GCS, heart rate, respiratory rate, and mean arterial pressure; laboratory results including glucose, creatinine, white blood cell, bilirubin level, etc.; urine output; clinical treatments received including mechanical ventilation (MV), continuous renal replacement therapy, and vasopressors. Only data measured during the first day of admission in the ICU was used. Representative statistical features were calculated based on the type of variable. Missing values were imputed using the median value of each feature except for FiO_2 (with the imputation of 21%). Additionally, we included a missing value indicator if a variable had missing values in 30% or more of patients. A total of 79 features were constructed. Additional information can be found including the proportions of missing raw data in Supplementary Tables 1 and 2.

Statistical Analysis

Continuous variables were reported as medians with interquartile ranges. The *t* test or Wilcoxon Rank Sum Test was used when appropriate to compare between surviving and nonsurviving elderly adults with MODS. Categorical variables were reported by the total

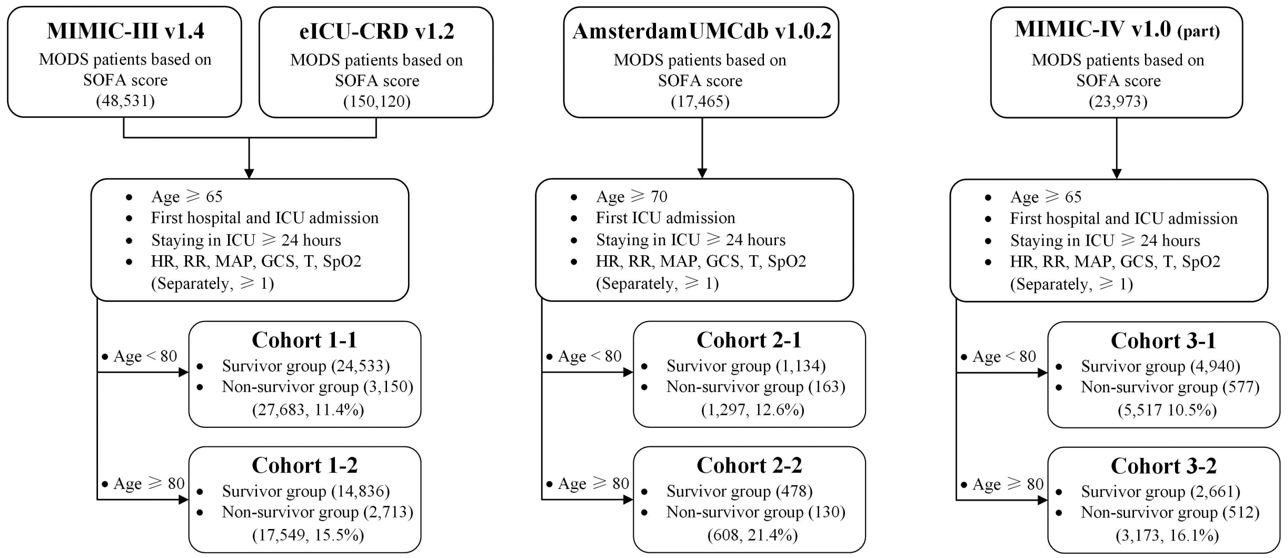

Figure 1. An overview of inclusion criteria with all study cohorts.

number and percentage. Two-sided *p*-values of less than .05 were considered statistically significant.

Model Development

We used the eXtreme Gradient Boosting (XGBoost) algorithm for the mortality prediction model. The XGBoost algorithm has previously been used in other health care applications with high performance, which is an optimized distributed tree gradient boosting method by converting weak learners to strong learners with iteratively refitting (24). Three other ML algorithms including logistic regression (LR), random forests (RF), and naive Bayesian (NB) were used as baseline models for comparison. We developed 2 early prediction models for young-old and old-old patients. For each patient subgroup, we used the combined cohort of patients from the MIMIC-III and eICU-CRD databases for model development (25). These patients were randomly sampled into an 80% training set for model training and 20% validation set for internal validation. The cohorts from the AmsterdamUMCdb and MIMIC-IV databases were used as separate external validation sets. Training and validation sets are terminology used in ML to denote the data used to develop the model and data used to evaluate the performance of the model, respectively. Internal and external validation refer to the evaluation of model performance within the same population in which a model was developed and within an external population, respectively. As data from the MIMIC-IV database were collected in a time period after the MIMIC-III and eICU-CRD databases, we define the external validation performed on this database as "temporal" validation, with the aim of estimating model performance when applied to data encountered in subsequent years. Hyperparameter tuning was performed using Bayesian optimization.

Model Evaluation

We performed internal and external validation, comparing against the baseline models and conventional clinical scoring systems including SOFA, Simplified Acute Physiology Score (SAPS), and Acute Physiology Score III (APSIII). Seven evaluation metrics were calculated along with their 95% confidence intervals (95% CI), including the area under the curve of the receiver operating characteristic curve (AUROC), sensitivity, specificity, accuracy, F1 score, precision (positive predictive value), and area under the precision-recall curve (AUPRC).

Model Interpretation

SHapley Additive exPlanations (SHAP) is a game theoretic approach to explain the predicted outcomes in ML models; it has been proven helpful for clinicians to understand the importance of model predictors, for example, for anesthesiologists to identify the cause of hypoxemia during surgery (26). The SHAP method uses the Shapley value to evaluate a feature's effect on model predictions and to measure its relative importance ranking (27). We used SHAP to identify important features that contributed to mortality predictions in our developed models.

Software Usage

The data extraction was accomplished with PostgreSQL Version 9.6. All calculations and analyses were performed utilizing Python software, version 3.7.1.

Results

Patient Characteristics

The combined MIMIC-III and eICU-CRD cohort included 45 232 older patients (5 863 nonsurvivors, 13.0%) with MODS. The AmsterdamUMCdb cohort included 1 905 older patients (293 nonsurvivors, 15.4%), and the MIMIC-IV included 8 690 (1 089 nonsurvivors, 12.5%). Detailed inclusion and exclusion criteria for each data set were provided in Supplementary Figures 1–4. Table 1 summarized the characteristics of 3 cohorts. The AmsterdamUMCdb cohort had the oldest median age, more severe disease as indicated by higher clinical severity scores, longest ICU median hospital stay, and highest proportion on MV. The specific type of ICU and length of hospital stay and ethnicity data were not available in this cohort. The combined MIMIC-III and eICU-CRD cohort and MIMIC-IV cohort were multiethnic with a higher proportion of White patients in the former. The combined MIMIC-III and eICU-CRD cohort had a higher proportion of patients admitted to the medical ICU. Old-old patients had on average lower BMI, proportion of patients on MV,

Table 1. The Comparison of the Total Study Cohorts' Baseline Characteristic

	Cohort 1 (Multicenter, United States)		Cohort 2 (Single Center, EUR)		Cohort 3 (Single Center, United States)	
	Young-Old (27 683)	Old-Old (17 549)	Young-Old (1 297)	Old-Old (608)	Young-Old (5 517)	Old-Old (3 173)
Demographic						
Age (y), (median, IQR)	72.0 [68.0, 76.0]	85.0 [82.0, 89.0]	74.0 [72.0, 76.0]	90.0 [84.0, 94.0]	72.0 [68.0, 75.0]	85.0 [82.0, 89.0]
Male, *n* (%)	15 481 (55.9)	7 935 (45.2)	813 (63.3)	338 (55.6)	3 391 (61.5)	1 568 (49.4)
BMI (kg/m^2), (median, IQR)	28.3 [24.4, 33.2]	25.7 [22.4, 29.4]	26.1 [23.4, 29.1]	25.1 [22.9, 28.4]	27.9 [24.3, 32.4]	26.0 [23.1, 29.6]
ICU type (%)						
CCU	5 481 (19.8)	3 392 (19.3)	0 (0.0)	0 (0.0)	2 402 (43.5)	1 068 (33.7)
CSRU	3 112 (11.2)	1 203 (6.9)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
MICU	13 988 (50.5)	9 644 (55.0)	0 (0.0)	0 (0.0)	695 (12.6)	460 (14.5)
NICU	1 645 (5.9)	951 (5.4)	0 (0.0)	0 (0.0)	556 (10.1)	450 (14.2)
SICU	2 554 (9.2)	1 620 (9.2)	0 (0.0)	0 (0.0)	616 (11.2)	392 (12.4)
TSICU	903 (3.3)	739 (4.2)	0 (0.0)	0 (0.0)	579 (10.5)	362 (11.4)
Mixed MICU/SICU	0 (0.0)	0 (0.0)	1 297 (100)	608 (100)	669 (12.1)	441 (13.9)
Ethnicity (%)						
Asian	438 (1.6)	256 (1.5)	0 (0.0)	0 (0.0)	148 (2.7)	86 (2.7)
Black	2 260 (8.2)	972 (5.5)	0 (0.0)	0 (0.0)	312 (5.7)	158 (5.0)
Hispanic	758 (2.7)	588 (3.4)	0 (0.0)	0 (0.0)	116 (2.1)	38 (1.2)
Other/Unknown	2 471 (8.9)	1 346 (7.7)	1 297 (100.0)	608 (100.0)	1 285 (23.3)	748 (23.5)
White	21 756 (78.6)	14 387 (82.0)	0 (0.0)	0 (0.0)	3 656 (66.3)	2 143 (67.5)
Treatments						
MV	12 929 (46.7)	6 282 (35.8)	1 102 (85.0)	497 (81.7)	4 593 (83.3)	2 506 (79.0)
CRRT	614 (2.2)	189 (1.1)	40 (3.1)	18 (3.0)	136 (2.5)	43 (1.4)
Severity of illness						
APSIII	41.0 [30.0, 57.0]	43.0 [32.0, 57.0]	54.0 [40.0, 71.0]	58.5 [44.0, 74.2]	42.0 [31.0, 59.0]	47.0 [37.0, 62.0]
SOFA	5.0 [3.0, 7.0]	5.0 [3.0, 7.0]	6.0 [4.0, 8.0]	6.0 [4.0, 8.0]	4.0 [3.0, 7.0]	4.0 [3.0, 6.0]
SAPS	20.0 [17.0, 24.0]	20.0 [17.0, 24.0]	24.5 [21.5, 27.5]	26.0 [22.0, 29.0]	19.0 [17.0, 22.0]	20.0 [18.0, 23.0]
Outcomes						
Days of hospital admission (d), (median, IQR)	7.1 [4.4, 11.5]	6.8 [4.2, 10.5]	—	—	8.2 [5.5, 13.8]	7.7 [4.9, 12.7]
Days of ICU admission (d), (median, IQR)	2.7 [1.7, 4.8]	2.5 [1.7, 4.1]	2.8 [1.5, 5.8]	2.9 [1.7, 6.2]	2.6 [1.5, 4.8]	2.7 [1.8, 4.7]
Hospital mortality, *n* (%)	3 150 (11.4)	2 713 (15.5)	163 (12.6)	130 (21.4)	577 (10.5)	512 (16.1)

Notes: APSIII = Acute Physiology Score III; BMI = body mass index; CCU = coronary care unit; CRRT = continuous renal replacement therapy; CSRU = cardiac surgery recovery unit; IQR = interquartile range; MICU = medical ICU; MV = mechanical ventilation; NICU = neurological intensive care unit; SAPS = simplified acute physiology score; SICU = surgical ICU; SOFA = sequential organ failure assessment score; TSICU = trauma/surgical ICU. Cohort 1 patients from the MIMIC-III, and eICU-CRD databases; Cohort 2 patients from the AmsterdamUMCdb; and Cohort 3 from the MIMIC-IV.

higher clinical scores (APSIII and SAPS), and higher mortality compared with the young-old across all cohorts.

We compared the characteristics of survivors and nonsurvivors in the combined MIMIC-III and eICU-CRD cohort (Supplementary Table 3). Nonsurvivors were significantly older in age, had higher severity scores, lower BMI upon ICU admission, and longer duration of ICU stay. Among young-old patients in AmsterdamUMCdb cohort (Supplementary Table 4), body weight, use of continuous renal replacement therapy (CRRT), higher severity scores, and longer ICU stay was associated with mortality. Old-old patients had similar risk factors for mortality with the exception of weight, and addition of MV. In the MIMIC-IV cohort (Supplementary Table 5), among young-old patients, age, and BMI were not significantly associated with mortality. In old-old patients, male gender and BMI were not significantly associated with mortality.

Model Evaluation

We present the internal and external evaluation of the final model stratified by the 2 age groups (Table 2). The model performed well in both internal validation (young-old: AUROC 0.866 [95% CI 0.849–0.881]; old-old: AUROC 0.821 [95% CI 0.801–0.841]), external validation of (young-old: AUROC 0.856 [95% CI 0.82–0.888]; old-old AUROC 0.853 [95% CI 0.813–0.891]), and temporal validation (young-old: AUROC 0.845 [95% CI 0.828–0.862]; old-old: AUROC 0.776 [95% CI 0.752–0.798]). Model performance was lower in the old-old compared with the young-old.

We then compared our model's performance against 3 baseline ML models and conventional clinical scores in 3 cohorts (Table 3). Consistently, our model had better performance compared with the baseline ML models and conventional clinical scores (Supplementary Table 6). We assessed model calibration visually using a calibration plot (Supplementary Figure 5), showing reasonable calibration results. We performed a sensitivity analysis to determine whether the use of a smaller subset of features chosen by feature importance ranking had an impact on model performance (Supplementary Table 7). Model performance decreased with the inclusion of fewer features but still outperformed conventional clinical scores. We assessed for racial bias comparing model performance between the whole population, White, Black and Hispanic subgroups with acceptable difference found between the subgroups (Supplementary Figure 6).

Interpretability

To improve the clinical utility of the model, we used the SHAP method to determine which features contributed to a prediction of mortality by the model and compared them between the 2 age groups (Supplementary Table 8). Figure 2A and B displays the top 20 risk factors in the 2 age groups. Features with greater overall importance appear higher (y axis). The SHAP value (x axis) indicates the impact of a feature in the model. A positive SHAP value indicates that a feature contributes to a prediction of mortality. For continuous features, a color gradient between red and blue represents a decreasing value of the feature from high to low. If a feature is binary (eg, yes or no), red indicates yes and blue indicates no. Risk factors including GCS (gcs_mean), Charlson Comorbidity Index (CCI, charlson), MV (vent_flag), respiratory rate (rr_mean), heart rate (hr_mean), shock index (si_mean), lowest temperature (t_min), and total urine output (uo_24hour) during the initial 24 hours of ICU stay were ranked as the 10 most important factors for all older patients. The top 4 features were common between the 2 groups being GCS, MV, CCI, and mean respiratory rate.

Table 2. Summary of Our Model's Validation Performance for Mortality Prediction in Multicenter Databases

	Internal Validation		External Validation in EUR		Temporal Validation in United States	
Indexes (95% CI)	Young-Old	Old-Old	Young-Old	Old-Old	Young-Old	Old-Old
AUROC	0.866 (0.849–0.881)	0.821 (0.801–0.841)	0.856 (0.82–0.888)	0.853 (0.813–0.891)	0.845 (0.828–0.862)	0.776 (0.752–0.798)
Sensitivity	0.816 (0.781–0.848)	0.807 (0.768–0.843)	0.847 (0.786–0.906)	0.815 (0.738–0.885)	0.821 (0.786–0.856)	0.738 (0.695–0.78)
Specificity	0.742 (0.727–0.754)	0.682 (0.663–0.7)	0.718 (0.688–0.749)	0.762 (0.716–0.803)	0.702 (0.686–0.715)	0.675 (0.655–0.695)
Accuracy	0.748 (0.736–0.761)	0.701 (0.684–0.718)	0.733 (0.706–0.761)	0.771 (0.733–0.807)	0.713 (0.7–0.726)	0.685 (0.667–0.703)
F1 score	0.425 (0.397–0.452)	0.456 (0.424–0.486)	0.444 (0.384–0.5)	0.604 (0.533–0.664)	0.375 (0.348–0.401)	0.431 (0.399–0.462)
Precision	0.287 (0.263–0.31)	0.317 (0.29–0.345)	0.301 (0.252–0.349)	0.48 (0.407–0.552)	0.243 (0.223–0.264)	0.304 (0.275–0.332)
AUPRC	0.521 (0.473–0.569)	0.478 (0.431–0.529)	0.498 (0.415–0.597)	0.595 (0.502–0.693)	0.416 (0.373–0.465)	0.412 (0.365–0.459)

Notes: AUROC = area under the receiver operating characteristic curve; AUPRC = area under the precision-recall curve; Precision = positive predictive value.

Table 3. Summary of Our Model's Performance Comparing With Machine Learning Methods and Clinical Scores in Multicenter Databases

AUROC (95% CI)	Internal Validation in United States (Cohort 1)		External Validation in EUR (Cohort 2)		Temporal Validation in United States (Cohort 3)	
	Young-Old	Old-Old	Young-Old	Old-Old	Young-Old	Old-Old
XGBoost	0.866 (0.849–0.881)	0.821 (0.801–0.841)	0.856 (0.82–0.888)	0.853 (0.813–0.891)	0.845 (0.828–0.862)	0.776 (0.752–0.798)
LR	0.844 (0.827–0.862)	0.793 (0.771–0.815)	0.836 (0.799–0.869)	0.831 (0.785–0.872)	0.822 (0.803–0.841)	0.723 (0.695–0.751)
RF	0.792 (0.77–0.813)	0.742 (0.714–0.768)	0.795 (0.753–0.834)	0.796 (0.752–0.838)	0.772 (0.749–0.793)	0.701 (0.673–0.727)
NB	0.784 (0.762–0.804)	0.731 (0.706–0.754)	0.767 (0.723–0.81)	0.784 (0.736–0.832)	0.772 (0.75–0.794)	0.697 (0.668–0.723)
APSIII	0.753 (0.729–0.777)	0.697 (0.669–0.725)	0.775 (0.728–0.82)	0.732 (0.681–0.781)	0.819 (0.799–0.839)	0.753 (0.727–0.78)
SAPS	0.742 (0.718–0.765)	0.708 (0.681–0.733)	0.766 (0.719–0.812)	0.774 (0.723–0.823)	0.733 (0.71–0.757)	0.687 (0.66–0.714)
SOFA	0.706 (0.679–0.731)	0.673 (0.643–0.702)	0.628 (0.572–0.684)	0.628 (0.561–0.691)	0.689 (0.662–0.716)	0.655 (0.628–0.685)

Notes: APSIII = Acute Physiology Score III; NB = naive Bayesian; LR = logistic regression; RF = random forests; SAPS = simplified acute physiology score, SOFA = sequential organ failure assessment score.

We found differences between young-old and old-old patients. Among kidney biomarkers, maximum blood urea nitrogen (BUN) was more important in young-old patients, whereas maximum creatinine was more important in old-old patients. Among liver biomarkers, maximum alkaline phosphatase was more important in young-old patients, while maximum AST was more important in old-old patients. Figure 2C and D shows the contribution of different features to an outcome of mortality in example patients from each age group and outcome. In the young-old group, a nonsurvivor had a high CCI (6 points), low urine output in the first 24 hours (150 mL), high BUN (54 mg/dL), and need for MV. A patient who survived had a normal GCS, was not mechanically ventilated, had good urine output (2 030 mL), and had low CCI (1 point). In the old-old group, a nonsurvivor required MV, required norepinephrine at a maximum rate of 0.20 mcg/kg/min, and had a high shock index of 1.1. The survivor did not require MV, had normal GCS (15 points), had low CCI (0 point), had mean respiratory rate (12.84 bpm), had low shock index (0.56), had normal peak creatinine (1.02 mg/dL), had normal heart rate (69.5 bpm), and had normal SpO_2 (98.9%).

Discussion

We leveraged large and international data sets to develop and externally validate predictive models for mortality tailored for older ICU patients with MODS. Incorporating a broad range of variables spanning physiologic and geriatric domains, our models consistently outperformed existing clinical risk scores for ICU patients. Moreover, our SHAP analysis revealed that cognitive status (GCS), pre-existing comorbidity (CCI), and CS—variables important in older patients—are just as if not more important than more traditionally used physiologic parameters in ICU clinical risk scores.

In the last decade, the median age of patients admitted to ICUs has been over 65 years (28). Most studies analyzing potential risk factors associated with ICU mortality have been derived from data sets comprising of younger adults, and these factors are extrapolated and incorporated into outcome prediction models for older patients (29). Our analyses revealed differential key prognostic factors for young-old and old-old patients upon admission to the ICU. As expected, physiologic variables remain prognostic in older adults. These variables included abnormal vital signs (temperature, heart rate, respiratory rate), low urine output, and markers of renal failure (BUN, creatinine). Specifically for older adults, mental status (GCS), comorbidity (CCI), and advance directives (code status, CS) emerged as top predictors of mortality alongside these physiologic variables. GCS, also included in SOFA and APACHE scores (30,31), stood as the most important predictor in young-old patients and second most important predictor in old-old patients. Impaired GCS can range from hypoactive and hyperactive delirium, coma, and medically induced sedation, all of which are prevalent and are a poor prognostic sign in ICU care (32,33). Older patients are more susceptible to delirium, yet this syndrome is often missed with harmful consequences. Our findings echo the call for system-wide interventions to prevent and manage delirium in the ICU (34).

Additionally, CCI emerged as one of the top 5 predictors in both young-old and old-old patients with MODS. Chronic conditions accumulate with age across multiple organ systems (35,36). This multimorbidity rests on a background of age-related depletion of physiologic reserves, contributing to states of frailty. In combination, multimorbidity and frailty predispose older patients to the development of MODS, with markedly increased risks of morbidity and mortality (14,15,37). Second, the presence of advanced or terminal

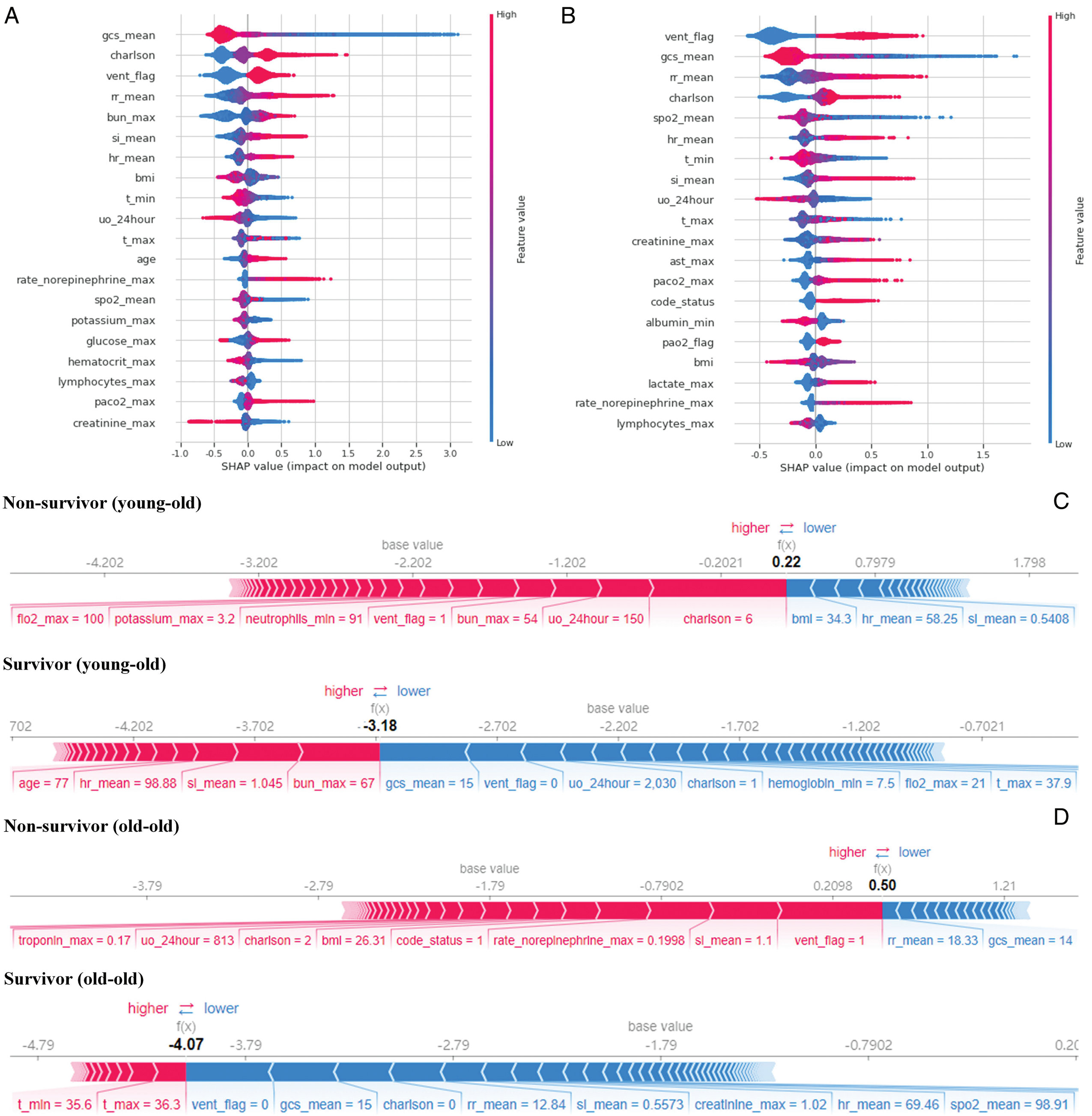

Figure 2. The model's interpretation. (**A**) and (**B**) The importance ranking of the top 20 risk factors with stability and interpretation using the optimal model of young-old and old-old patients. The higher the SHAP value a feature is given, the higher the risk of death for the patient. The red part in feature value represents a higher value (**C**) and (**D**). The interpretation of model prediction results with the 2 samples of nonsurvivor and survivor in 2 age groups, respectively. Charlson = Charlson Comorbidity Index; vent = mechanical ventilation; rr = respiratory rate; si = shock index; hr = heart rate; t = temperature; uo = urine output; ast = aspartate aminotransferase; max = the maximum value on the first day of ICU admission; min = the minimum value on the first day of ICU admission; mean = the average value on the first day of ICU admission; flag = the indicator vector representing measurements.

chronic conditions probably affects subsequent treatment decisions in older patients in the ICU. Patients and their families are less likely to pursue aggressive and prolonged resuscitation in the ICU in the presence of advanced stage heart failure or cancer, compared with patients with minimal comorbidity and a better baseline prognosis. Patient and family preferences, partially reflected by CS on admission, also drive the clinical decision making that in turn drives the intent and extent of interventions delivered in the ICU. We recommend for GCS to be prioritized and comorbidity burden to be integrated into clinical tools for older patients admitted to the ICU with MODS.

A strength of our analysis is its use of large and globally representative data from older adults admitted to the ICU. We utilized a large, multinational sample containing a broad range of variables. By doing so, we averted the problem of sample size that is commonly encountered in prediction modeling. A recent review of 129 studies focusing on mortality of older patients in ICU showed that multicenter analyses from a single country accounted for nearly a third of studies, whereas multinational analysis accounted only for a select few (8%) (29). To our knowledge, our sample size of 55 827 older patients in 198 hospitals across 2 countries is the most comprehensive to date for building and evaluating predictive models for older patients with MODS admitted to the ICU. We evaluated model performance within different countries (United States vs Amsterdam), different races (Caucasian vs Black and Hispanic), and over time (from 2014 to 2019). The results of these evaluations demonstrate model robustness across geography and across time.

We also conducted probability calibration curves and evaluated model performance using a subset of the features. We adopted an ensemble ML model, XGBoost, to represent the nonlinear and complex correlations between risk factors and outcome. In comparison, previous studies have mainly used regression models to characterize complex physiological states, which assumes monotonic relationships between independent variables (38). These assumptions may not always hold true for all clinical variables and limit the ability to obtain more accurate weighting of risk factors (7,26). With our approach, our model is superior to the linear regression model across all validations while providing interpretability, with modal discrimination (AUROC) of 0.82 and greater for internal and external validations across the 198 hospitals in different countries, compared with the lower AUROC of 0.71–0.88 in LR models of previous works and our baseline LR models (39).

Taken together, our models would aid in the provision of more calibrated prognostication of older patients admitted to the ICU with MODS. Examples of learning health systems that include ML to improve decision making are steadily rising (40–42), and our model integrates a broad array of important physiologic and health parameters that can be rapidly synthesized and presented to ICU clinicians. Frail, multimorbid older patients presenting with MODS are complex; integrating a broad array of variables would allow busy ICU clinicians to focus more on decision making and communication with patients and families (43).

Our study has a number of limitations. First, the disparity in the contribution of CCI and CS to our model's performance and SHAP analyses may be explained by the model's ceiling effect. Second, our model demonstrated relatively poor precision, which has been seen in other disease prediction models as well (44,45). Accordingly, our model's utility is to aid clinical decision making and not to replace the clinician (46). Third, we did not include admission diagnosis and subsequent ICU treatments, which can be incorporated in future versions of our model. Fourth, the results of temporal validation are somewhat biased, due to the not entirely consistent population distribution with the development set. Fifth, we recommend that the models need to be calibrated using local data before using. Finally, models would be further improved by including a validated measure of frailty (47).

In conclusion, this study developed and validated predictive models for mortality in older patients admitted to the ICU with MODS using ML methods in a large and international multicenter data set. Our models outperformed several risk scores traditionally used in the ICU setting and demonstrated that cognitive status, comorbidity, and code status emerge as powerful predictors when combined with physiologic and laboratory data routinely collected in the ICU. Our models represent a proof of concept of how ML using broad-ranging data could potentially streamline data synthesis for busy ICU clinicians and optimize decision making for complex older adults admitted with MODS. Future work would include refining our model and calibrating for drift, as well as pilot implementation in clinical settings.

Supplementary Material

Supplementary data are available at *The Journals of Gerontology, Series A: Biological Sciences and Medical Sciences* online.

Funding

This work had been carried out when X.L. was at the Laboratory for Computational Physiology, Massachusetts Institute of Technology as a visiting student with the support of the China Scholarship Council. The study was funded by the National Natural Science Foundation of China (62171471) and Big Data Research and Development Project of Chinese PLA General Hospital (2018MBD-009), and National Clinical Research Center for Geriatric Diseases of China (NCRCG-PLAGH-2017008). L.A.C. is funded by the National Institutes of Health through NIBIB R01 EB017205.

Conflict of Interest

None declared.

Acknowledgments

The authors wish to acknowledge the excellent technical assistance of Dr. Alistair Johnson and Dr. Tom Pollard. And thanks to Dr. Joanne Cohn for wonderful writing guidance.

Author Contributions

X.L., C.D., and L.A.C. contributed to the conception and design of the work; P.H. and Z.Z. contributed to collected data; C.L., W.Y., Z.M., and V.H. contributed to analyze data; P.J.T., P.-C.K., J.H., D.L., and D.C. contributed to interpret results; X.L., C.D., and W.Y. wrote the manuscript. Z.Z., F.Z., R.G.M., and L.A.C. reviewed the manuscript. All authors read and approved the final manuscript.

Data Availability

Code: the code that was used to extract code from the MIMIC-III, eICU-CRD, AmsterdamUMCdb, and MIMIC-IV databases, develop machine learning models and calculate statistical analysis are available at https://github.com/liuxiaoliXRZS/MODSE. Data set: we shared them on the PhysioNet website (https://physionet.org/about/database/).

References

1. Murray MJ, Coursin DB. Multiple organ dysfunction syndrome. *Yale J Biol Med.* 1993;66(5):501–510.
2. Seely AJ, Christou NV. Multiple organ dysfunction syndrome: exploring the paradigm of complex nonlinear systems. *Crit Care Med.* 2000;28(7):2193–2200. doi:10.1097/00003246-200007000-00003
3. Soo A, Zuege DJ, Fick GH, et al. Describing organ dysfunction in the intensive care unit: a cohort study of 20,000 patients. *Crit Care.* 2019;23(1):186. doi:10.1186/s13054-019-2459-9
4. Shiels MS, Almeida JS, García-Closas M, Albert PS, Freedman ND, Berrington de González A. Impact of population growth and aging on estimates of excess U.S. deaths during the COVID-19 pandemic, March to August 2020. *Ann Intern Med.* 2021;174(4):437–443. doi:10.7326/M20-7385
5. Koff WC, Williams MA. Covid-19 and immunity in aging populations—a new research agenda. *N Engl J Med.* 2020;383(9):804–805. doi:10.1056/NEJMp2006761
6. Poole D, Rossi C, Latronico N, et al. Comparison between SAPS II and SAPS 3 in predicting hospital mortality in a cohort of 103 Italian ICUs. Is new always better? *Intensive Care Med.* 2012;38(8):1280–1288. doi:10.1007/s00134-012-2578-0

7. Alaa AM, Yoon J, Hu S, van der Schaar M. Personalized risk scoring for critical care prognosis using mixtures of Gaussian processes. *IEEE Trans Biomed Eng*. 2018;65(1):207–218. doi:10.1109/TBME.2017.2698602
8. Ghassemi M, Pimentel M, Naumann T, et al. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. *Proc AAAI Conf Artif Intell*. 2015;29(1):446–453.
9. Xie F, Chakraborty B, Ong MEH, Goldstein BA, Liu N. Autoscore: a machine learning–based automatic clinical score generator and its application to mortality prediction using electronic health records. *JMIR Med Inform*. 2020;8(10):e21798. doi:10.2196/21798
10. Thorsen-Meyer HC, Nielsen AB, Nielsen AP, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *Lancet Digit Health*. 2020;2(4):e179–e191. doi:10.1016/S2589-7500(20)30018-2
11. Liu X, Liu T, Zhang Z, et al. TOP-Net prediction model using bidirectional long short-term memory and medical-grade wearable multisensor system for tachycardia onset: algorithm development study. *JMIR Med Inform*. 2021;9(4):e18803. doi:10.2196/18803
12. Dang TK, Tan KC, Choo M, Lim N, Weng J, Feng M. Building ICU in-hospital mortality prediction model with federated learning. In: Goebel R, Tanaka Y, W Wahlster W, ed. *Federated Learning*. Cham, Switzerland: Springer; 2020:255–268.
13. Biehl M, Takahashi PY, Cha SS, Chaudhry R, Gajic O, Thorsteinsdottir B. Prediction of critical illness in elderly outpatients using elder risk assessment: a population-based study. *Clin Interv Aging*. 2016;11:829–34. doi:10.2147/CIA.S99419
14. Damluji AA, Forman DE, van Diepen S, et al. Older adults in the cardiac intensive care unit: factoring geriatric syndromes in the management, prognosis, and process of care: a scientific statement from the American Heart Association. *Circulation*. 2020;141(2):e6–e32. doi:10.1161/CIR.0000000000000741
15. Turcotte LA, Zalucky AA, Stall NM, et al. Baseline frailty as a predictor of survival after critical care: a retrospective cohort study of older adults receiving home care in Ontario, Canada. *Chest*. 2021;160(6):2101–2111. doi:10.1016/j.chest.2021.06.009
16. Lee SB, Oh JH, Park JH, Choi SP, Wee JH. Differences in youngest-old, middle-old, and oldest-old patients who visit the emergency department. *Clin Exp Emerg Med*. 2018;5(4):249–255. doi:10.15441/ceem.17.261
17. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035. doi:10.1038/sdata.2016.35
18. Johnson AE, Stone DJ, Celi LA, Pollard TJ. The MIMIC Code Repository: enabling reproducibility in critical care research. *J Am Med Inform Assoc*. 2018;25(1):32–39. doi:10.1093/jamia/ocx084
19. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci Data*. 2018;5:180178. doi:10.1038/sdata.2018.178
20. Thoral PJ, Peppink JM, Driessen RH, et al. Sharing ICU patient data responsibly under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: the Amsterdam University Medical Centers Database (AmsterdamUMCdb) example. *Crit Care Med*. 2021;49(6):e563–e577. doi:10.1097/CCM.0000000000004916
21. Dann T. Global elderly care in crisis. *Lancet*. 2014;383(9921):927. doi:10.1016/S0140-6736(14)60463-3
22. Schuler A, Wulf DA, Lu Y, et al. The impact of acute organ dysfunction on long-term survival in sepsis. *Crit Care Med*. 2018;46(6):843–849. doi:10.1097/CCM.0000000000003023
23. Zhou CJ, Chen FF, Zhuang CL, et al. Feasibility of radical gastrectomy for elderly patients with gastric cancer. *Eur J Surg Oncol*. 2016;42(2):303–311. doi:10.1016/j.ejso.2015.11.013
24. Chen T, Guestrin CJA. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY: ACM; 2016:785–794.
25. Liu X, Hu P, Mao Z, et al. Interpretable machine learning model for early prediction of mortality in elderly patients with multiple organ dysfunction syndrome (MODS): a multicenter retrospective study and cross validation, *arXiv*, 2020, arXiv:2001.10977, preprint: not peer reviewed.
26. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2018;2(10):749–760. doi:10.1038/s41551-018-0304-0
27. Lundberg SM, Erion GG, Lee S-I. Consistent individualized feature attribution for tree ensembles. *arXiv*, 2018, arXiv:1802.03888, preprint: not peer reviewed.
28. Flaatten H, de Lange DW, Artigas A, et al. The status of intensive care medicine research and a future agenda for very old patients in the ICU. *Intensive Care Med*. 2017;43(9):1319–1328. doi:10.1007/s00134-017-4718-z
29. Vallet H, Schwarz GL, Flaatten H, de Lange DW, Guidet B, Dechartres A. Mortality of older patients admitted to an ICU: a systematic review. *Crit Care Med*. 2021;49(2):324–334. doi:10.1097/CCM.0000000000004772
30. Ferreira FL, Bota DP, Bross A, Mélot C, Vincent JL. Serial evaluation of the SOFA score to predict outcome in critically ill patients. *JAMA*. 2001;286(14):1754–1758. doi:10.1001/jama.286.14.1754
31. 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(5):1297–1310. doi:10.1097/01.CCM.0000215112.84523.F0
32. Inouye SK, Westendorp RG, Saczynski JS. Delirium in elderly people. *Lancet*. 2014;383(9920):911–922. doi:10.1016/S0140-6736(13)60688-1
33. Veiga D, Luis C, Parente D, et al. Postoperative delirium in intensive care patients: risk factors and outcome. *Rev Bras Anestesiol*. 2012;62(4):469–483. doi:10.1016/S0034-7094(12)70146-0
34. Inouye SK. Delirium—a framework to improve acute care for older persons. *J Am Geriatr Soc*. 2018;66(3):446–451. doi:10.1111/jgs.15296
35. Boast J. The Lancet. Making more of multimorbidity: an emerging priority. *Lancet*. 2018;391(10131):1637. doi:10.1016/S0140-6736(18)30941-3
36. Whitty CJM, MacEwen C, Goddard A, et al. Rising to the challenge of multimorbidity. *BMJ*. 2020;368–369:l6964. doi:10.1136/bmj.l6964
37. Romero-Ortuno R, Wallis S, Biram R, Keevil V. Clinical frailty adds to acute illness severity in predicting mortality in hospitalized older adults: an observational study. *Eur J Intern Med*. 2016;35:24–34. doi:10.1016/j.ejim.2016.08.033
38. Stoltzfus JC. Logistic regression: a brief primer. *Acad Emerg Med*. 2011;18(10):1099–1104. doi:10.1111/j.1553-2712.2011.01185.x
39. Minne L, Ludikhuize J, de Jonge E, de Rooij S, Abu-Hanna A. Prognostic models for predicting mortality in elderly ICU patients: a systematic review. *Intensive Care Med*. 2011;37(8):1258–1268. doi:10.1007/s00134-011-2265-6
40. Wongvibulsin S, Garibaldi BT, Antar AAR, et al. Development of severe COVID-19 adaptive risk predictor (SCARP), a calculator to predict severe disease or death in hospitalized patients with COVID-19. *Ann Intern Med*. 2021;174(6):777–785. doi:10.7326/M20-6754
41. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*. 2019;572(7767):116–119. doi:10.1038/s41586-019-1390-1
42. Hyland SL, Faltys M, Hüser M, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat Med*. 2020;26(3):364–373. doi:10.1038/s41591-020-0789-4
43. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med*. 2020;3:1–10. doi:10.1038/s41746-020-0221-y
44. Churpek MM, Carey KA, Edelson DP, et al. Internal and external validation of a machine learning risk score for acute kidney injury. *JAMA Netw Open*. 2020;3(8):e2012892. doi:10.1001/jamanetworkopen.2020.12892
45. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med*. 2018;46(4):547–553. doi:10.1097/CCM.0000000000002936
46. Oliver D. David Oliver: what has the pandemic taught us about using frailty scales? *BMJ*. 2021;374:n1683. doi:10.1136/bmj.n1683
47. Flaatten H, Guidet B, Andersen FH, et al. Reliability of the Clinical Frailty Scale in very elderly ICU patients: a prospective European study. *Ann Intensive Care*. 2021;11(1):22. doi:10.1186/s13613-021-00815-7