Title: Evaluating Prognostic Bias of Critical Illness Severity Scores Based on Age, Sex, and Primary Language in the USA: A Retrospective Multicenter Study

Authors: Xiaoli Liu PhD[1,2,3*]; Max Shen MD[4*]; Margaret Lie MD[4*]; Zhongheng Zhang MD[5]; Chao Liu MD[6]; Deyu Li PhD[2]; Roger G. Mark PhD[3]; Zhengbo Zhang PhD[1,2#]; Leo Anthony Celi MD[3,4,7§]

co-first authors[*], corresponding author[#], senior author[§]

[1]Center for Artificial Intelligence in Medicine, The General Hospital of PLA, Beijing, China;

[2]School of Biological Science and Medical Engineering, Beihang University, Beijing, China;

[3]Laboratory for Computational Physiology, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, United States of America;

[4]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, United States of America;

[5]Department of Emergency Medicine, Key Laboratory of Precision Medicine in Diagnosis and Monitoring Research of Zhejiang Province, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China;

[6]Department of Critical Care Medicine, The First Medical Center, The General Hospital of PLA, Beijing, China;

[7]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, United States of America

Corresponding:

Zhengbo Zhang

E-mail: zhangzhengbo@301hospital.com.cn

Telephone: (86) 13693321644

Fax: 86/010-82317613

Center for Artificial Intelligence in Medicine, The General Hospital of PLA

No. 28 Fuxing Road, Haidian District, Beijing, 100853, CN.

Funding

ZBZ is funded by the National Natural Science Foundation of China (62171471). LAC is funded by the National Institute of Health through NIBIB R01 EB017205.

Conflicts of interest

The authors declare that they have no competing interests.

Keywords: Illness Severity Scores, Bias Evaluation, Hospital Mortality, Discrimination, Calibration.

Table of Contents

Supplementary Material: Evaluating Prognostic Bias of Critical Illness Severity Scores Based on Age, Sex, and Primary Language in the USA: A Retrospective Multicenter Study

Xiaoli Liu[1,2,3*]; Max Shen[4*]; Margaret Lie[4*]; Zhongheng Zhang[5]; Chao Liu[6]; Deyu Li[2]; Roger Mark[3]; Zhengbo Zhang[1,2#]; Leo Anthony Celi[3,4,7§]

Supplemental Methods

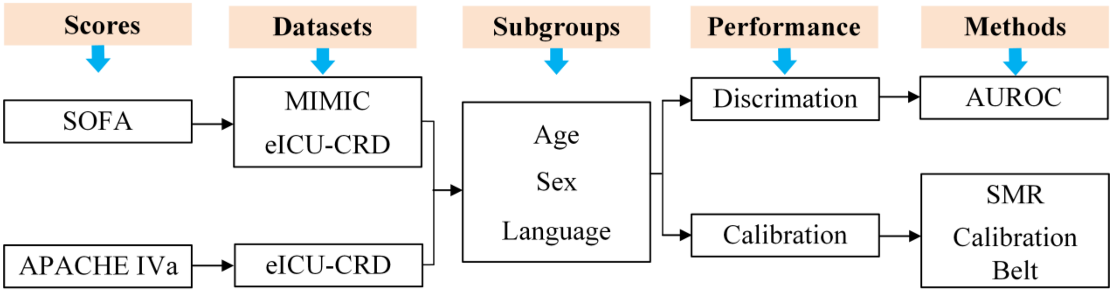
In this section, we present an expanded narrative of our statistical analysis methods and other details. Details regarding data sources were described in the main Method section. We conducted a retrospective analysis study to evaluate the performance of SOFA and APACHE IVa scores in based on two large ICU databases, Medical Information Mart for Intensive Care (MIMIC) and eICU Collaborative Research Database (eICU-CRD).

Descriptive statistics of patient characteristics were reported using median (25th, 75th) percentiles (IQR) or proportions. Groups were compared using student's *t*-test or X^2 test for categorical variables and Wilcoxon rank-sum test or Kruskal-Wallis test for continuous variables, as appropriate.

Analysis was conducted across subgroups created from the following variables: age (16-44, 45-64, 75-79, and 80 and older), sex (female and male), and primary language (English and Non-English). In hospital mortality was selected as the outcome of interest. Since SOFA was not initially created for mortality prediction, we utilized 20% of randomly selected encounters fitted a univariate logistic regression model in MIMIC and eICU-CRD, respectively [1-3]. For APACHE IVa score, the mortality prediction of each eICU patient encounter had already been calculated in the databases based on published algorithm and therefore was directly imported.

Discrimination and calibration of both SOFA and APACHE IVa scores in mortality prediction were evaluated for the overall databases as well as by subgroups described above. For discrimination, we evaluated SOFA and APACHE IVa's performance using area under receiver operating characteristic (AUROC) curve. AUROC is commonly used to assess the ability of a classifier to discriminate between binary outcomes at each threshold [4,5]. It is calculated for both scores across all cohorts, and differences between subgroups are compared using Kruskal-Wallis test. For calibration, standardised mortality ratio (SMR) was calculated to analyze each score's performance. Forest plot was generated to compare SMRs between subgroups. Significant difference between SMRs were evaluated using Kruskal-Wallis test. For SOFA, calibration was additionally assessed according to the increasing severity levels with score categories (≤7, 8-11, and >11) and predicted mortality categories (0-5%, 5-10%, 10-20%, 20-50%, >50%) within each subgroup [4]. The GiViTI (Italian Group for the Evaluation of Intervention in Intensive Care Medicine) calibration belt, another tool for calibration assessment, was also adopted to detect deviations of logits of predicted probabilities, generated by a model, from observed probabilities. It is a graphical tool to display calibration curve with confidence level, fitted by a polynomial function. The resulting coefficient (polynomial degree) and belt deviation from the bisector (under or over) with various confidence levels (such as 80% and 95%) were used together to evaluate for potential miscalibration of a predictive model [6].

Univariate and multivariate LR models were constructed using mortality as dependent variable and SOFA or APACHE IVa, age, sex, and primary language as independent variable in MIMIC and eICU-CRD cohorts. One thousand-fold Bootstrap resampling iteration was used to calculate 95% CI, and 2-tailed $P < 0.05$ was used as a threshold for statistical significance for all analyses described above.



eFigure 1. Study Analysis Process Map

The de-identification and anonymization were both strictly implemented in the MIMIC and eICU-CRD databases. This study was exempt from institutional review board approval due to the retrospective design, and the security schema that certified re-identification risk to meet safe harbor standards by an independent privacy expert (Institutional Review Boards (IRBs): No. 0403000206 [MIT, MIMIC], 2001-P-001699/14 [BIDMC, MIMIC], and No. 1031219-2 [Health Insurance Portability & Accountability Act, eICU-CRD]). All statistical analyses were performed using Python version 3.8 (sklearn, pyroc, scipy, and tableone package) and R version 4.1.1 (ems, dplyr, forestplot, givitiR, gbm, and rsq package).

Code and data sharing: we extracted data based on the MIMIC-III, eICU-CRD, and MIMIC-IV databases, bias evaluations, and statistical analysis are available at https://github.com/liuxiaoliXRZS/clinical_scores_bias. Data set: we plan to share it on the PhysioNet website (https://physionet.org/about/database/).

Reference

[1] Soo, Andrea, et al. "Describing organ dysfunction in the intensive care unit: a cohort study of 20,000 patients." Critical Care 23.1 (2019): 1-15.

[2] Minne, Lilian, Ameen Abu-Hanna, and Evert de Jonge. "Evaluation of SOFA-based models for predicting mortality in the ICU: A systematic review." Critical care 12.6 (2008): 1-13.

[3] Ferreira, Flavio Lopes, et al. "Serial evaluation of the SOFA score to predict outcome in critically ill patients." Jama 286.14 (2001): 1754-1758.

[4] Sarkar, Rahuldeb, et al. "Performance of intensive care unit severity scoring systems across different ethnicities in the USA: a retrospective observational study." The Lancet Digital Health 3.4 (2021): e241-e249.

[5] Alba, Ana Carolina, et al. "Discrimination and calibration of clinical prediction models: users' guides to the medical literature." Jama 318.14 (2017): 1377-1384.

[6] Nattino, Giovanni, Stefano Finazzi, and Guido Bertolini. "A new calibration test and a reappraisal of the calibration belt for the assessment of prediction models based on dichotomous outcomes." Statistics in medicine 33.14 (2014): 2390-2407.

eTable 1. Patient Characteristics of MIMIC and eICU-CRD Study Cohorts

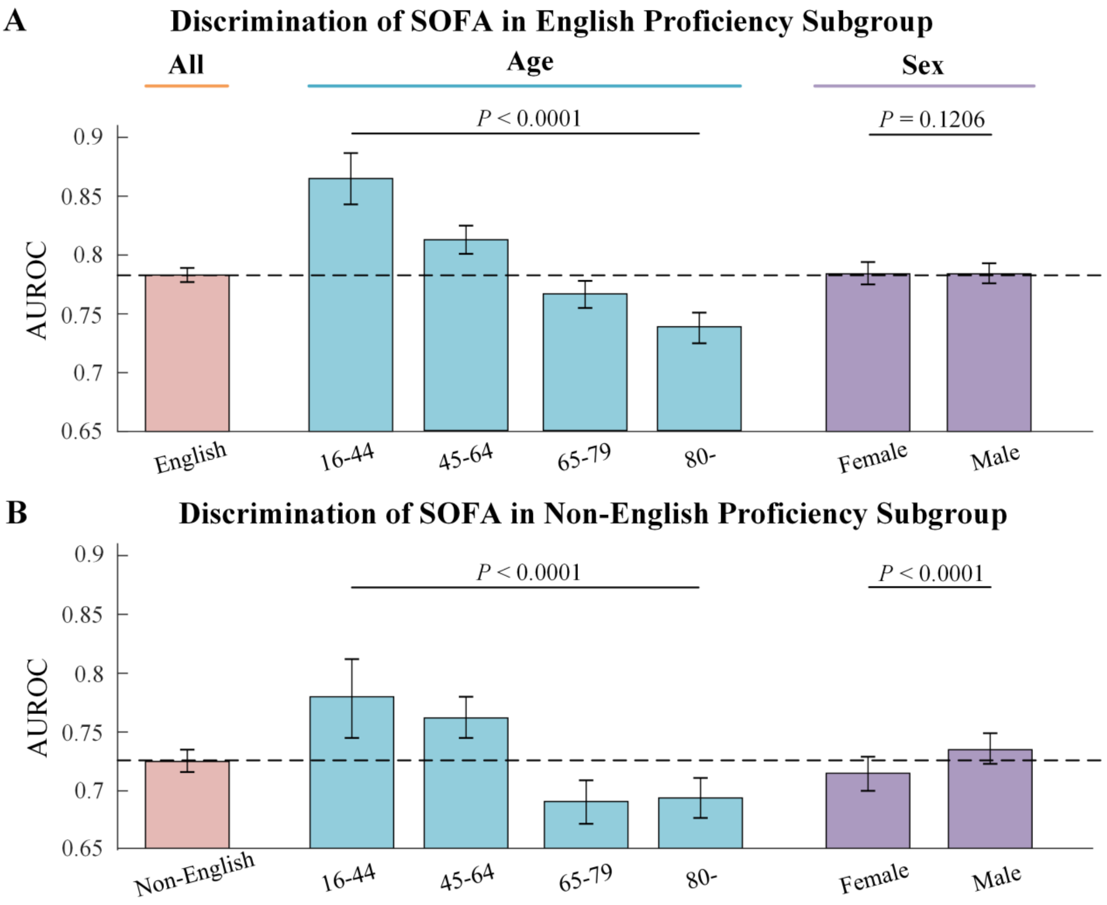
Variable	MIMIC (96,029, 10.4% mortality)				eICU-CRD (100,281, 8.8% mortality)			
	Overall	Survivors	Non-Survivors	*P*-Value	Overall	Survivors	Non-Survivors	*P*-Value
Total number (%)	96029	86067 (89.6)	9962 (10.4)		100281	91436 (91.2)	8845 (8.8)	
Age, median [Q1, Q3]	66.0 [54.0,78.0]	65.0 [53.0,77.0]	73.0 [61.0,83.0]	<0.001	65.0 [53.0,76.0]	65.0 [53.0,76.0]	71.0 [60.0,81.0]	<0.001
Age group (%)				<0.001				<0.001
16-44	12838 (13.4)	12214 (14.2)	624 (6.3)		13473 (13.4)	12895 (14.1)	578 (6.5)	
45-64	31743 (33.1)	29211 (33.9)	2532 (25.4)		34658 (34.6)	32228 (35.2)	2430 (27.5)	
65-79	31198 (32.5)	27807 (32.3)	3391 (34.0)		33714 (33.6)	30340 (33.2)	3374 (38.1)	
80-	20250 (21.1)	16835 (19.6)	3415 (34.3)		18436 (18.4)	15973 (17.5)	2463 (27.8)	
Sex (%)				<0.001				0.757
Female	42330 (44.1)	37731 (43.8)	4599 (46.2)		45539 (45.4)	41508 (45.4)	4031 (45.6)	
Male	53699 (55.9)	48336 (56.2)	5363 (53.8)		54742 (54.6)	49928 (54.6)	4814 (54.4)	
BMI, median [Q1, Q3]	27.5 [23.9,32.0]	27.6 [24.0,32.1]	26.5 [22.8,31.2]	<0.001	27.5 [23.5,32.9]	27.6 [23.6,32.9]	26.7 [22.7,32.2]	<0.001
Primary language (%)				<0.001				
English	70269 (73.2)	63900 (74.2)	6369 (63.9)		--	--	--	
Non-English	25760 (26.8)	22167 (25.8)	3593 (36.1)		--	--	--	
Ethnicity (%)				<0.001				0.001
Asian	2611 (2.7)	2315 (2.7)	296 (3.0)		1461 (1.5)	1316 (1.4)	145 (1.6)	
Black	9879 (10.3)	9071 (10.5)	808 (8.1)		11696 (11.7)	10761 (11.8)	935 (10.6)	
Hispanic	3411 (3.6)	3178 (3.7)	233 (2.3)		3860 (3.8)	3523 (3.9)	337 (3.8)	
Other	14381 (15.0)	12229 (14.2)	2152 (21.6)		6253 (6.2)	5747 (6.3)	506 (5.7)	

White	65747 (68.5)	59274 (68.9)	6473 (65.0)		77011 (76.8)	70089 (76.7)	6922 (78.3)	
SOFA, median [Q1, Q3]	4.0 [2.0,6.0]	4.0 [2.0,6.0]	7.0 [5.0,11.0]	<0.001	5.0 [3.0,7.0]	5.0 [3.0,7.0]	8.0 [5.0,10.0]	<0.001
APACHE IVa, median [Q1, Q3]	--	--	--	--	53.0 [39.0,71.0]	51.0 [38.0,67.0]	83.0 [64.0,106.0]	<0.001
CCI score, median [Q1, Q3]	5.0 [3.0,7.0]	5.0 [3.0,7.0]	7.0 [5.0,9.0]	<0.001	4.0 [2.0,5.0]	4.0 [2.0,4.0]	4.0 [3.0,5.0]	<0.001
Ventilation (%)	33808 (35.2)	28271 (32.8)	5537 (55.6)	<0.001	38732 (38.6)	32652 (35.7)	6080 (68.7)	<0.001
Advance directives (%)	4666 (4.9)	3064 (3.6)	1602 (16.1)	<0.001	9272 (9.2)	7060 (7.7)	2212 (25.0)	<0.001
DNR/DNI	4210 (4.4)	2932 (3.4)	1278 (12.8)	<0.001	9237 (9.2)	7041 (7.7)	2196 (24.8)	<0.001
Comfort measures only	456 (0.5)	132 (0.2)	324 (3.3)	<0.001	328 (0.3)	136 (0.1)	192 (2.2)	<0.001
Pre ICU LOS day, median [Q1, Q3]	0.1 [0.0,0.6]	0.1 [0.0,0.6]	0.1 [0.0,0.8]	<0.001	0.2 [0.0,0.4]	0.2 [0.0,0.4]	0.1 [0.0,0.7]	0.001
LOS ICU day, median [Q1, Q3]	2.0 [1.1,3.9]	1.9 [1.1,3.5]	3.2 [1.6,7.1]	<0.001	2.2 [1.5,3.9]	2.1 [1.4,3.8]	3.5 [1.9,6.7]	<0.001
LOS hospital day, median [Q1, Q3]	6.8 [4.0,11.8]	6.8 [4.1,11.6]	6.8 [2.8,13.9]	<0.001	6.1 [3.6,10.3]	6.1 [3.6,10.2]	5.9 [3.0,11.6]	<0.001

CCI: Charlson Comorbidity Index, LOS: length of stay. DNR: do not resuscitate, DNI: do not intubate, Advance directives: DNR/DNI/Comfort measures only

eTable 2. Discrimination Performance of SOFA and APACHE IVa by Age, Sex, and Primary Language in MIMIC and eICU-CRD Cohorts

Name	Subgroup (*p value*)	AUROC (MIMIC, SOFA)	AUROC (eICU, SOFA)	AUROC (eICU, APACHE IVa)
All		0.761 (0.755-0.766)	0.73 (0.724-0.736)	0.828 (0.823-0.833)
Age	16-44	0.827 (0.806-0.846)	0.812 (0.793-0.83)	0.886 (0.87-0.903)
	45-64	0.792 (0.782-0.803)	0.768 (0.758-0.779)	0.844 (0.835-0.852)
	65-79	0.74 (0.73-0.75)	0.712 (0.702-0.723)	0.81 (0.802-0.819)
	>=80	0.721 (0.711-0.732)	0.678 (0.665-0.69)	0.761 (0.751-0.772)
	P value	<0.001	<0.001	<0.001
Sex	Female	0.759 (0.751-0.767)	0.727 (0.718-0.736)	0.823 (0.815-0.831)
	Male	0.764 (0.757-0.771)	0.733 (0.725-0.74)	0.832 (0.825-0.838)
	P value	<0.001	<0.001	<0.001
Language	English	0.783 (0.776-0.789)	--	--
	Non-English	0.726 (0.716-0.735)	--	--
	P value	<0.001	--	--



eFigure 2. Discrimination Performance of SOFA in Subgroups Divided by Primary

Language and Age or Sex in MIMIC Cohort

eTable 3. Comparison of Discrimination Performance by Primary Language and Age or Sex in MIMIC Cohort

Subgroup (Primary Language)	Cross subgroup (*P* value)	AUROC
English	Age (16-44)	0.865 (0.843-0.886)
	Age (45-64)	0.813 (0.801-0.825)
	Age (65-79)	0.767 (0.755-0.778)
	Age (80-)	0.739 (0.725-0.751)
	P value	<0.001
Non-English	Age (16-44)	0.78 (0.745-0.812)
	Age (45-64)	0.762 (0.745-0.78)
	Age (65-79)	0.691 (0.672-0.709)
	Age (80-)	0.694 (0.677-0.711)
	P value	<0.001
English	Sex (Female)	0.784 (0.775-0.794)
	Sex (Male)	0.784 (0.776-0.793)
	P value	0.1206
Non-English	Sex (Female)	0.715 (0.7-0.729)
	Sex (Male)	0.735 (0.723-0.749)
	P value	<0.001

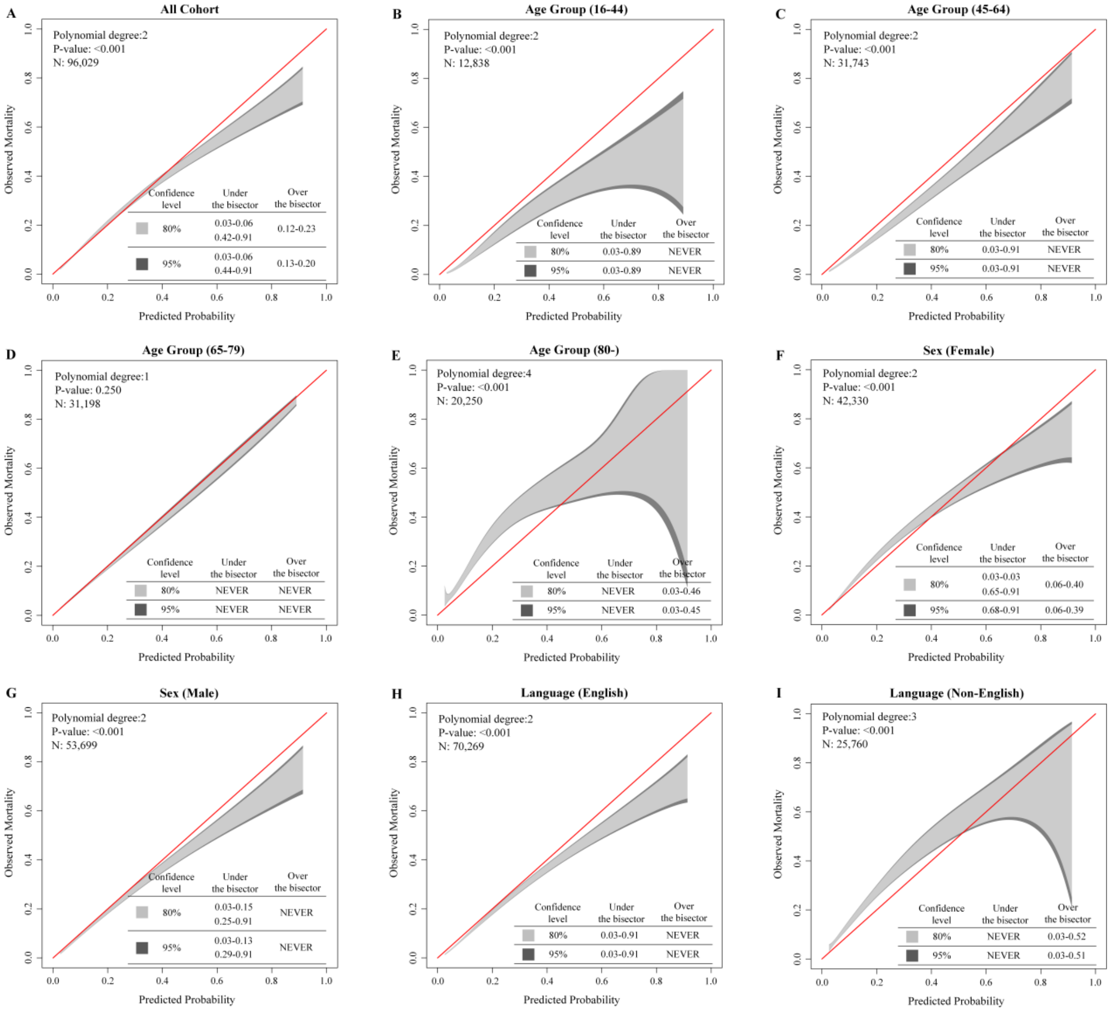
eTable 4. Expected Mortality Predicted by SOFA and APACHE IVa Score Compared to Observed Mortality by Age, Sex, and Primary Language Subgroups

Dataset (score)	Subgroup	Total Patients	Observation	Expectation	SMR (95% CI)	*P* value
MIMIC (SOFA)	All	96029	9962	10167	0.98 (0.96-1.00)	--
	Age					<0.001
	16-44	12838	624	1109	0.56 (0.51-0.62)	
	45-64	31743	2532	3414	0.74 (0.71-0.77)	
	65-79	31198	3391	3438	0.99 (0.96-1.02)	
	80-	20250	3415	2206	1.55 (1.51-1.59)	
	Sex					<0.001

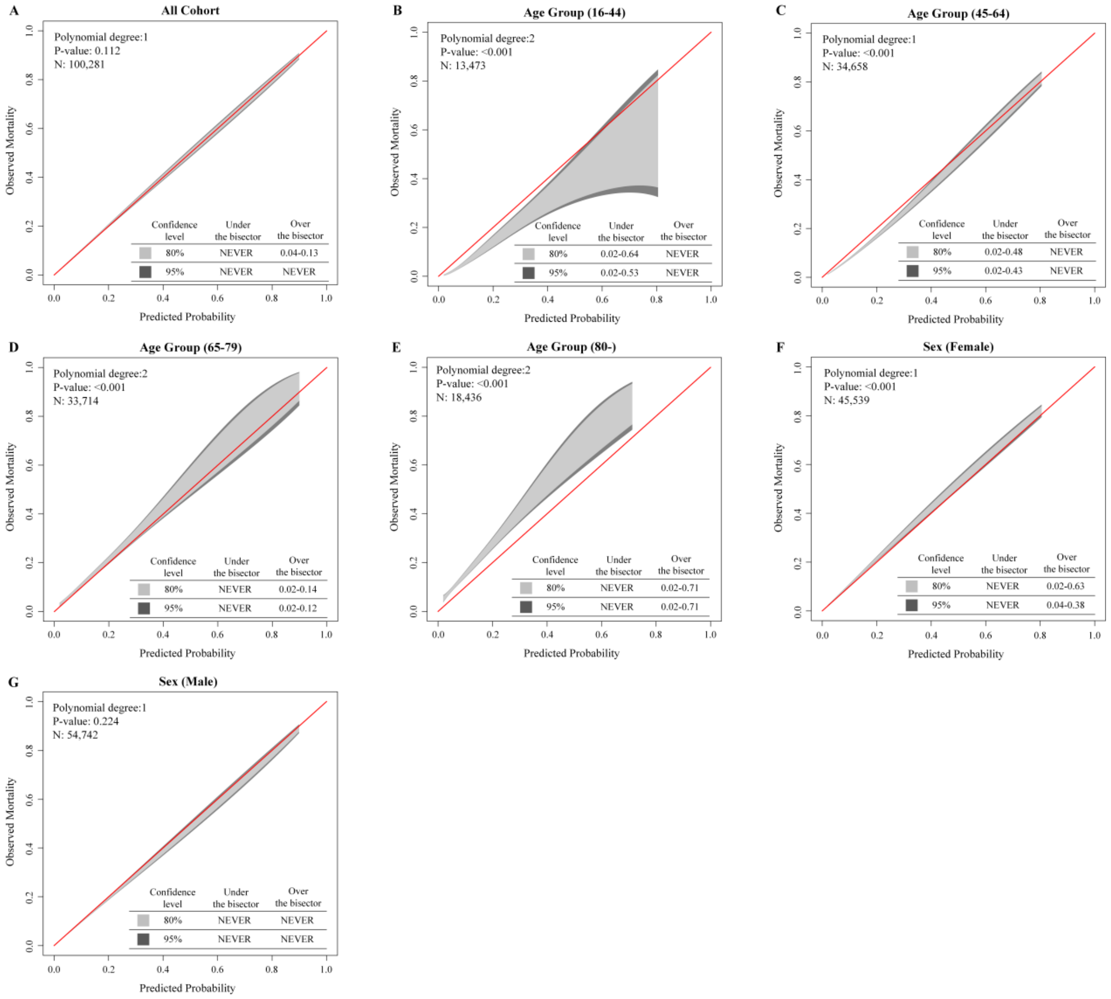
	Female	42330	4599	4214	1.09 (1.06-1.12)	
	Male	53699	5363	5954	0.90 (0.88-0.92)	
	Language					<0.001
	English	70269	6369	7565	0.84 (0.82-0.86)	
	Non-English	25760	3593	2603	1.38 (1.35-1.42)	
eICU (SOFA)	All	100281	8845	8679	1.02 (1.00-1.04)	--
	Age					<0.001
	16-44	13473	578	1031	0.56 (0.51-0.62)	
	45-64	34658	2430	3055	0.80 (0.76-0.83)	
	65-79	33714	3374	3028	1.11 (1.08-1.15)	
	80-	18436	2463	1565	1.57 (1.53-1.62)	
	Sex					<0.001
	Female	45539	4031	3778	1.07 (1.04-1.10)	
	Male	54742	4814	4901	0.98 (0.96-1.01)	
eICU (APACHE IVa)	All	100281	8845	12496	0.71 (0.69-0.72)	--
	Age					<0.001
	16-44	13473	578	929	0.62 (0.57-0.68)	
	45-64	34658	2430	3530	0.69 (0.66-0.72)	
	65-79	33714	3374	4581	0.74 (0.71-0.76)	
	80-	18436	2463	3456	0.71 (0.69-0.74)	
	Sex					<0.001
	Female	45539	4031	5837	0.69 (0.67-0.71)	
	Male	54742	4814	6659	0.72 (0.70-0.74)	

Calibration of SOFA and APACHE IVa models was also evaluated via calibration belts shown in **eFigure 3-5**. Results were generally similar to that of SMRs previously described. For SOFA score in both MIMIC and eICU-CRD cohorts, calibration belts of 16-44 years, 45-

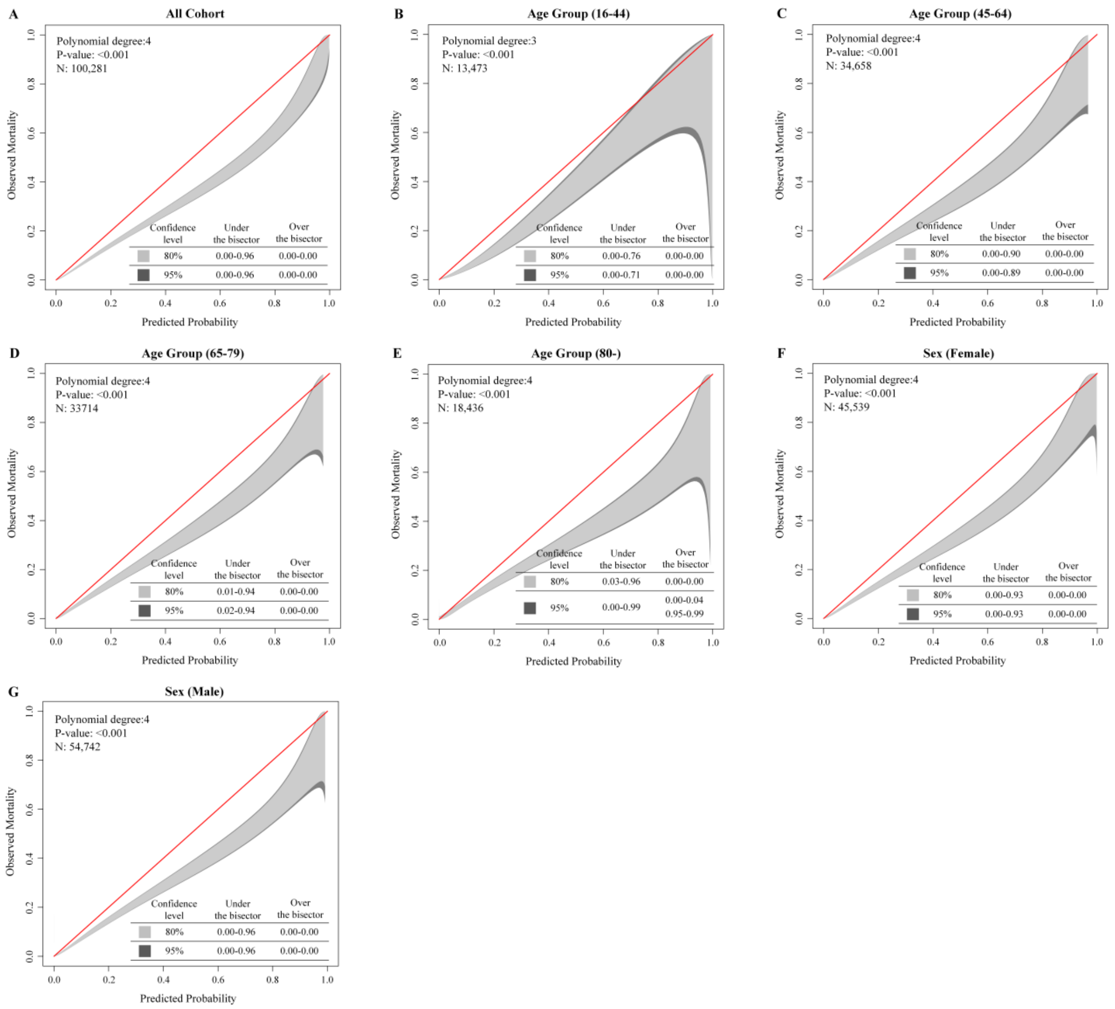
64 years, male, and English primary language speakers (MIMIC only) subgroups were consistently under the bisector suggesting overestimation of in-hospital mortality. In contrast, the ≥80 years and non-English primary language speakers group (MIMIC only) were consistently over the bisector suggesting underestimation of mortality. For APACHE IVa score in eICU-CRD, calibration belts for all subgroups were under the bisector, suggesting an overestimation of mortality overall.



eFigure 3. GiViTI Calibration Belt for SOFA Score in MIMIC by Age, Sex, and Primary Language. *(A)* Overall Cohort, *(B)-(E)* Age Groups (16-44, 45-64, 65-79, and ≥80), *(F)* and *(G)* Sex (Female and Male), *(H)* and *(I)* Primary Language (English and Non-English)

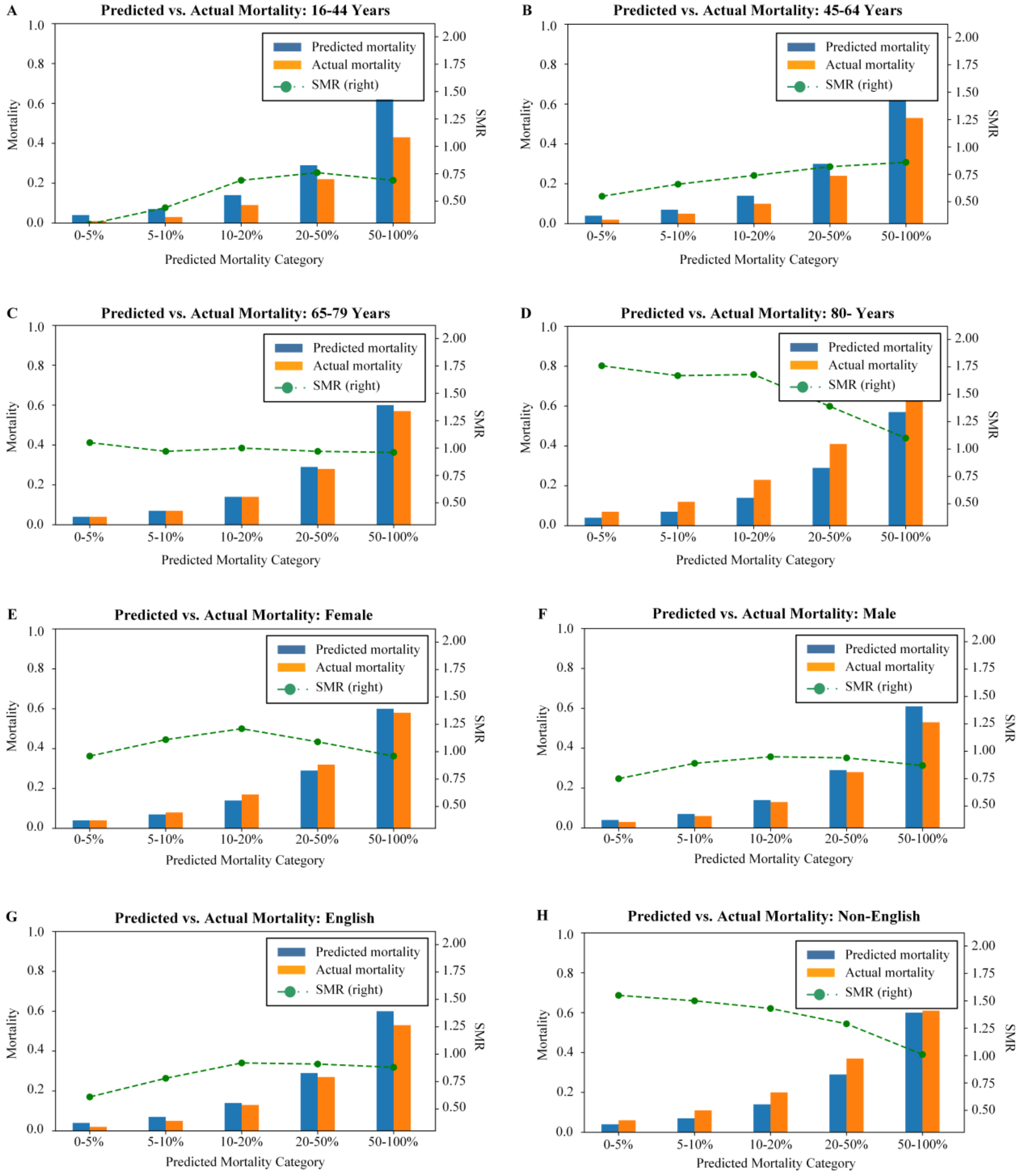


eFigure 4. GiViTI Calibration Belt for SOFA Score in eICU-CRD by Age and Sex. *(A)* Overall Cohort, *(B)-(E)* Age Groups (16-44, 45-64, 65-79, and ≥80), *(F)* and *(G)* Sex (Female and Male)

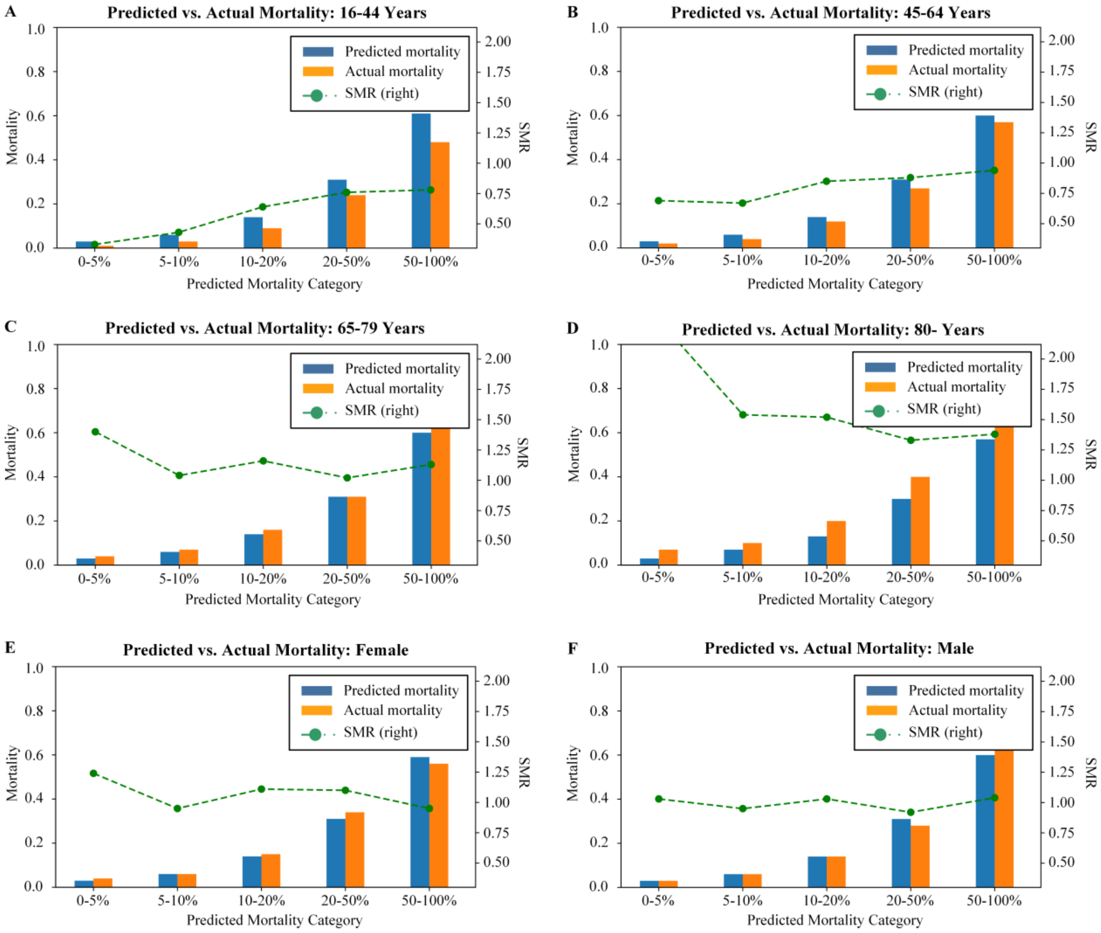


eFigure 5. GiViTI Calibration Belt for APACHE IVa Score in eICU-CRD by Age and Sex. *(A)* Overall Cohort, *(B)-(E)* Age Groups (16-44, 45-64, 65-79, and ≥80), *(F)* and *(G)* Sex (Female and Male)

Additionally, we evaluated calibration performance of SOFA score for each subgroup, further divided by predicted mortality of 0-5%, 5-10%, 10-20%, 20-50%, and 50-100%, as presented in **eFigure 6**, **eFigure 7**, and **eTable 5** for both databases. For age groups 16-44 and 45-64, SMR was lower when compared to their respective age group's average SMR for lower risk patients and opposite for higher risk patients, while SMR was higher compared to average for lower risk patients in ≥80. In particular, SMR was as low as around 0.3 in 16-44 /0-5% mortality group but as high as 2.3 in ≥80/0-5% mortality group. There was no consistent pattern of SMR across predicted mortality for both male and female subgroups. For primary English speakers, SMR was as low as 0.61 for 0-5% mortality group but increased to ~0.9 for mortality >10%. For non-English primary speakers, SMR was as high as 1.55 for 0-5% mortality group but decreased to 1.01 for 50-100% mortality group. We also performed a similar analysis utilizing three SOFA score categories (0-7, 8-11, over 11) of increasing disease severity. Results were in **eTable 6** and are largely similar to the above analysis.



eFigure 6. Predicted Mortality, Observed Mortality, and SMR for SOFA Score by Predicted Mortality Risk Groups and Age, Sex, or Primary Language Subgroups in MIMIC Cohort. *(A)-(D)* Age Groups 16-44, 45-64, 65-79, and ≥ 80; *(E)* and *(F)* Sex (Female and Male); *(G)* and *(H)* Primary Language (English and Non-English); Left y-axis: Predicted Mortality; Right, y-axis: SMR



eFigure 7. Predicted Mortality, Observed Mortality, and SMR for SOFA Score by Predicted Mortality Risk Groups and Age, Sex, or Primary Language Subgroups in eICU-CRD Cohort. *(A)-(D)* Age Groups 16-44, 45-64, 65-79, and ≥ 80; *(E)* and *(F)* Sex (Female and Male); Predicted Mortality: y-axis, Left; SMR: y-axis, Right

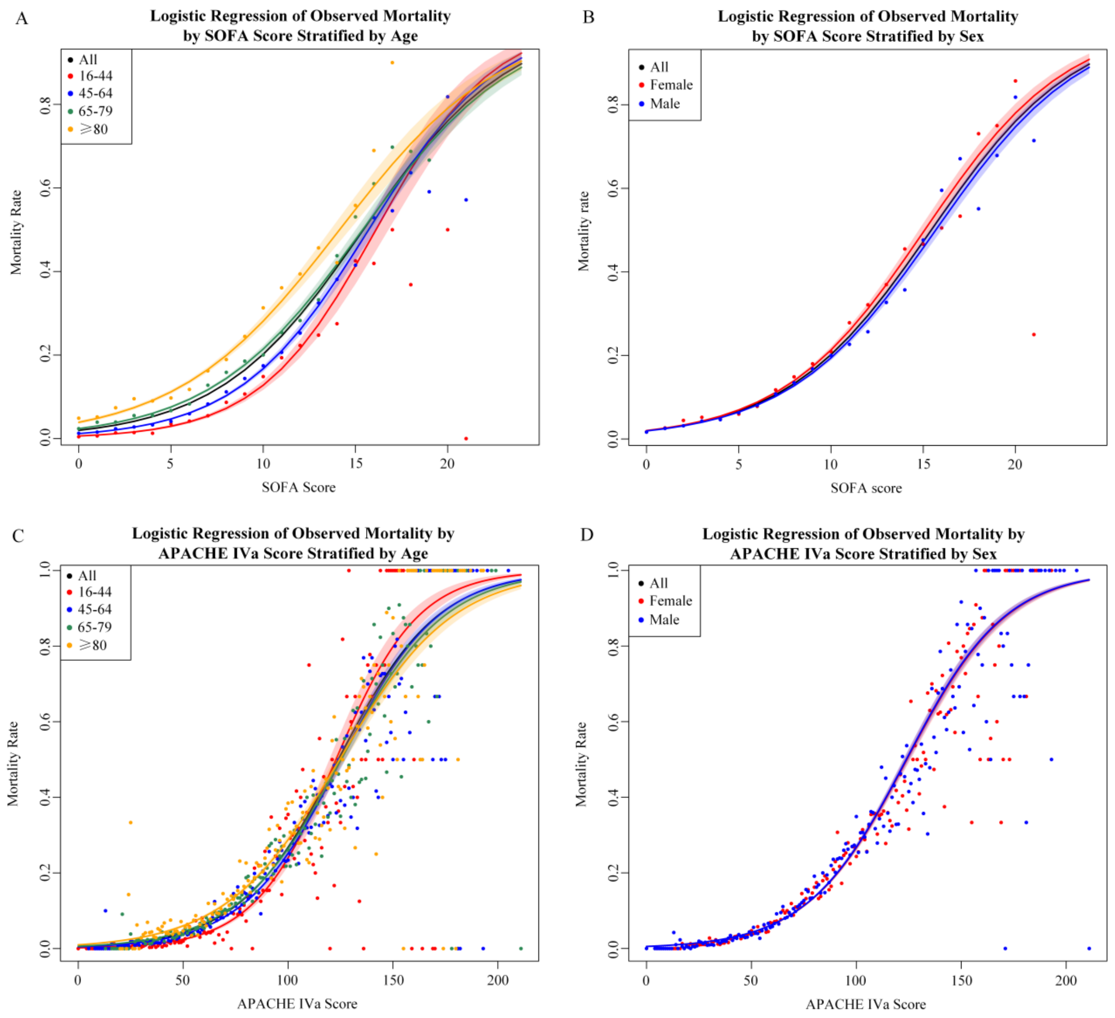
eTable 5. SMR of SOFA Score by Predicted Mortality Categories in Various Subgroups

Predicted Mortality Category	**Data set**	**Age (16-44)**	**Age (45-64)**	**Age (65-79)**	**Age (80-)**	**Female**	**Male**	**English**	**Non-English**
0-5%	MIMIC	0.29 (0.19-0.45)	0.55 (0.46-0.65)	1.05 (0.96-1.16)	1.76 (1.63-1.9)	0.96 (0.89-1.05)	0.75 (0.68-0.84)	0.61 (0.54-0.67)	1.55 (1.44-1.66)
	eICU	0.33 (0.21-0.53)	0.69 (0.59-0.8)	1.4 (1.29-1.52)	2.32 (2.18-2.48)	1.24 (1.15-1.33)	1.03 (0.95-1.12)	--	--
5-10%	MIMIC	0.44 (0.34-0.57)	0.66 (0.6-0.74)	0.97 (0.91-1.04)	1.67 (1.59-1.75)	1.11 (1.05-1.17)	0.89 (0.84-0.94)	0.78 (0.74-0.83)	1.5 (1.43-1.57)

	eICU	0.43 (0.33-0.55)	0.67 (0.6-0.73)	1.04 (0.98-1.11)	1.54 (1.46-1.63)	0.95 (0.89-1.0)	0.95 (0.9-1.01)	--	--
10-20%	MIMIC	0.69 (0.57-0.83)	0.74 (0.68-0.81)	1.0 (0.94-1.07)	1.68 (1.61-1.76)	1.21 (1.16-1.27)	0.95 (0.9-1.0)	0.92 (0.88-0.96)	1.43 (1.36-1.5)
	eICU	0.64 (0.55-0.75)	0.85 (0.8-0.91)	1.16 (1.1-1.21)	1.52 (1.45-1.6)	1.11 (1.07-1.17)	1.03 (0.99-1.07)	--	--
20-50%	MIMIC	0.76 (0.66-0.87)	0.82 (0.77-0.88)	0.97 (0.92-1.03)	1.39 (1.33-1.46)	1.09 (1.05-1.14)	0.94 (0.9-0.98)	0.91 (0.87-0.94)	1.29 (1.23-1.35)
	eICU	0.76 (0.65-0.89)	0.88 (0.82-0.95)	1.02 (0.96-1.08)	1.33 (1.24-1.43)	1.1 (1.04-1.16)	0.92 (0.87-0.97)	--	--
50-100%	MIMIC	0.69 (0.6-0.79)	0.86 (0.81-0.91)	0.96 (0.9-1.02)	1.1 (1.01-1.19)	0.96 (0.91-1.02)	0.87 (0.83-0.91)	0.88 (0.84-0.92)	1.01 (0.93-1.08)
	eICU	0.78 (0.63-0.97)	0.94 (0.85-1.05)	1.13 (1.0-1.27)	1.38 (1.15-1.65)	0.95 (0.84-1.07)	1.04 (0.95-1.13)	--	--

eTable 6. SMR of SOFA Score by Score Categories in Various Subgroups

SOFA Score	**Dataset**	**Age (16-44)**	**Age (45-64)**	**Age (65-79)**	**Age (80-)**	**Female**	**Male**	**English**	**Non-English**
0-7	MIMIC	0.43 (0.36-0.51)	0.65 (0.6-0.69)	1.0 (0.96-1.05)	1.7 (1.65-1.76)	1.1 (1.06-1.14)	0.88 (0.84-0.91)	0.78 (0.75-0.81)	1.49 (1.44-1.54)
	eICU	0.42 (0.34-0.51)	0.7 (0.65-0.75)	1.15 (1.1-1.2)	1.69 (1.63-1.76)	1.04 (1.0-1.09)	0.99 (0.95-1.03)	--	--
8-11	MIMIC	0.82 (0.72-0.95)	0.83 (0.77-0.89)	0.98 (0.93-1.04)	1.5 (1.43-1.57)	1.17 (1.12-1.22)	0.95 (0.91-1.0)	0.94 (0.9-0.98)	1.32 (1.26-1.39)
	eICU	0.71 (0.61-0.82)	0.87 (0.81-0.93)	1.11 (1.06-1.17)	1.5 (1.42-1.58)	1.11 (1.06-1.17)	1.0 (0.96-1.05)	--	--
>11	MIMIC	0.67 (0.6-0.76)	0.84 (0.8-0.89)	0.96 (0.91-1.01)	1.18 (1.11-1.25)	0.97 (0.92-1.01)	0.89 (0.86-0.93)	0.87 (0.83-0.9)	1.13 (1.07-1.19)
	eICU	0.76 (0.66-0.87)	0.91 (0.85-0.97)	1.02 (0.96-1.09)	1.27 (1.17-1.38)	1.05 (0.99-1.11)	0.93 (0.89-0.98)	--	--



eFigure 8. Logistic Regression Models of Observed Mortality by SOFA or APACHE IVa Scores in eICU-CRD. (A) SOFA Score Stratified by Age, (B) SOFA Score Stratified by Sex, (C) APACHE IVa Score Stratified by Age, (D) APACHE IVa Score Stratified by Sex

Details of LR models that relate SOFA or APACHE IVa scores to mortality were shown in **eTable 7**. In **eTable 8**, age groups were added to LR models, using youngest patients (16-44) as baseline. A higher relative risk was observed with increasing age in all cohorts, and R^2 improved when compared to models without age factor. In **eTable 9**, sex was added to the original LR models using female patients as baseline. **eTable 10** showed LR mortality model with SOFA score and primary language as variable. In terms of mortality, non-English primary speakers have an odds ratio of 1.9 (1.81-1.99) compared to English primary speakers that only decreased to 1.82 (1.74-1.91) if age and sex are included in the LR model (**eTable 11**).

eTable 7. Logistic Regression Models of Mortality by SOFA or APACHE IVa Scores

Variable	Estimate	Std. Error	z value	Pr(>|z|)	R^2
MIMIC (SOFA)					
(Intercept)	-3.642574426	0.022538841	-161.613211	<0.0001	0.138676
SOFA	0.258815824	0.002861707	90.44105136	<0.0001	
eICU (SOFA)					
(Intercept)	-3.90392169	0.02631375	-148.3605205	<0.0001	0.105191
SOFA	0.252956185	0.003263728	77.50528108	<0.0001	
eICU (APACHE IVa)					
(Intercept)	-5.216240121	0.035977443	-144.9864065	<0.0001	0.1823068
APACHE IVa	0.042097441	0.000436572	96.42735278	<0.0001	

eTable 8. Logistic Regression Models of Mortality by Age and SOFA or APACHE IVa Score

Variable	Estimate	Std. Error	z value	Pr(>|z|)	R^2
MIMIC					
(Intercept)	-4.35326327	0.048284667	-90.15829518	<0.0001	0.1600516
SOFA	0.263074946	0.002930015	89.78622051	<0.0001	
Age (45-64)	0.321576382	0.049176121	6.539279077	<0.0001	
Age (65-79)	0.690038382	0.047906552	14.40384151	<0.0001	
Age (80-)	1.282156899	0.048218309	26.5906652	<0.0001	
eICU					
(Intercept)	-4.630682118	0.050584535	-91.54343638	<0.0001	0.1220283
SOFA	0.257927224	0.003338335	77.26222807	<0.0001	
Age (45-64)	0.394132081	0.049291485	7.995946555	<0.0001	
Age (65-79)	0.798145649	0.048017777	16.62187824	<0.0001	
Age (80-)	1.218837178	0.049630344	24.55830617	<0.0001	
eICU					
(Intercept)	-5.535874621	0.056013646	-98.83082048	<0.0001	0.1851033
APACHE IVa	0.041549651	0.000441742	94.05865117	<0.0001	
Age (45-64)	0.254887686	0.051348195	4.963907425	<0.0001	

Age (65-79)	0.37998587	0.050064149	7.589979587	<0.0001	
Age (80-)	0.572195994	0.051594526	11.09024626	<0.0001	

eTable 9. Logistic Regression Models of Mortality by Sex and SOFA or APACHE IVa Score

Variable	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**	**R^2**
MIMIC					
(Intercept)	-3.519732714	0.024890204	-141.410361	<0.0001	0.1404939
SOFA	0.261078582	0.002874977	90.81067071	<0.0001	
Male	-0.246797773	0.022852574	-10.79956125	<0.0001	
eICU					
(Intercept)	-3.855610679	0.028620569	-134.7146769	<0.0001	0.1054817
SOFA	0.253725016	0.003271114	77.56532187	<0.0001	
Male	-0.097671345	0.023397524	-4.174430846	<0.0001	
eICU					
(Intercept)	-5.219732919	0.038294648	-136.3045011	<0.0001	0.1851033
APACHE IVa	0.04209699	0.000436565	96.42768585	<0.0001	
Male	0.006511843	0.024406776	0.2668047	0.789619535	

eTable 10. Logistic Regression Model of Mortality by Primary Language and SOFA Score

Variable	**OR (95% CI)**	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**	**R^2**
(Intercept)	0.02 (0.02-0.02)	-3.879371635	0.025154169	-154.2238023	<0.0001	0.1493607
SOFA	1.30 (1.30-1.31)	0.264908149	0.002904562	91.20417227	<0.0001	
Non-English	1.90 (1.81-1.99)	0.640793854	0.024082032	26.60879517	<0.0001	

eTable 11. Logistic Regression Model of Mortality by Primary Language, Age, Sex and SOFA Score

Variable	**OR (95% CI)**	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**	**R^2**
(Intercept)	0.01 (0.01-0.01)	-4.480833316	0.050974187	-87.90396897	<0.0001	0.1700973
SOFA	1.31 (1.30-1.32)	0.270039812	0.002978514	90.66258211	<0.0001	

Age (45-64)	1.40 (1.27-1.54)	0.337401239	0.049379502	6.832819793	<0.0001	
Age (65-79)	1.99 (1.81-2.19)	0.688750529	0.048098745	14.31951149	<0.0001	
Age (80-)	3.46 (3.14-3.80)	1.239842674	0.048484017	25.57219373	<0.0001	
Male	0.84 (0.81-0.88)	-0.169994841	0.023276072	-7.303416307	<0.0001	
Non-English	1.82 (1.74-1.91)	0.600327823	0.024352328	24.65176297	<0.0001	

eTable 12. Patient Characteristics by Primary Language in MIMIC

Variable	**MIMIC (96,029, 10.4% mortality)**			
	Overall	English	Non-English	*P*-Value
Number (%)	96029	70269	25760	
Outcome (%)				
Survivor	86067 (89.6)	63900 (90.9)	22167 (86.1)	<0.001
Non-survivor	9962 (10.4)	6369 (9.1)	3593 (13.9)	
Age, median [Q1,Q3]	66.0 [54.0,78.0]	66.0 [54.0,77.0]	68.0 [54.0,79.0]	<0.001
Age group (%)				
16-44	12838 (13.4)	9491 (13.5)	3347 (13.0)	<0.001
45-64	31743 (33.1)	23917 (34.0)	7826 (30.4)	
65-79	31198 (32.5)	22881 (32.6)	8317 (32.3)	
80-	20250 (21.1)	13980 (19.9)	6270 (24.3)	
Sex (%)				
Female	42330 (44.1)	30993 (44.1)	11337 (44.0)	0.796
Male	53699 (55.9)	39276 (55.9)	14423 (56.0)	
BMI, median [Q1,Q3]	27.5 [23.9,32.0]	27.8 [24.1,32.5]	26.8 [23.4,31.1]	<0.001
Ethnicity (%)				
Asian	2611 (2.7)	961 (1.4)	1650 (6.4)	<0.001
Black	9879 (10.3)	7811 (11.1)	2068 (8.0)	
Hispanic	3411 (3.6)	1285 (1.8)	2126 (8.3)	
Other	14381 (15.0)	8943 (12.7)	5438 (21.1)	
White	65747 (68.5)	51269 (73.0)	14478 (56.2)	
Insurance (%)				
Medicaid	7377 (7.7)	4723 (6.7)	2654 (10.3)	<0.001
Medicare	45946 (47.8)	32624 (46.4)	13322 (51.7)	
Other	42706 (44.5)	32922 (46.9)	9784 (38.0)	
SOFA, median [Q1,Q3]	4.0 [2.0,6.0]	4.0 [2.0,6.0]	4.0 [2.0,6.0]	0.001
CCI score, median [Q1,Q3]	5.0 [3.0,7.0]	5.0 [3.0,7.0]	5.0 [3.0,7.0]	<0.001
Ventilation (%)	33808 (35.2)	22515 (32.0)	11293 (43.8)	<0.001

Advance directives (%)	4666 (4.9)	2666 (3.8)	2000 (7.8)	<0.001
Pre ICU LOS day, median [Q1,Q3]	0.1 [0.0,0.6]	0.1 [0.0,0.6]	0.1 [0.0,0.7]	<0.001
LOS ICU day, median [Q1,Q3]	2.0 [1.1,3.9]	1.9 [1.1,3.6]	2.2 [1.2,4.6]	<0.001
LOS hospital day, median [Q1,Q3]	6.8 [4.0,11.8]	6.6 [3.9,11.3]	7.3 [4.3,13.1]	<0.001

In MIMIC, non-English primary speakers were older, had higher mortality rate, had higher percentages of ethnic minorities, holding Medicaid or Medicare, requiring ventilatory support, and needed longer ICU and hospital duration of stay when compared to English primary speakers (**eTable 12**).

Transparent Reporting of Multivariable Prediction Model for Individual Prognosis or Diagnosis: the TRIPOD Checklist

Section/Topic	ı		Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	3
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	3-4
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	5
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	6
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	6
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	6-8
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centers.	6-8
	5b	D;V	Describe eligibility criteria for participants.	6-8
	5c	D;V	Give details of treatments received, if relevant.	NA
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	8
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	NA
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	7-8
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	7-8
Sample size	8	D;V	Explain how the study size was arrived at.	7-8
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	7-8
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	7-8
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	7-8
	10c	V	For validation, describe how the predictions were calculated.	7-8
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	7-8

	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	7-8
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	7-8
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	7-8
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	9-9
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	9-9
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	9-9
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	9-11
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	9-11
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	9-11
	15b	D	Explain how to the use the prediction model.	9-11
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	9-11
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	9-11
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	15-16
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	12-16
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	12-16
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	12, 16
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	Supplement 1-22
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	2