OPEN

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

Xiaoli Liu, PhD[1–3]
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]

OBJECTIVES: Although illness severity scoring systems are widely used to support clinical decision-making and assess ICU performance, their potential bias across different age, sex, and primary language groups has not been well-studied.

DESIGN, SETTING, AND PATIENTS: We aimed to identify potential bias of Sequential Organ Failure Assessment (SOFA) and Acute Physiology and Chronic Health Evaluation (APACHE) IVa scores via large ICU databases.

SETTING/PATIENTS: This multicenter, retrospective study was conducted using data from the Medical Information Mart for Intensive Care (MIMIC) and eICU Collaborative Research Database. SOFA and APACHE IVa scores were obtained from ICU admission. Hospital mortality was the primary outcome. Discrimination (area under receiver operating characteristic [AUROC] curve) and calibration (standardized mortality ratio [SMR]) were assessed for all subgroups.

INTERVENTIONS: Not applicable.

MEASUREMENTS AND MAIN RESULTS: A total of 196,310 patient encounters were studied. Discrimination for both scores was worse in older patients compared with younger patients and female patients rather than male patients. In MIMIC, discrimination of SOFA in non-English primary language speakers patients was worse than that of English speakers (AUROC 0.726 vs. 0.783, $p < 0.0001$). Evaluating calibration via SMR showed statistically significant underestimations of mortality when compared with overall cohort in the oldest patients for both SOFA and APACHE IVa, female patients (1.09) for SOFA, and non-English primary language patients (1.38) for SOFA in MIMIC.

CONCLUSIONS: Differences in discrimination and calibration of two scores across varying age, sex, and primary language groups suggest illness severity scores are prone to bias in mortality predictions. Caution must be taken when using them for quality benchmarking and decision-making among diverse real-world populations.

KEYWORDS: bias evaluation; calibration; discrimination; hospital mortality; illness severity scores

Illness severity scoring systems (ISSSs) such as Sequential Organ Failure Assessment (SOFA) and Acute Physiology and Chronic Health Evaluation (APACHE) IVa are commonly used in critical care (CC) settings (1–3). These scores provide benchmarks to predict patient outcomes, triage patients, support clinical decision-making, assess ICU performance, and allocate scarce resources (4–7). Despite the need for fair, unbiased systems, these illness severity scores (ISSs) are limited by population-level prognostic estimation, leading to variable performance across subgroups such as ethnicity (6–11).

DOI: 10.1097/CCE.0000000000001033

KEY POINTS

Question: Is there potential bias of illness severity scores (ISSs) in patients divided by age, sex, and primary language?

Findings: Discrimination performance of Sequential Organ Failure Assessment (SOFA) and Acute Physiology and Chronic Health Evaluation (APACHE) IVa scores decreased significantly with increasing age. Mortality was underestimated for older patients and overestimated for younger patients. Both scores demonstrated slightly better discrimination for male patients. For non-English primary speakers, discrimination was decreased and mortality was significantly underestimated by SOFA score when compared with English primary speakers.

Meaning: Our study sheds light on significant disparities in ISSs and points out a serious need for a new generation of scoring systems that can provide accurate prognostication for our most vulnerable patients.

Additional sources of bias could be found in subgroups categorized by age, sex, and English language proficiency, but bias in these subgroups has not been fully explored (12–15).

Age plays a significant role in ICU care, and patients over 80 years old demonstrate the fastest population growth rate in the CC setting (12, 16). According to Daniele et al (17), Simplified Acute Physiology Score (SAPS) III was prone to underestimate the risk of death for patients older than 80 years old when compared with patients younger than 40 years old. A study conducted by Fernando et al (18) indicated that the accuracy of SAPS III score could be improved by accounting for performance status and comorbidities of older patients. Separately, female patients are often subject to inequalities such as delays in treatment (13). Recently, Todorov et al (13) showed that female patients have significantly higher observed mortality compared with male patients for every increase in SAPS II score. Additionally, a growing portion of today's patient population has limited English proficiency (LEP) in the United States (14, 19). Although some studies have evaluated the impact of LEP on the quality of healthcare delivery, there is limited literature in the CC setting (20–22). One recent study analyzed the effect of LEP on mortality in patients with sepsis and found an association with increased mortality after adjusting for illness severity (23).

We are only at the beginning stages of understanding the potential disparities that various social groups are currently facing while battling their critical illness. Current literature is mostly limited to small sample sizes from a few centers. Additionally, most studies did not focus on the evaluation of bias in ISSSs commonly used in CC settings. Therefore, in this large, multicenter retrospective study of ICU patients, we seek to evaluate the discrimination and calibration of two widely used ISSSs, SOFA, and APACHE IVa, in predicting in-hospital mortality in multiple patient subgroups divided by age (16–44, 45–64, 65–79, and over 80-yr-old), sex (female and male), and primary language (English and non-English) to assess for any potential bias.

MATERIALS AND METHODS

Data Sources

Two high-quality clinical databases were used. Medical Information Mart for Intensive Care (MIMIC) is an open, shared clinical database containing ICU admission data at Beth Israel Deaconess Medical Center (BIDMC) (24–26). Our analysis included both MIMIC-IV data (2008–2019) and MIMIC-III data (2001–2008). Altogether, MIMIC contains 83,478 patients with 113,873 ICU admissions across 19 years. In contrast, eICU Collaborative Research Database (eICU-CRD) is generated from the Philips telehealth system which covers 208 hospitals across the United States (10, 27). eICU-CRD contains 139,367 patients with 200,859 ICU admissions from 2014 to 2015 in the latest release of V.2. This study was exempt from institutional review board approval due to eICU-CRD's retrospective design, lack of direct patient intervention, and the security schema that certified reidentification risk to meet safe harbor standards by an independent privacy expert (Privacert, Cambridge, MA) based on the Health Insurance Portability and Accountability Act Certification 1031219-2. MIMIC has also been previously deidentified and deemed to be approved for the use of research by both institutional review boards of the Massachusetts Institute of Technology (number 04030000206) and BIDMC (2001-P-001699/14).

From the two databases, we excluded patients younger than 16 years old, ICU admission durations less than 4 hours, and patients without age, sex, or outcome information. Records of repeat ICU admissions during the same hospitalization were not included. We also excluded patient encounters without documented APACHE IVa scores in eICU-CRD and encounters where SOFA scores could not be calculated in both databases. Remaining data were then pooled in aggregate for analysis within each database. Age, sex, race, admission type, and discharge status were extracted from the data. English proficiency was determined by patient-reported preferred language found in the MIMIC database; eICU-CRD data does not contain information on language preference. To facilitate subgroup analysis, patient age was divided into four categories: 16–44, 45–64, 65–79, and over 80 years old. Patient's primary language spoken was divided into English and non-English proficiency in MIMIC.

Statistical Analysis

Descriptive statistics of patient characteristics were reported using median (25th, 75th) percentiles (interquartile range [IQR]) or proportions. Groups were compared using the Student *t* test or χ^2 test for categorical variables and the Wilcoxon rank-sum test or the Kruskal-Wallis test for continuous variables, as appropriate.

In-hospital mortality was selected as our outcome of interest. For SOFA score, we used logistic regression (LR) to characterize its relationship with mortality. LRs were fitted using the score's original continuous form from 0 to 24 points based on 20% of randomly selected patients in MIMIC and eICU-CRD, respectively (7–9). For APACHE IVa score, the mortality prediction of each eICU encounter had already been calculated in the database based on a published algorithm and therefore was directly imported (27).

We assessed discrimination of SOFA and APACHE IVa in mortality prediction for each subgroup of age, sex, and primary language via area under receiver operating characteristic (AUROC); associated 95% CIs were calculated. We measured calibration using standardized mortality ratio (SMR) and calibration belt plot for both scores (28). For SOFA, calibration was additionally assessed in two ways: 1) three increasing severity levels with score categories (≤ 7, 8–11, and > 11) and 2) four separate groups of increasing predicted mortality categories (0–5%, 5–10%, 10–20%, 20–50%, > 50%) within each subgroup (10).

To assess and quantify the impact of age, sex, and primary language on the ability of both SOFA and APACHE IVa to predict mortality, we fit LR models of both scores against observed mortality of databases. One thousand-fold Bootstrap resampling iteration was used to calculate 95% CI, and two-tailed *p* value of less than 0.05 was used as a threshold for statistical significance. Additional methodological details on statistical analysis are provided in the online data supplement, including an overall study flowchart (**eFig. 1**, http://links.lww.com/CCX/B292).

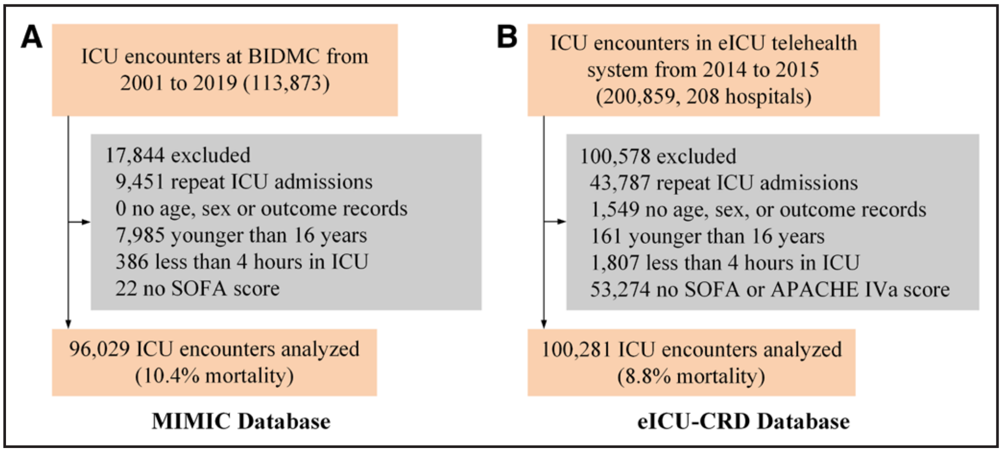
A
ICU encounters at BIDMC from 2001 to 2019 (113,873)
17,844 excluded
9,451 repeat ICU admissions
0 no age, sex or outcome records
7,985 younger than 16 years
386 less than 4 hours in ICU
22 no SOFA score
96,029 ICU encounters analyzed (10.4% mortality)
MIMIC Database
B
ICU encounters in eICU telehealth system from 2014 to 2015 (200,859, 208 hospitals)
100,578 excluded
43,787 repeat ICU admissions
1,549 no age, sex, or outcome records
161 younger than 16 years
1,807 less than 4 hours in ICU
53,274 no SOFA or APACHE IVa score
100,281 ICU encounters analyzed (8.8% mortality)
eICU-CRD Database

Figure 1. Data selection flowchart. **A**, Medical information mart for intensive care (MIMIC) data selection. **B**, eICU Collaborative Research Database (eICU-CRD) data selection. APACHE IV = Acute Physiology and Chronic Health Evaluation IV, BIDMS = Beth Israel Deaconess Medical Center, SOFA = Sequential Organ Failure Assessment.

RESULTS

Figure 1 describes the selection process used to arrive at our dataset. A total of 96,029 MIMIC ICU encounters were included with a 10.4% in-hospital mortality, and 100,281 eICU-CRD cases were included with an 8.8% in-hospital mortality. Baseline characteristics of the two study cohorts are shown in **eTable 1** (http://links.lww.com/CCX/B292). Nonsurvivors were significantly older than survivors (73 yr vs. 65 yr, $p < 0.001$ [MIMIC]; 71 yr vs. 65 yr, $p < 0.001$ [eICU-CRD]). Female patients represented 44.1% of the MIMIC cohort and 45.4% of the eICU-CRD cohort. In MIMIC, there was a higher percentage of non-English primary language speakers among nonsurvivors when compared with survivors (36.1% vs. 25.8%, $p < 0.001$), and non-English primary language speakers had a higher in-hospital mortality compared with English speakers (13.9% vs. 9.1%, $p < 0.001$). Nonsurvivors had higher SOFA scores during the first day of ICU admission (7 points vs. 4 points, $p < 0.001$ [MIMIC], 8 points vs. 5 points, $p < 0.001$ [eICU-CRD]) compared with survivors. In the eICU-CRD cohort, nonsurvivors had a higher APACHE IVa score on the first day of ICU admission compared with survivors (83 points vs. 51 points, $p < 0.001$).

In **Figure 2**, discrimination performance (i.e., AUROC) of both SOFA and APACHE IVa scores was presented for the two databases as well as different subgroups based on age, sex, and primary language. A higher AUROC indicates that SOFA or APACHE IVa score is better at discriminating between the two possible patient outcomes within the cohort studied. The overall performance was 0.761 (0.755–0.766) for SOFA in MIMIC, 0.73 (0.724–0.736) for SOFA in eICU-CRD, and 0.828 (0.823–0.833) for APACHE IVa in eICU-CRD. For different age categories, AUROC was noted to decrease with increasing age ($p < 0.0001$) as shown in **Figure 2** and **eTable 2** (http://links.lww.com/CCX/B292) (16–44, 45–64, 65–79, and 80 and older: 0.827, 0.792, 0.740, and 0.721 in MIMIC for SOFA; 0.812, 0.768, 0.712, and 0.678 in eICU-CRD for SOFA; 0.886, 0.844, 0.810, and 0.761 in eICU-CRD for APACHE IVa). The discrimination performance of male patients was better than that of female patients (0.764 vs. 0.759; 0.733 vs. 0.727; 0.832 vs. 0.823, $p < 0.0001$). For primary language, AUROC was higher for English speakers than for non-English speakers in the MIMIC cohort (0.783 vs. 0.726, $p < 0.0001$). Primary language subgroups were further

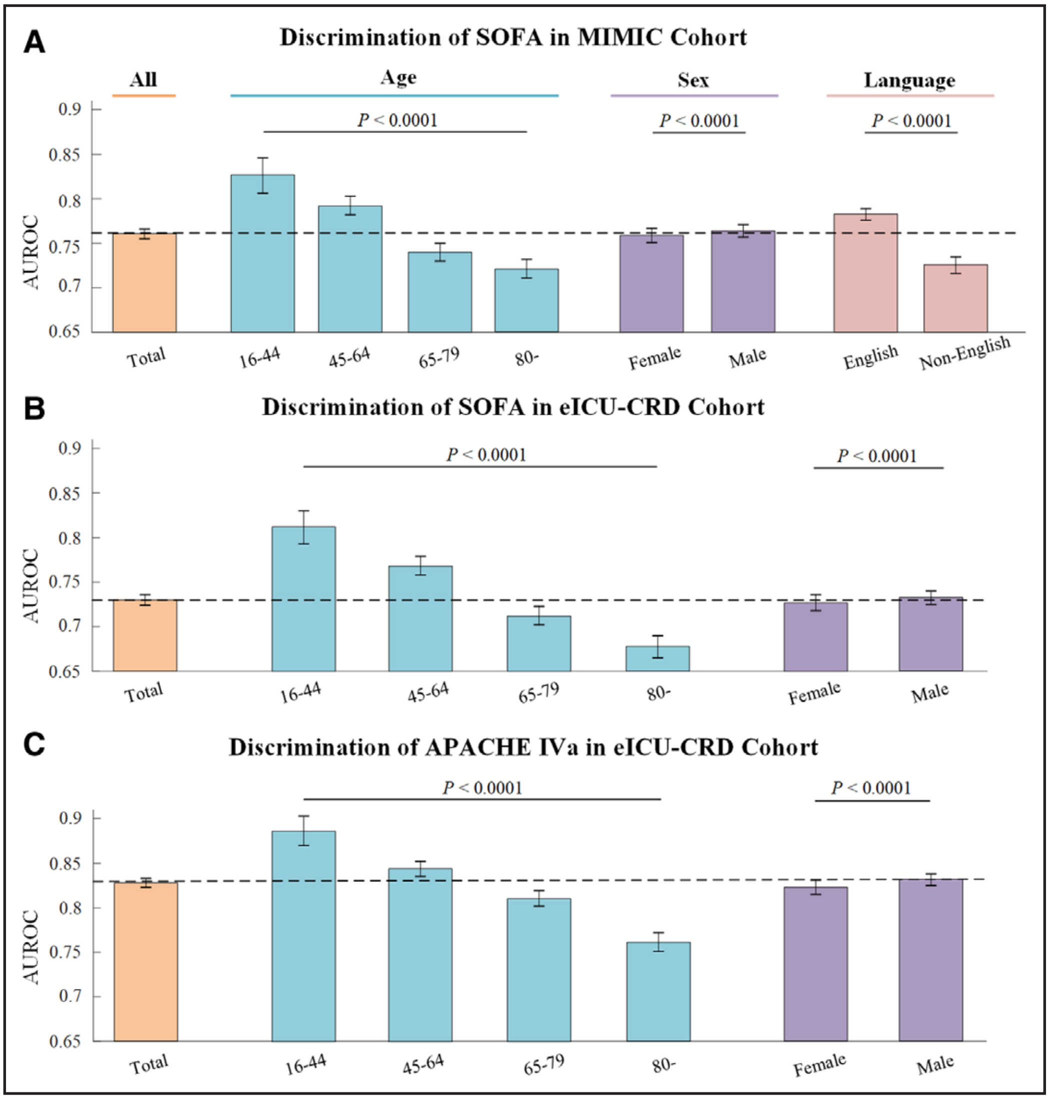

Figure 2. Discrimination performance of Sequential Organ Failure Assessment (SOFA) and Acute Physiology and Chronic Health Evaluation (APACHE) IVa on prediction of in-hospital mortality grouped by age, sex, and primary language spoken. **A**, SOFA score in medical information mart for intensive care (MIMIC) cohort. **B**, SOFA score in eICU Collaborative Research Database (eICU-CRD) cohort. **C**, Acute Physiology and Chronic Health Evaluation IVa (APACHE IVa) score in eICU-CRD cohort. AUROC = Area Under Receiver Operating Characteristic.

subdivided by age and sex, as shown in **eFigure 2** and **eTable 3** (http://links.lww.com/CCX/B292). In both language subgroups, increasing age was associated with decreased discrimination performance. AUROCs were not significantly different between female and male English speakers ($p = 0.12$) whereas female non-English speakers had lower AUROC than male non-English speakers (0.715 vs. 0.735, $p < 0.001$).

Figure 3 and **eTable 4** (http://links.lww.com/CCX/B292) detail the calibration of SOFA and APACHE IVa scores in mortality prediction across different subgroups as assessed by SMR. The fitted SOFA LR mortality function provided a good estimation of the entire MIMIC cohort's mortality risk, evidenced by an SMR of 0.98 (0.96–1.00). However, significant differences existed across different subgroups. In terms of age groups, SMR increased with age from 0.56 for 16–44 to 1.55 for 80 years and older, indicating overprediction of in-hospital mortality for younger patients and underprediction for older patients. Mortality was overestimated for male patients while underestimated for female patients (SMR: 0.90 vs. 1.09, $p < 0.001$). SMR for English speakers was 0.84, suggesting overprediction of mortality as compared with 1.38 for non-English speakers, suggesting underprediction ($p < 0.001$). SOFA score also demonstrated good calibration in the eICU-CRD cohort with an SMR of 1.02 (1.00–1.04). The SMR trend across various subgroups was similar to the MIMIC cohort. Mortality was overestimated in younger patients (SMR 0.56 in 16–44) but underestimated in older (1.57 in 80 yr and older) and female patients (1.07). In eICU-CRD, APACHE IVa overestimated the entire cohort's mortality with an SMR of 0.71. Within subgroups, mortality of younger patients continued to be overestimated (SMR 0.62 for 16–44) and older patients underestimated (0.74 for 65–79) using overall SMR as baseline. Mortality of female patients was overestimated whereas mortality of male patients was underestimated (SMR: 0.69 vs. 0.72, $p < 0.001$). Calibration was also assessed using calibration plots, which revealed similar findings as the SMR analysis described above (**eFigs. 3–5**, http://links.lww.com/CCX/B292).

Additionally, we evaluated calibration performance of SOFA score for each subgroup, further divided by predicted mortality categories of 0–5%, 5–10%, 10–20%, 20–50%, and 50–100%, as presented in **eFigures 6** and **7**, and **eTable 5** (http://links.lww.com/CCX/B292). A similar analysis was performed using three SOFA score categories (0–7, 8-11, over 11) of increasing disease severity, shown in **eTable 6** (http://links.lww.com/CCX/B292). Detailed results were presented in Supplemental Results (http://links.lww.com/CCX/B292).

Lastly, we assessed the significance and impact of age, sex, and primary language on mortality prediction by SOFA and APACHE IVa via LR models in both MIMIC and eICU-CRD. As presented in **Figure 4*A*, eFigure 8*A*** and ***C*** (http://links.lww.com/CCX/B292), mortality increased with increasing age for the same SOFA score ($p < 0.0001$). In **Figure 4*B*, eFigure 8*B*** and ***D*** (http://links.lww.com/CCX/B292), increased mortality was noted in female patients when compared with male patients with the same SOFA score in both databases ($p < 0.0001$), but no relationship was found when APACHE IVa score was substituted for SOFA ($p = 0.79$). Non-English primary speakers demonstrated higher mortality compared with English primary speakers for the same SOFA score ($p < 0.0001$), shown in **Figure 4*C***. Further details regarding each LR model were presented in **eTables 7–11** (http://links.lww.com/CCX/B292).

DISCUSSION

We conducted a retrospective study using two large ICU databases (MIMIC and eICU-CRD) to evaluate mortality prediction performance by two ISSSs (SOFA and APACHE IVa) stratified by age, sex, and primary language. To date, this is the first large multicenter cohort study that evaluated the disparity of ISSSs in these groups. AUROC was used to determine a score's ability to distinguish between two possible outcomes. The relative performance of these scores in distinguishing between alive or dead in each subgroup of age, sex, and primary language was assessed by comparing AUROCs. We found a consistent pattern in which both scores demonstrated superior discrimination performance in subcohort of younger patients when compared with older patients. Discrimination in subcohort of male patients was superior to female patients. Additionally, AUROC for subcohort of non-English primary speakers was decreased when compared with English speakers although analysis was limited to SOFA score with MIMIC data. We also found that the calibration of both scores was likely inadequate for most subgroups,

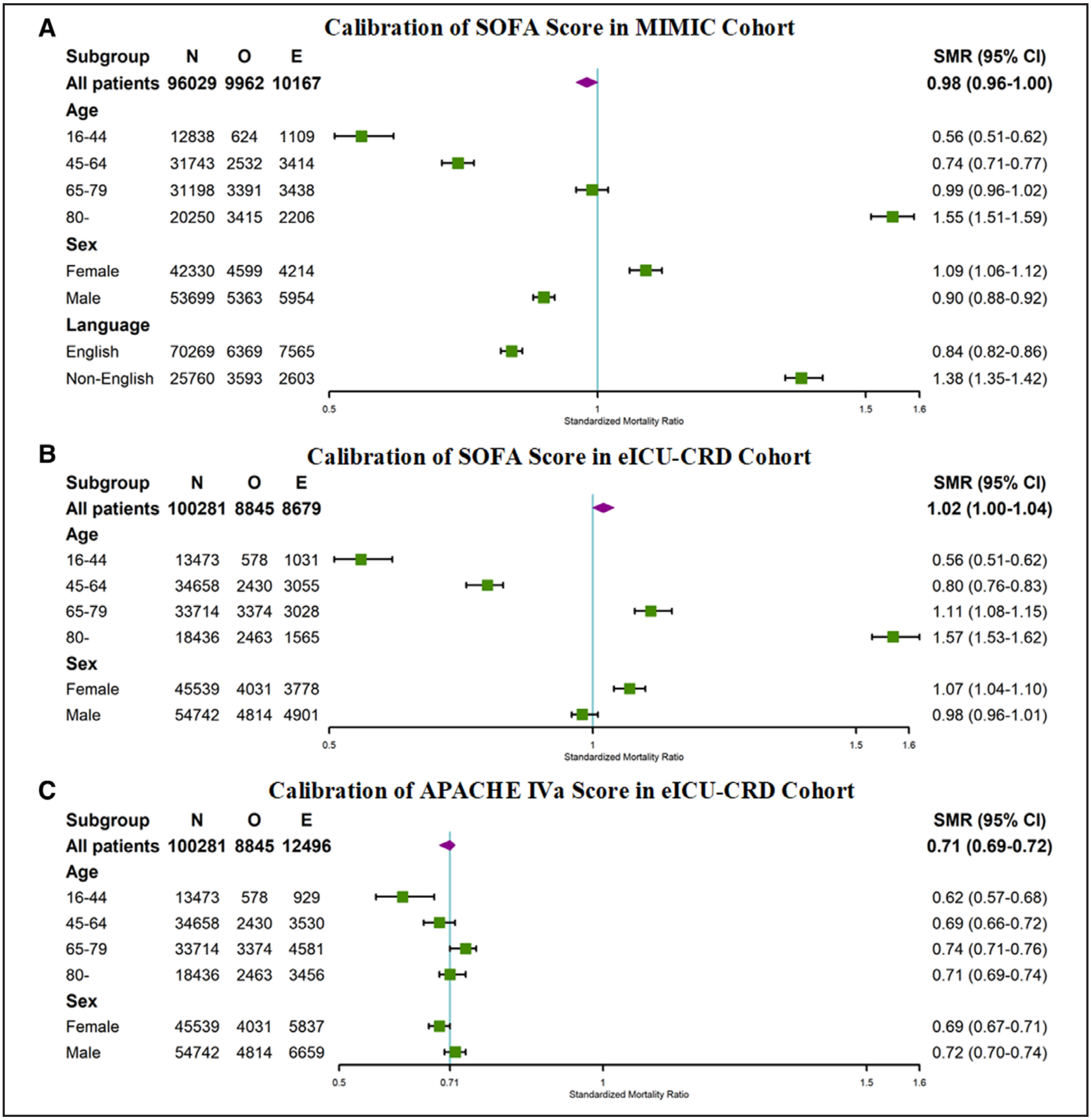

Figure 3. Calibration performance of Sequential Organ Failure Assessment (SOFA) and Acute Physiology and Chronic Health Evaluation (APACHE) IVa assessed via standardized mortality ratio (SMR). **A**, SOFA in medical information mart for intensive care (MIMIC) cohort. **B**, SOFA in eICU Collaborative Research Database (eICU-CRD) cohort. **C**, APACHE IVa score in eICU-CRD cohort. E = expected mortality, *N* = total number of patients, O = observed mortality.

or the quality of care delivered to various subgroups may be different. For example, overall mortality was overpredicted in younger patients and underpredicted in older patients. A higher than expected mortality was also found in non-English primary speakers compared with English speakers when evaluated by SOFA score.

With an increasingly aging society, ICUs are admitting more elderly patients, a population that brings a unique set of management challenges (12, 16). Popular ISSSs currently used in clinical practice may be less applicable to the elderly. SOFA score is based on expert consensus whereas APACHE IVa score was generated

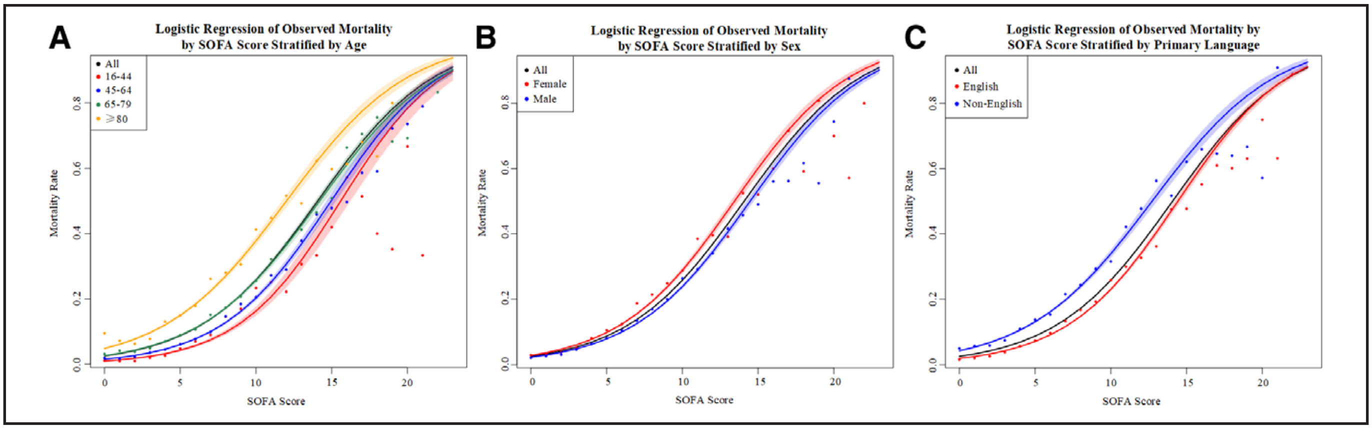

Figure 4. Logistic regression (LR) models of observed mortality by Sequential Organ Failure Assessment (SOFA) score in medical information mart for intensive care. **A**, Stratified by age. **B**, Stratified by sex. **C**, Stratified by primary language.

based on a younger population with a median age of 61.5 years (2). Our study revealed poor discrimination performance as well as significant calibration issues of ISSSs for elderly patients. This is especially true for SOFA, where SMR showed a significant underestimation of mortality for the oldest patients. This finding may be due to difficulty in predicting outcomes in the elderly due to many factors unique to this population, including differences in functional status, cognition, comorbidities, and frailty (12, 16, 29). Clinicians must be cautious when using ISSSs as sole basis for decision-making in elderly patients since they may in fact be sicker than their ISS represents, and consider incorporating additional factors mentioned above.

Disparities between sexes are receiving increasing attention. For example, a recent study showed female patients were overall less likely to receive ICU-level care (13). Our study showed that SOFA underestimated predicted mortality for female patients. The opposite was true when looking at APACHE IVa, so no consistent trend was observed. However, we did observe slightly decreased discrimination for female patients in both ISSSs. Regardless, evidence relating to mortality bias based on sex in ISSSs appears to be conflicting and can possibly be explained by geographic differences in the underlying dataset and the specific ISSS characteristics. This finding demonstrates the need for additional research into disparities based on large, multicenter datasets.

As the U.S. population increases in diversity, hospital systems will care for a larger proportion of patients with LEP. Studies have shown that language barriers adversely affect patient outcomes, leading to increased readmission rates (30). Our study in the MIMIC cohort also demonstrated worse outcomes for patients with non-English primary language; they have longer ICU and hospital lengths of stay as well as higher mortality (**eTable 12**, http://links.lww.com/CCX/B292). Discrimination performance of SOFA is also worse for non-English speakers when compared with English speakers. Additionally, underestimation of mortality by SOFA in non-English speakers suggests many variables specific to this population are currently overlooked. Therefore, using SOFA to predict mortality in LEP patients could further exacerbate disparities and cause harm to these minority groups.

ISSSs are standardized, validated, and user-friendly tools for assessment of disease severity and risk stratification (31). However, recent studies have revealed potential biases in these scores when used in certain ethnicities (10). As detailed above, our study further extends prior findings and sheds light on potential disparities of two scores when used across different age, sex, and primary language groups. Our findings have important clinical implications since these ISSSs are aimed to give clinicians a gestalt of how sick a patient is. Inaccuracies and biases exacerbated by these ISSSs for certain patient groups can lead to changes in clinical management, therapies offered, and most importantly triage decisions (13). One can easily imagine an 84-year-old Mandarin-speaking patient whose SOFA or APACHE IVa score may underestimate the severity of his or her mortality and illness severity, resulting in delays in care that may be associated with worse outcomes. Therefore, it is important for clinicians to be aware of these biases when using ISSSs for clinical decision-making. Overestimation of severity may lead to a disconnect between the patient's true severity of illness and physician recommendations to withdraw care, potentially interfering with a patient's best opportunity for treatment (10, 32). Finally, ISSSs are also used as benchmarks of ICU quality, which in turn support the evaluation of organizational management and development of care quality improvement plans (17, 33). However, given our finding of poor discrimination and calibration of these scores for many patient subgroups, caution must be exercised when they are used as a quality metric for specific populations. Our findings suggest these scores may need to be further refined and should only be used in conjunction with additional clinical data and quality metrics. Modification of these scores should be evaluated to see if applicable locally, and validation for different demographic subgroups would improve generalizability.

Our research has some limitations. First, we excluded patient records without either APACHE IVa (eICU-CRD only) or SOFA scores. The number of records with missing APACHE IVa scores was much higher than that of records without SOFA scores, indicating potential bias. However, MIMIC and eICU-CRD data have similar baseline characteristics suggesting that such exclusion criteria did not materially affect the composition of patient population and results. We also did not exclude patients with advance directives. These patients may not have received full aggressive care per patient and/or family preference, creating bias and impacting results. However, the overall proportion of patients with advanced directives is relatively small and should not materially impact our analysis. SOFA score was initially designed to assess severity of organ dysfunction not to predict mortality; therefore, our findings of biases are only applicable to the current use of SOFA in predicting mortality not its original intention. Additionally, composition of the original database used to formulate the APACHE IVa score likely differs from that of eICU-CRD data since the mortality of APACHE IVa database was higher at 13.6% compared with 8.8% in eICU-CRD. Although this discrepancy impacts calibration assessment when using the overall population, differences when comparing subgroups are likely preserved. Our study focused on in-hospital mortality and did not account for longer-term outcomes such as 30-day and 90-day mortality which are also important when evaluating ISSSs bias among various subgroups. Finally, the databases we used contained only U.S. hospitals, and thus applicability in Europe or other countries may be more limited.

CONCLUSIONS

Our large retrospective, multicenter study found decreased discrimination performance of SOFA and APACHE IVa scores when predicting mortality in elderly and female patients. In addition, SOFA showed impaired discrimination for non-English primary speakers as well. Significant underestimation of mortality risk in older and non-English primary speakers was found when calibration was analyzed. Our findings suggest that there are inherent inequities for various patient populations when using SOFA and APACHE IVa, two ISSSs commonly used for clinical management, patient triage, and ICU quality assessment. These scores can introduce biases that lead to inaccurate assessments and underestimation of illness severity in minority groups, which in turn can delay lifesaving interventions and result in inappropriate treatment plans. Therefore, clinicians and administrators must be cautious when using these scores and thoughtfully address specific characteristics underlying certain patient populations. It is also imperative to develop more accurate predictive systems for all patient subgroups and situations to reduce harm and improve the quality of medical care.

ACKNOWLEDGMENTS

We thank all research scientists in the Massachusetts Institute of Technology Laboratory for Computational Physiology for their efforts and contributions to the open-source databases.

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, MA.

4 Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA.

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, MA.*

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (http://journals.lww.com/ccejournal).

Dr. Zhang is funded by the National Natural Science Foundation of China (62171471) and the National Special Health Science Program (22BJZ42). Dr. Celi is funded by the National Institute of Health through the NIBIB R01 EB017205. The remaining authors have disclosed that they do not have any potential conflict of interest.

All authors contributed to data interpretation, writing, editing, and reviewing of the article. Drs. Liu, Shen, and Lie were in charge of data extraction, data analysis, and writing the initial draft article. Drs. Celi and Zhang contributed to the study design and study permissions. Drs. Zhang, Li, and Mark contributed to the interpretation of results and visualization of data. Drs. Liu, Shen, and Lie verified data and reviewed the literature. All authors contributed to statistical analysis and discussion of key findings. All authors had access to the raw study databases and had final responsibility for the decision to submit for publication.

Dr. Liu, Shen, and Lie are co-first authors.

Dr. Celi is senior author.

For information regarding this article, E-mail: zhangzhengbo@301hospital.com.cn

REFERENCES

1. Vincent JL, Moreno R, Takala J, et al: The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure on behalf of the working group on sepsis-related problems of the European Society of Intensive Care Medicine. *Intensive Care Med* 1996; 22:707–710
2. Zimmerman JE, Kramer AA, McNair DS, et al: Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Crit Care Med* 2006; 34:1297–1310
3. Raith EP, Udy AA, Bailey M, et al; Australian and New Zealand Intensive Care Society (ANZICS) Centre for Outcomes and Resource Evaluation (CORE): Prognostic Accuracy of the SOFA score, SIRS criteria, and qSOFA score for in-hospital mortality among adults with suspected infection admitted to the intensive care unit. *JAMA* 2017; 317:290–300
4. Pellathy TP, Pinsky MR, Hravnak M: Intensive care unit scoring systems. *Crit Care Nurse* 2021; 41:54–64
5. Desai N, Gross J: Scoring systems in the critically ill: Uses, cautions, and future directions. *BJA Educ*. 2019; 19:212–218
6. Miller WD, Han X, Peek ME, et al: Accuracy of the Sequential Organ Failure Assessment Score for in-hospital mortality by race and relevance to crisis standards of care. *JAMA Netw Open*. 2021; 4:e2113891
7. Ashana DC, Anesi GL, Liu VX, et al: Equitably allocating resources during crises: Racial differences in mortality prediction models. *Am J Respir Crit Care Med* 2021; 204:178–186
8. Minne L, Abu-Hanna A, de Jonge E: Evaluation of SOFA-based models for predicting mortality in the ICU: A systematic review. *Crit Care* 2008; 12:R161
9. Ferreira FL, Bota DP, Bross A, et al: Serial evaluation of the SOFA score to predict outcome in critically ill patients. *JAMA* 2001; 286:1754–1758
10. Sarkar R, Martin C, Mattie H, et al: Performance of intensive care unit severity scoring systems across different ethnicities in the USA: A retrospective observational study. *Lancet Digit Health* 2021; 3:e241–e249
11. Gershengorn HB, Holt GE, Rezk A, et al: Assessment of disparities associated with a crisis standards of care resource allocation algorithm for patients in 2 US Hospitals during the COVID-19 pandemic. *JAMA Netw Open*. 2021; 4:e214149
12. Guidet B, Vallet H, Boddaert J, et al: Caring for the critically ill patients over 80: A narrative review. *Ann Intensive Care*. 2018; 8:114
13. Todorov A, Kaufmann F, Arslani K, et al; Swiss Society of Intensive Care Medicine: Gender differences in the provision of intensive care: A Bayesian approach. *Intensive Care Med* 2021; 47:577–587
14. Divi C, Koss RG, Schmaltz SP, et al: Language proficiency and adverse events in US hospitals: A pilot study. *Int J Qual Health Care* 2007; 19:60–67
15. Bell SK, Dong J, Ngo L, et al: Diagnostic error experiences of patients and families with limited English-language health literacy or disadvantaged socioeconomic position in a cross-sectional US population-based survey. *BMJ Qual Saf* 2022; 32:644–654
16. Guidet B, de Lange DW, Boumendil A, et al; VIP2 study group: The contribution of frailty, cognition, activity of daily life and comorbidities on outcome in acutely admitted patients over 80 years in European ICUs: the VIP2 study. *Intensive Care Med* 2020; 46:57–69
17. Poole D, Rossi C, Anghileri A, et al: External validation of the Simplified Acute Physiology Score (SAPS) 3 in a cohort of 28,357 patients from 147 Italian intensive care units. *Intensive Care Med* 2009; 35:1916–1924
18. Zampieri FG, Colombari F: The impact of performance status and comorbidities on the short-term prognosis of very elderly patients admitted to the ICU. *BMC Anesthesiol*. 2014; 14:59
19. John-Baptiste A, Naglie G, Tomlinson G, et al: The effect of English language proficiency on length of stay and in-hospital mortality. *J Gen Intern Med* 2004; 19:221–228
20. Malevanchik L, Wheeler M, Gagliardi K, et al: Disparities after discharge: The Association of limited English proficiency and postdischarge patient-reported issues. *Jt Comm J Qual Patient Saf*. 2021; 47:775–782
21. Herbert BM, Johnson AE, Paasche-Orlow MK, et al: Disparities in reporting a history of cardiovascular disease among adults with limited English proficiency and Angina. *JAMA Netw Open*. 2021; 4:e2138780
22. Woods AP, Alonso A, Duraiswamy S, et al: Limited English proficiency and clinical outcomes after hospital-based care in English-speaking countries: A systematic review. *J Gen Intern Med* 2022; 37:2050–2061
23. Jacobs ZG, Prasad PA, Fang MC, et al: The association between limited English proficiency and sepsis mortality. *J Hosp Med* 2019; 14:E1–E7

24. Goldberger AL, Amaral LA, Glass L, et al: PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* 2000; 101:E215–E220

25. Johnson AE, Pollard TJ, Shen L, et al: MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3:160035

26. Johnson AEW, Bulgarelli L, Shen L, et al: MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data* 2023; 10:1

27. Pollard TJ, Johnson AEW, Raffa JD, et al: The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci Data* 2018; 5:180178

28. Finazzi S, Poole D, Luciani D, et al: Calibration belt for quality-of-care assessment based on dichotomous outcomes. *PLoS One* 2011; 6:e16110

29. Jain S, Murphy TE, O'Leary JR, et al: Association between socioeconomic disadvantage and decline in function, cognition, and mental health after critical illness among older adults: A cohort study. *Ann Intern Med* 2022; 175:644–655

30. Squires A, Ma C, Miner S, et al: Assessing the influence of patient language preference on 30 day hospital readmission risk from home health care: A retrospective analysis. *Int J Nurs Stud* 2022; 125:104093

31. Falcão ALE, Barros AGA, Bezerra AAM, et al: The prognostic accuracy evaluation of SAPS 3, SOFA and APACHE II scores for mortality prediction in the surgical ICU: An external validation study and decision-making analysis. *Ann Intensive Care*. 2019; 9:18

32. Keegan MT, Gajic O, Afessa B: Comparison of APACHE III, APACHE IV, SAPS 3, and MPM0III and influence of resuscitation status on model performance. *Chest* 2012; 142:851–858

33. Lam, Kam W, Kang Yiu L: Evaluation of outcome and performance of an intensive care unit in Hong Kong by APACHE IV model: 2007–2014. *J Emerg Crit Care Med* 2017; 1:10–21037.