Predictive Value of Sequential Organ Failure Assessment Score across Patients with and without COVID-19 Infection

Hayley B. Gershengorn[1,2], Samira Patel[3], Bhavarth Shukla[4], Prem R. Warde[3], Shane M. Soorus[3], Gregory E. Holt[1], Daniel H. Kett[1], Dipen J. Parekh[5], and Tanira Ferreira[1]

[1]Division of Pulmonary, Critical Care, and Sleep Medicine, [4]Division of Infectious Diseases, Department of Medicine, and [5]Department of Urology, Miller School of Medicine, University of Miami, Miami, Florida; [2]Division of Critical Care Medicine, Albert Einstein College of Medicine, Bronx, New York; and [3]Care Transformation, University of Miami Hospital and Clinics, Miami, Florida

ORCID ID: 0000-0002-7360-2489 (H.B.G.).

Abstract

Rationale: Sequential organ failure assessment (SOFA) scores are commonly used in crisis standards of care policies to assist in resource allocation. The relative predictive value of SOFA by coronavirus disease (COVID-19) infection status and among racial and ethnic subgroups within patients infected with COVID-19 is unknown.

Objectives: To evaluate the accuracy and calibration of SOFA in predicting hospital mortality by COVID-19 infection status and across racial and ethnic subgroups.

Methods: We performed a retrospective cohort study of adult admissions to the University of Miami Hospital and Clinics inpatient wards (July 1, 2020–April 1, 2021). We primarily considered maximum SOFA within 48 hours of hospitalization. We assessed accuracy using the area under the receiver operating characteristic curve (AUROC) and created calibration belts. Considered subgroups were defined by COVID-19 infection status (by severe acute respiratory syndrome coronavirus 2 polymerase chain reaction testing) and prevalent racial and ethnic minorities. Comparisons across subgroups were made with DeLong testing for discriminative accuracy and visualization of calibration belts.

Results: Our primary cohort consisted of 20,045 hospitalizations, of which 1,894 (9.5%) were COVID-19 positive. SOFA was similarly accurate for COVID-19–positive (AUROC, 0.835) and COVID-19–negative (AUROC, 0.810; $P = 0.15$) admissions but was slightly better calibrated in patients who were positive for COVID-19. For those with critical illness, maximum SOFA score accuracy at critical illness onset also did not differ by COVID-19 status (AUROC, COVID-19 positive vs. negative: intensive care unit admissions, 0.751 vs. 0.775; $P = 0.46$; mechanically ventilated, 0.713 vs. 0.792, $P = 0.13$), and calibration was again better for patients positive for COVID-19. Among patients with COVID-19, SOFA accuracy was similar between the non-Hispanic White population (AUROC, 0.894) and racial and ethnic minorities (Hispanic White population: AUROC, 0.824 [P vs. non-Hispanic White = 0.05]; non-Hispanic Black population: AUROC, 0.800 [$P = 0.12$]; Hispanic Black population: AUROC, 0.948 [$P = 0.31$]). This similar accuracy was also found for those without COVID-19 (non-Hispanic White population: AUROC, 0.829; Hispanic White population: AUROC, 0.811 [$P = 0.37$]; Hispanic Black population: AUROC, 0.828 [$P = 0.97$]; non-Hispanic Black population: AUROC, 0.867 [$P = 0.46$]). SOFA was well calibrated for all racial and ethnic groups with COVID-19 but estimated mortality more variably and performed less well across races and ethnicities without COVID-19.

Conclusions: SOFA accuracy does not differ by COVID-19 status and is similar among racial and ethnic groups both with and without COVID-19. Calibration is better for COVID-19–infected patients and, among those without COVID-19, varies by race and ethnicity.

Keywords: organ dysfunction scores; calibration; COVID-19; race factors; ethnic groups

(Received in original form June 7, 2021; accepted in final form November 15, 2021)

Supported by the University of Miami Hospital and Clinics through the UHealth-DART Research Group (H.B.G., S.P., B.S., P.R.W., and T.F., all of whom are members).

Ann Am Thorac Soc Vol 19, No 5, pp 790–798, May 2022

DOI: 10.1513/AnnalsATS.202106-680OC
Internet address: www.atsjournals.org

The importance of accurate predictions of short-term mortality in the setting of acute illness has become clear during the coronavirus disease (COVID-19) pandemic. Crisis standards of care (CSC) policies create frameworks to allocate life-saving resources when demand exceeds supply. Most such policies are based, at least in part, on expected short-term patient survival with the sequential organ failure assessment (SOFA) score (1) being commonly used for mortality predictions (2–4).

The accuracy of SOFA for predicting mortality in the setting of COVID-19 has been found to vary. In Wuhan, China, early in the pandemic, SOFA was shown to have a poor predictive accuracy (area under the receiver operating characteristic curve [AUROC], 0.69) for hospitalized patients with COVID-19 (5), but to be excellent for critically ill patients with COVID-19 (AUROC, 0.89) (6, 7). In the United States, SOFA scores had excellent predictive accuracy for hospitalized patients with COVID-19 (AUROC, 0.85) and performed even better in a larger cohort inclusive of patients without COVID-19 (AUROC, 0.90) (8). Whether the accuracy of SOFA differs for patients based on COVID-19 status is unknown.

In diverse U.S. populations of hospitalized patients without COVID-19, SOFA accuracy has been demonstrated to vary little across racial and ethnic groups (9, 10). However, in this same study, SOFA was shown to be miscalibrated, overestimating mortality among individuals of Black race and underestimating it for White people. Whether this differential miscalibration exists among individuals with COVID-19 is uncertain.

Recognizing that use of any CSC policy during the COVID-19 pandemic will necessarily affect patients of all racial and ethnic groups both with and without COVID-19, we sought to directly compare the predictive accuracy and calibration of SOFA across these groups. We hypothesized that SOFA would perform similarly for patients with and without COVID-19 but would perform less well for persons of non-White race and/or Hispanic ethnicity independent of COVID-19 status.

Methods

We performed a retrospective cohort study of admissions to the University of Miami Hospital and Clinics inpatient wards from July 1, 2020, to April 1, 2021. This hospital system consists of three inpatient facilities: a general tertiary care academic hospital (550 beds), a cancer-specialty hospital (40 beds), and an ophthalmology-care hospital (75 beds).

Cohort

Our primary cohort consisted of all admissions who were discharged by April 1, 2021 (as patients discharge after then would be missing hospital mortality data). Admissions were excluded if they were less than 18 years old or had missing SOFA data. We considered two secondary cohorts: *1*) patients admitted to an ICU; and *2*) patients who received invasive mechanical ventilation (MV).

Exposure

Our exposure was maximum SOFA score. Starting on June 15, 2020, we implemented an automated SOFA score calculation into our electronic health record (Epic). Total SOFA scores (range 0 [best]–24 [worst]) as well as each organ system-based component (respiratory, cardiovascular, renal, liver, hematological, and neurological; range 0 [best]–4 [worst] each) were calculated and recorded hourly for all admissions throughout their hospital stay (Table E1 in the online supplement). Owing to incomplete documentation of urine output, the renal component of the SOFA score was based solely on creatinine; all patients with an active order for dialysis were given a renal SOFA score of 4 (worst value). When arterial blood gas testing was not available, conversion of oxygen saturation by pulse oximetry to an estimated partial pressure of oxygen in the arterial blood was used (11).

Our primary exposure was maximum SOFA score within 48 hours following hospitalization for the full cohort. We evaluated two additional exposures for sensitivity analyses, within 24 hours and within 72 hours of hospitalization. For the subcohort of ICU patients, we considered maximum SOFA within 48 hours surrounding ICU admission (defined as within 24 hours before and 24 hours after ICU admission), recognizing that early ICU care may change illness trajectory such that need for invasive mechanical ventilation, a main resource for potential allocation, may be altered. As a *post hoc* sensitivity analysis, we also considered maximum SOFA in solely the 24 hours before ICU admission. For the subcohort of MV patients, we considered maximum SOFA within 48 hours prior to MV initiation (12).

Statistical Analyses

We described our cohort using standard summary statistics. Comparisons across SOFA groupings commonly used in CSC policies (SOFA <6, 6–8, 9–11, or ≥ 12 [3, 4, 13]) were performed using *t* and chi-square testing. Model discrimination was assessed using the AUROC for SOFA predictions of hospital mortality (inclusive of patients who died during hospitalization or were discharged to hospice). Model calibration was assessed through evaluation of hospital mortality rates across SOFA groupings as well as calibration belts, which visually display type, range, and magnitude of miscalibration (10, 14). As SOFA score is used without adjustment for other covariables as a predictor of short-term mortality in CSC policies, no adjustment for other covariables was included in either analysis.

SOFA discrimination and calibration were first calculated for the full cohort. Each was then calculated for patients with and without COVID-19; by hospital protocol, all patients received a severe acute respiratory syndrome coronavirus 2

Author Contributions: H.B.G. and S.P. had full access to all data. H.B.G. and S.P. performed all statistical analyses. H.B.G. drafted the primary manuscript. All authors were involved in primary study design, results interpretation, and critical manuscript review. H.B.G. is responsible for the integrity of the work.

Correspondence and requests for reprints should be addressed to Hayley B. Gershengorn, M.D., Professor, University of Miami, Miller School of Medicine, Division of Pulmonary, Critical Care, and Sleep Medicine, 1951 NW 7th Avenue, Miami, FL 33136. E-mail: hbg20@med.miami.edu.

This article has an online supplement, which is accessible from this issue's table of contents at www.atsjournals.org.

polymerase chain reaction test on hospital admission, and patients with a positive test at any time during hospitalization were considered COVID-19 positive. Finally, for both patients with and, separately, without COVID-19, discrimination and calibration were calculated for racial and ethnic groups (non-Hispanic White, Hispanic White, non-Hispanic Black, Hispanic Black, unknown [inclusive of unknown and refused], and other [inclusive of Asian, American Indian, Pacific Islander, and multiple races]); all race and ethnicity designations were provided by the patient or their family member and captured in the electronic health record. Categorization of discrimination accuracy used a previously defined framework (AUROC for poor, <0.7; acceptable, 0.7–0.8; excellent, 0.8–0.9; or outstanding, >0.9) (7, 10), and comparisons of discrimination were made by DeLong testing. Likelihood ratio testing was used to assess statistical differences from perfect calibration (10, 14). Comparison of mortality rates across SOFA groupings was made by chi-square testing.

Post hoc, we decided to evaluate the accuracy among COVID-19 admissions of the individual components of the SOFA score: respiratory, cardiovascular, liver, kidney, and coagulation; we did not include neurological as it is less consistently captured and, thus, often assumed to be normal. For those components with at least acceptable accuracy, we investigated whether accuracy or calibration differed by race and ethnicity.

We repeated the above total SOFA score assessments of discrimination and calibration for each secondary cohort (ICU and MV patients). We then conducted sensitivity analyses using the alternate timeframes for maximum SOFA (24 h and 72 h) and, separately, an alternate definition of hospital mortality (reclassifying patients discharged to hospice as survivors). Finally, we evaluated the differential accuracy of SOFA score near the time of critical illness onset for the ICU and MV cohorts across racial and ethnic groups.

This study was approved by the Institutional Review Board of the University of Miami (#20200739). *P* values were considered significant if less than 0.05; correction for multiple comparisons was not used, and, therefore, all secondary analyses should be considered hypothesis generating. Statistical analyses were performed using R version 3.6.2.

Results

Our primary cohort consisted of 20,045 hospitalizations, of which 1,894 (9.5%) were COVID-19 positive (Figure E1 and Table 1). Nearly half the cohort (9,849 admissions [49.1%]) were Hispanic White; the remainder were non-Hispanic White (4,358 [21.7%]), non-Hispanic Black (3,791 [18.9%]), Hispanic Black (513 [2.6%]), other race and ethnicity (796 [4.0%]), or of unknown race (738 [3.7%]). Most patients had maximum SOFA scores within 48 hours of hospital admissions of less than 6 (18,446 [92.0%]), with 1,052 (5.3%) having SOFA 6–8, 337 (1.7%) SOFA 9–11, and 210 (1.1%) SOFA 12 or more. There were substantial differences in the maximum SOFA score across COVID-19 positivity ($P < 0.001$) and racial and ethnic groups ($P < 0.001$). Admissions with lower maximum SOFA scores were more likely to be COVID-19–positive (9.2% of all patients with a SOFA lower than 6, 13.9% of those with a SOFA of 6–8, and 9.5% of those with a SOFA of 9–11 vs. 6.7% of those with a SOFA of 12 or more were COVID-19–positive) and of non-Hispanic Black race and ethnicity (19.0% of all patients with a SOFA lower than 6 and 21.5% of those with a SOFA of 6–8 vs. 8.9% of those with a SOFA of 9–11 and 14.3% of those with a SOFA of 12 or more were of non-Hispanic Black race and ethnicity).

SOFA and COVID-19 Positivity

Maximum SOFA score within 48 hours of hospital admission had excellent accuracy in predicting hospital mortality for the full cohort (AUROC, 0.820) (Figure 1A). SOFA was similarly accurate for COVID-19–positive admissions (AUROC, 0.835) (Figure 1B) and those without COVID-19 infection (AUROC, 0.810; $P = 0.15$) (Figure 1C).

Increasing SOFA score was associated with increasing mortality rates for the full cohort from 2.9% for SOFA 6 or lower to 33.8% for SOFA 12 or more (Table 2); similar trends were seen in patients with and without COVID-19. SOFA was poorly calibrated for the full cohort, substantially underestimating mortality for those at low risk and overestimating mortality for those at more moderate risk of death (Figure 1D). Calibration was good for patients positive for COVID-19 (Figure 1E) but for patients without COVID-19 infection resulted in similar under and overestimations of mortality as for the full cohort (not unexpected, as patients negative for COVID-19 comprised 90.5% of the full cohort) (Figure 1F).

SOFA, COVID-19 Positivity, and Race and Ethnicity

SOFA accuracy among admissions with COVID-19 varied by race and ethnicity, ranging from excellent for non-Hispanic Black people (AUROC, 0.800) to outstanding for Hispanic Black people (AUROC, 0.948) (Figures 2A–2D). The accuracy of SOFA did not differ statistically between non-Hispanic White admissions (AUROC, 0.894) and those of any other racial and ethnic group: Hispanic White (AUROC, 0.824; $P = 0.05$); non-Hispanic Black ($P = 0.12$); or Hispanic Black ($P = 0.31$). Mortality rates increased as SOFA score increased for each race and ethnic group, and good calibration (albeit with lower confidence for non-Hispanic and Hispanic Black patients) was observed for all races and ethnicities.

SOFA accuracy among admissions without COVID-19 also varied by race and ethnicity but was excellent across groups (AUROC for non-Hispanic White patients = 0.829; Hispanic White patients = 0.811 [*P* vs. non-Hispanic White = 0.37]; non-Hispanic Black patients = 0.828 [$P = 0.97$]; and Hispanic Black patients = 0.867 [$P = 0.46$]) (Figures 2E–2H). Mortality rates for patients negative for COVID-19 were less consistently correlated with SOFA score; for non-Hispanic patients with higher illness severity (SOFA ⩾ 9), increased SOFA score was not coincident with increased mortality, although sample size was small. Calibration was poor for most racial and ethnic groups with underestimation of mortality for patients without COVID-19 at lower risk and overestimation for those at more moderate risk of death. Calibration was better for Hispanic than for non-Hispanic patients.

SOFA Components and Race and Ethnicity among COVID-19 Admissions

Among admissions with COVID-19, the accuracy of the respiratory component of the SOFA score was excellent (AUROC, 0.843), and that for the cardiovascular component was acceptable (AUROC, 0.720); the

Table 1. Primary cohort characteristics

Characteristic	Full Cohort, *n* (%)	SOFA <6, *n* (%)	SOFA 6–8, *n* (%)	SOFA 9–11, *n* (%)	SOFA ⩾12, *n* (%)
No. of patients (row %)	20,045 (100.0)	18,446 (92.0)	1,052 (5.3)	337 (1.7)	210 (1.1)
COVID-19 positive	1,894 (9.5)	1,702 (9.2)	146 (13.9)	32 (9.5)	14 (6.7)
Sex, male	10,179 (50.8)	9,289 (50.4)	565 (53.7)	201 (59.6)	124 (59.0)
Age, median (IQR), yr	61 (49–72)	61 (48–72)	68 (58–79)	68 (58–77)	69 (60–79)
Race and ethnicity					
Non-Hispanic White	4,358 (21.7)	4,050 (22.0)	165 (15.7)	91 (27.0)	52 (24.8)
Hispanic White	9,849 (49.1)	9,021 (48.9)	554 (52.7)	166 (49.3)	108 (51.4)
Non-Hispanic Black	3,791 (18.9)	3,505 (19.0)	226 (21.5)	30 (8.9)	30 (14.3)
Hispanic Black	513 (2.6)	473 (2.6)	30 (2.9)	7 (2.1)	3 (1.4)
Other*	796 (4.0)	729 (4.0)	42 (4.0)	17 (5.0)	8 (3.8)
Unknown	738 (3.7)	668 (3.6)	35 (3.3)	26 (7.7)	9 (4.3)
Primary insurer type					
Commercial Insurance	7,572 (37.8)	7,186 (39.0)	218 (20.7)	111 (32.9)	57 (27.1)
Medicaid	2,496 (12.5)	2,339 (12.7)	123 (11.7)	18 (5.3)	16 (7.6)
Medicare	8,588 (42.8)	7,601 (41.2)	657 (62.5)	196 (58.2)	134 (63.8)
Other	428 (2.1)	408 (2.2)	14 (1.3)	4 (1.2)	2 (1.0)
Not Recorded	961 (4.8)	912 (4.9)	40 (3.8)	8 (2.4)	1 (0.5)
Elixhauser comorbidities					
Congestive heart failure	3,363 (16.8)	2,695 (14.6)	454 (43.2)	122 (36.2)	92 (43.8)
Valvular disease	3,631 (18.1)	2,962 (16.1)	376 (35.7)	165 (49.0)	128 (61.0)
Pulmonary circulation dis.	1,045 (5.2)	931 (5.0)	92 (8.7)	14 (4.2)	8 (3.8)
Peripheral vascular dis.	3,516 (17.5)	3,064 (16.6)	305 (29.0)	82 (24.3)	65 (31.0)
Hypertension	13,396 (66.8)	12,034 (65.2)	914 (86.9)	277 (82.2)	171 (81.4)
Paralysis	1,136 (5.7)	1,013 (5.5)	86 (8.2)	21 (6.2)	16 (7.6)
Other neurologic dis.	4,593 (22.9)	3,981 (21.6)	446 (42.4)	103 (30.6)	63 (30.0)
Chronic pulmonary dis.	4,815 (24.0)	4,296 (23.3)	377 (35.8)	84 (24.9)	58 (27.6)
Diabetes mellitus	14,395 (71.8)	13,394 (0.0)	611 (0.0)	236 (0.0)	154 (0.0)
Uncomplicated	5,650 (28.2)	5,052 (27.4)	441 (41.9)	101 (30.0)	56 (26.7)
Complicated	4,916 (24.5)	4,155 (22.5)	553 (52.6)	131 (38.9)	77 (36.7)
Hypothyroidism	3,108 (15.5)	2,747 (14.9)	240 (22.8)	70 (20.8)	51 (24.3)
Renal failure	3,974 (19.8)	3,174 (17.2)	581 (55.2)	122 (36.2)	97 (46.2)
Liver disease	3,350 (16.7)	2,943 (16.0)	279 (26.5)	73 (21.7)	55 (26.2)
Peptic ulcer disease	1,061 (5.3)	939 (5.1)	80 (7.6)	29 (8.6)	13 (6.2)
AIDS	348 (1.7)	313 (1.7)	32 (3.0)	2 (0.6)	1 (0.5)
Lymphoma	1,195 (6.0)	1,088 (5.9)	75 (7.1)	16 (4.7)	16 (7.6)
Metastatic cancer	3,684 (18.4)	3,431 (18.6)	189 (18.0)	40 (11.9)	24 (11.4)
Solid tumor without metastasis	6,116 (30.5)	5,737 (31.1)	297 (28.2)	56 (16.6)	26 (12.4)
Rheumatoid arthritis/CVD	1,176 (5.9)	1,070 (5.8)	76 (7.2)	18 (5.3)	12 (5.7)
Coagulopathy	3,568 (17.8)	2,673 (14.5)	492 (46.8)	235 (69.7)	168 (80.0)
Obesity	6,220 (31.0)	5,605 (30.4)	425 (40.4)	114 (33.8)	76 (36.2)
Weight loss	3,736 (18.6)	3,328 (18.0)	300 (28.5)	70 (20.8)	38 (18.1)
Fluid and electrolyte dis.	9,045 (45.1)	7,725 (41.9)	916 (87.1)	247 (73.3)	157 (74.8)
Blood loss anemia	1,315 (6.6)	1,141 (6.2)	121 (11.5)	33 (9.8)	20 (9.5)
Deficiency anemia	8,416 (42.0)	7,336 (39.8)	794 (75.5)	175 (51.9)	111 (52.9)
Alcohol abuse	838 (4.2)	737 (4.0)	71 (6.7)	19 (5.6)	11 (5.2)
Drug abuse	1,000 (5.0)	936 (5.1)	53 (5.0)	8 (2.4)	3 (1.4)
Psychoses	1,225 (6.1)	1,112 (6.0)	91 (8.7)	8 (2.4)	14 (6.7)
Depression	4,381 (21.9)	3,978 (21.6)	304 (28.9)	61 (18.1)	38 (18.1)
Ever admitted to the ICU	2,861 (14.3)	1,909 (10.3)	465 (44.2)	288 (85.5)	199 (94.8)
Ever mechanically ventilated	1,044 (5.2)	391 (2.1)	259 (24.6)	225 (66.8)	169 (80.5)
Hospital mortality	942 (4.7)	532 (2.9)	240 (22.8)	99 (29.4)	71 (33.8)
Disposition for survivors					
Facility	1,722 (9.0)	1,476 (8.2)	178 (21.9)	35 (14.7)	33 (23.7)
Home	17,381 (91.0)	16,438 (91.8)	634 (78.1)	203 (85.3)	106 (76.3)

Definition of abbreviations: AIDS = acquired immunodeficiency syndrome; COVID-19 = coronavirus disease; CVD = collagen vascular disease; dis. = disorder; ICU = intensive care unit; IQR = interquartile range; SOFA = sequential organ failure assessment.
*Inclusive of American Indian, Asian, multiple races, Pacific islander and other.

accuracies for the liver (AUROC, 0.559), kidney (AUROC, 0.657), and coagulation (AUROC, 0.487) components were poor (Figure E2). The respiratory component was well calibrated, but the cardiovascular component both under- and overestimated mortality for admissions with COVID-19.

Accuracy of the respiratory component varied by race and ethnicity, ranging from acceptable for non-Hispanic Black patients (AUROC, 0.796) to outstanding for Hispanic

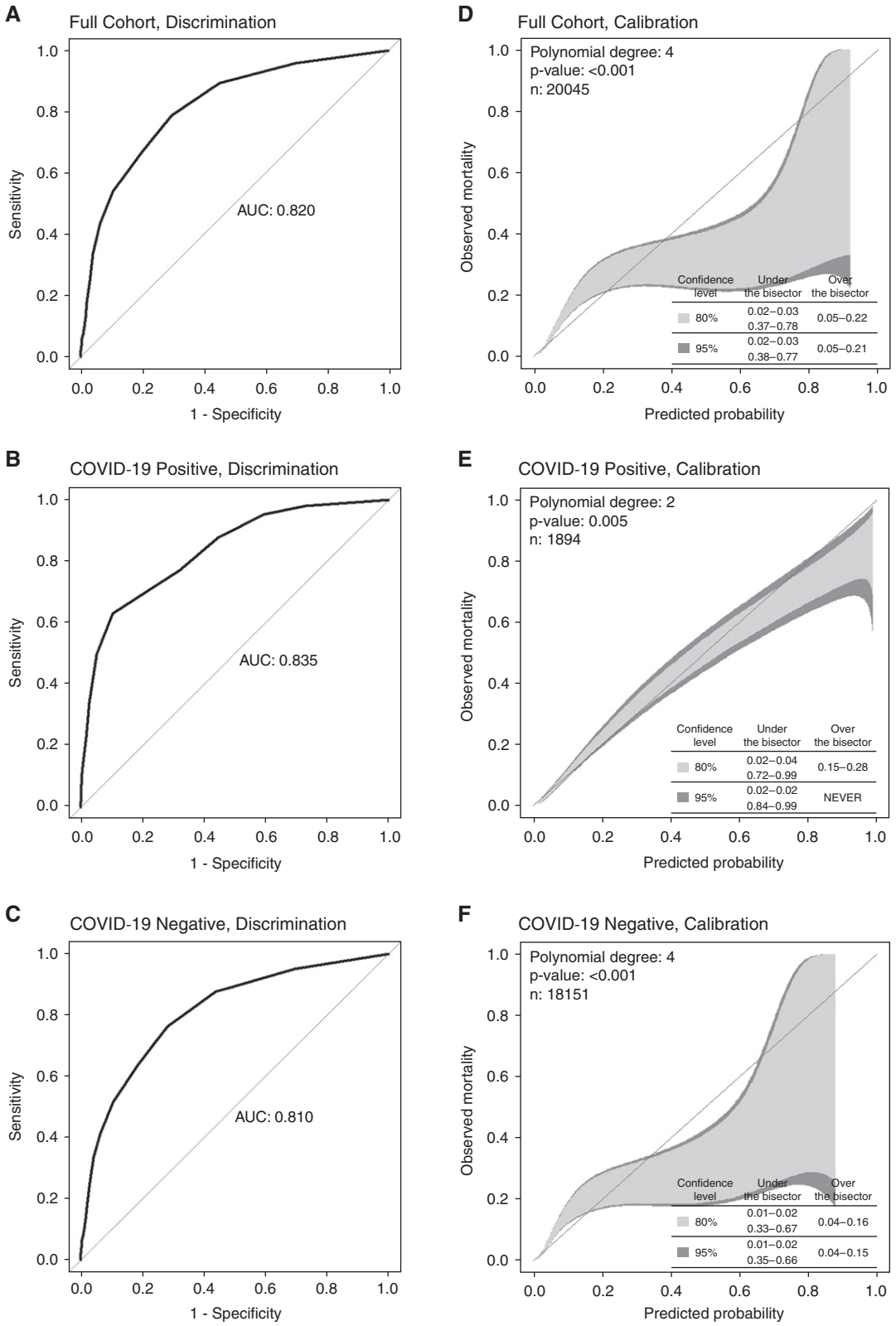

Figure 1. (*A–F*) Discrimination (*A–C*) and calibration (*D–F*) of maximum sequential organ failure assessment score for all admissions and by coronavirus disease (COVID-19) status. $P = 0.15$ for Delong testing of discrimination for COVID-19 positive versus COVID-19 negative. AUC = area under the curve.

Table 2. Mortality rates stratified by maximum SOFA score

Cohort	Full Cohort	SOFA <6	SOFA 6–8	SOFA 9–11	SOFA ⩾12	*P* Value*
All patients	4.7%	2.9%	22.8%	29.4%	33.8%	<0.001
COVID-19–positive patients	11.3%	6.3%	50.7%	65.6%	78.6%	<0.001
Race and ethnicity						
Non-Hispanic White	15.1%	6.6%	76.2%	50.0%	100.0%	<0.001
Hispanic White	11.9%	7.4%	46.9%	70.6%	83.3%	<0.001
Non-Hispanic Black	6.2%	4.0%	29.6%	33.3%	0.0%	<0.001
Hispanic Black	6.8%	1.8%	75.0%	n/a	n/a	<0.001
COVID-19–negative patients	4.0%	2.5%	18.3%	25.6%	30.6%	<0.001
Race and ethnicity						
Non-Hispanic White	3.2%	2.1%	18.1%	25.3%	14.3%	<0.001
Hispanic White	4.4%	2.7%	20.1%	25.5%	36.3%	<0.001
Non-Hispanic Black	3.3%	2.0%	13.1%	40.7%	34.5%	<0.001
Hispanic Black	4.2%	1.9%	23.1%	42.9%	66.7%	<0.001

Definition of abbreviations: COVID-19 = coronavirus disease; SOFA = sequential organ failure assessment.
*Chi-square testing to determine overall effect of SOFA on mortality for each grouping.

Black patients (AUROC, 0.939). Compared with non-Hispanic White patients (AUROC, 0.902), accuracy was worse for Hispanic White patients (AUROC 0.840; $P = 0.019$) and non-Hispanic Black patients ($P = 0.038$), but not Hispanic Black patients ($P = 0.29$) (Figure E3). Accuracy of the cardiovascular component also varied by race and ethnicity but was overall lower, ranging from poor for Hispanic Black patients (AUROC, 0.534) to excellent for non-Hispanic White patients (AUROC, 0.816). Compared with non-Hispanic White patients, accuracy was lower for Hispanic White patients (AUROC, 0.696; $P = 0.008$) and Hispanic Black patients ($P = 0.038$), but not for non-Hispanic Black patients (AUROC, 0.712; $P = 0.12$) (Figure E4). Both respiratory and cardiovascular SOFA components were largely well calibrated (albeit with large confidence intervals) for each racial and ethnic group.

Alternate Cohorts

For the 2,862 evaluable ICU admissions (Table E2), SOFA accuracy was acceptable (AUROC, 0.772) (Figure 3A) as it was for the 629 evaluable admissions receiving MV (AUROC, 0.781) (Table E3 and Figure 3B). Accuracy was not significantly different for either cohort based on COVID-19 positivity (Figures 3C–3F); however, in both cohorts, accuracy compared with non-Hispanic White patients (ICU cohort AUROC, 0.817; MV cohort, 0.871) was lower in racial and ethnic minorities (ICU cohort, Hispanic White AUROC, 0.746; $P = 0.004$; MV cohort, Hispanic White AUROC, 0.769; $P = 0.012$; non-Hispanic Black AUROC, 0.719; $P = 0.014$) (Figure E5). SOFA was miscalibrated (both over- and underestimating mortality) for ICU admissions both with and without COVID-19 in patterns similar to those seen for the hospitalized cohort. Calibration for MV patients, independent of COVID-19 status, was better. In a sensitivity analysis of the ICU cohort evaluating maximum SOFA within only the 24 hours prior to admission, SOFA had a slightly improved accuracy (AUROC, 0.813), yet differences were found by COVID-19 status (COVID-19–positive AUROC, 0.695, vs. COVID-19–negative AUROC, 0.815; $P = 0.001$); better, but still imperfect, calibration was observed (Figure E6).

Sensitivity Analyses

Our results were robust to timeframe of maximum SOFA score. For the full cohort, accuracy of maximum SOFA within 24 hours and 72 hours of hospitalization was acceptable, and accuracy was similar for COVID-19–positive and COVID-19–negative admissions (Figures E7 and E8). Miscalibration, which was more pronounced for patients without COVID-19, was observed for both SOFA scores. Reclassifying hospice discharges as survivors led to slightly improved accuracy of the maximum SOFA within 48 hours of hospital admission (our primary exposure, AUROC, full cohort, 0.862; COVID-19–positive, 0.843; and COVID-19–negative, 0.859) and similar calibration (Figure E9).

Discussion

We found maximum SOFA scores early during hospitalization had excellent accuracy at predicting hospital mortality and, consistent with our hypothesis, had similar accuracy for hospital admissions with and without COVID-19. Likewise, no difference based on COVID-19 positivity was apparent in our primary analyses of subgroups of critically ill patients (ICU admissions and those requiring MV) using SOFA at time of critical illness onset; yet differential accuracy by COVID-19 status arose when the timeframe of SOFA evaluation was altered. Among admissions both COVID-19 positive and negative, SOFA was similarly accurate in underrepresented minorities as in non-Hispanic White patients. Maximum SOFA score at time of hospital admission and critical illness onset were both better calibrated for patients positive for COVID-19 than for those without COVID-19, where they both underestimated (for low-risk patients) and overestimated (for more moderate-risk patients) mortality. SOFA was well calibrated across racial and ethnic subgroups of patients positive for COVID-19, although confidence was lower for Black versus White individuals, and calibration varied for patients negative for COVID-19 (better for Hispanic White and Black than for non-Hispanic White or Black individuals).

The accuracy of SOFA in the setting of COVID-19 has been a subject of concern because of its use in CSC resource-allocation policies (2). Our findings of excellent SOFA accuracy (AUROC, 0.835) among COVID-19–positive hospitalizations are consistent with those of Sottile and colleagues using data from Colorado (AUROC, 0.85) (8). Ma and colleagues noted a substantially poor accuracy for COVID-19

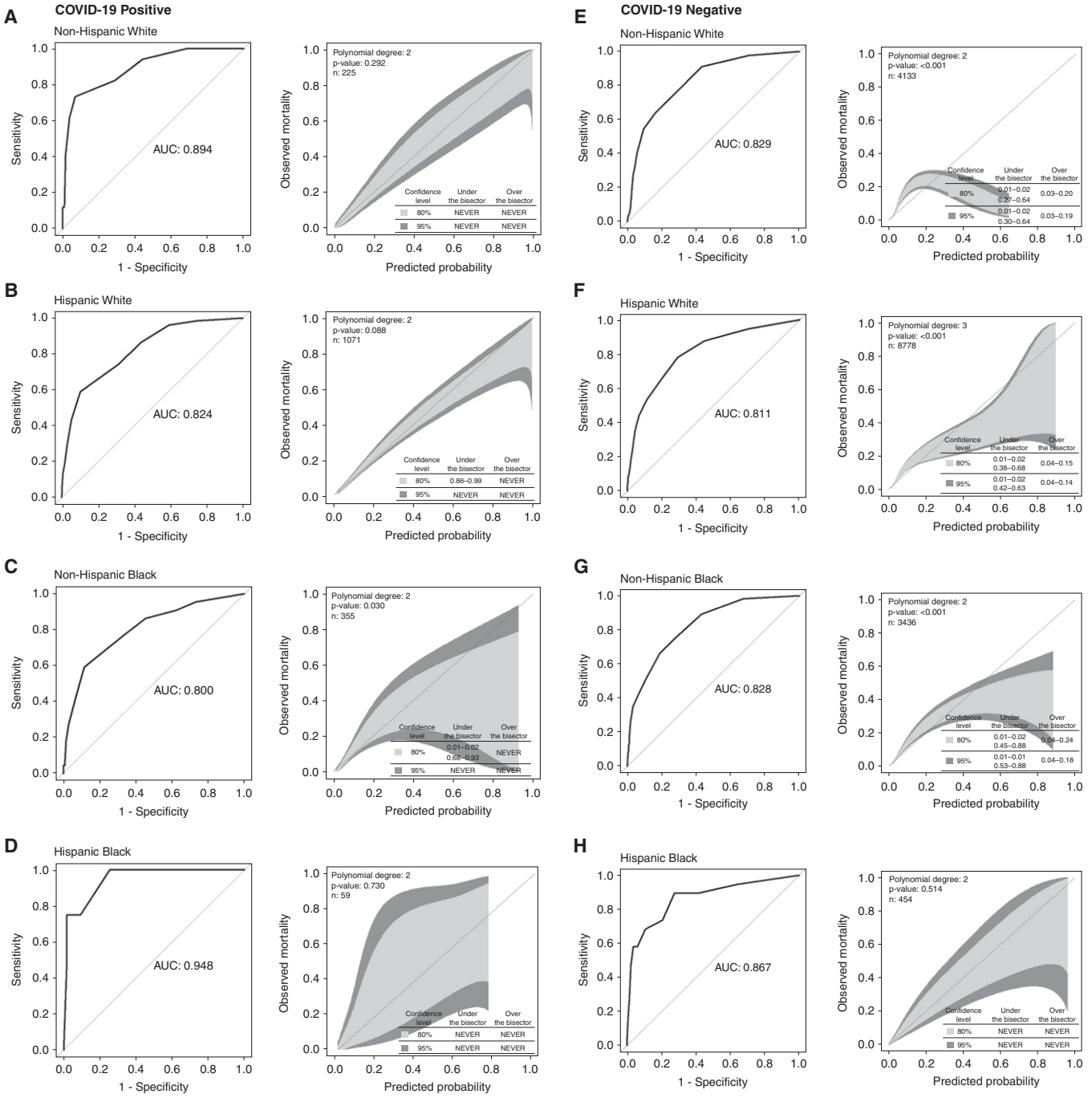

Figure 2. Discrimination and calibration of maximum sequential organ failure assessment score by coronavirus disease (COVID-19) status and race and ethnicity using Delong testing to compare discrimination with non-Hispanic White patients. (*A–D*) COVID-19 positive: Hispanic White, $P = 0.05$; non-Hispanic Black, $P = 0.12$; Hispanic Black, $P = 0.31$. (*E–H*) COVID-19 negative: Hispanic White, $P = 0.37$; non-Hispanic Black, $P = 0.97$; Hispanic Black, $P = 0.46$. AUC = area under the curve.

hospitalizations (5), yet their cohort was based in China and included patients from early on in the pandemic whose treatment and outcomes likely differed from those infected with COVID-19 more recently.

We found SOFA to be of similarly excellent accuracy (AUROC, 0.810) among admissions without COVID-19. Although not previously compared head to head, this accuracy is lower than the outstanding accuracy found by Sottile and colleagues (AUROC, 0.90) in a cohort of hospitalizations inclusive of patients both with and without COVID-19 (8). In this latter study, however, the SOFA scores evaluated were the maximum at any point during hospitalization and not just within

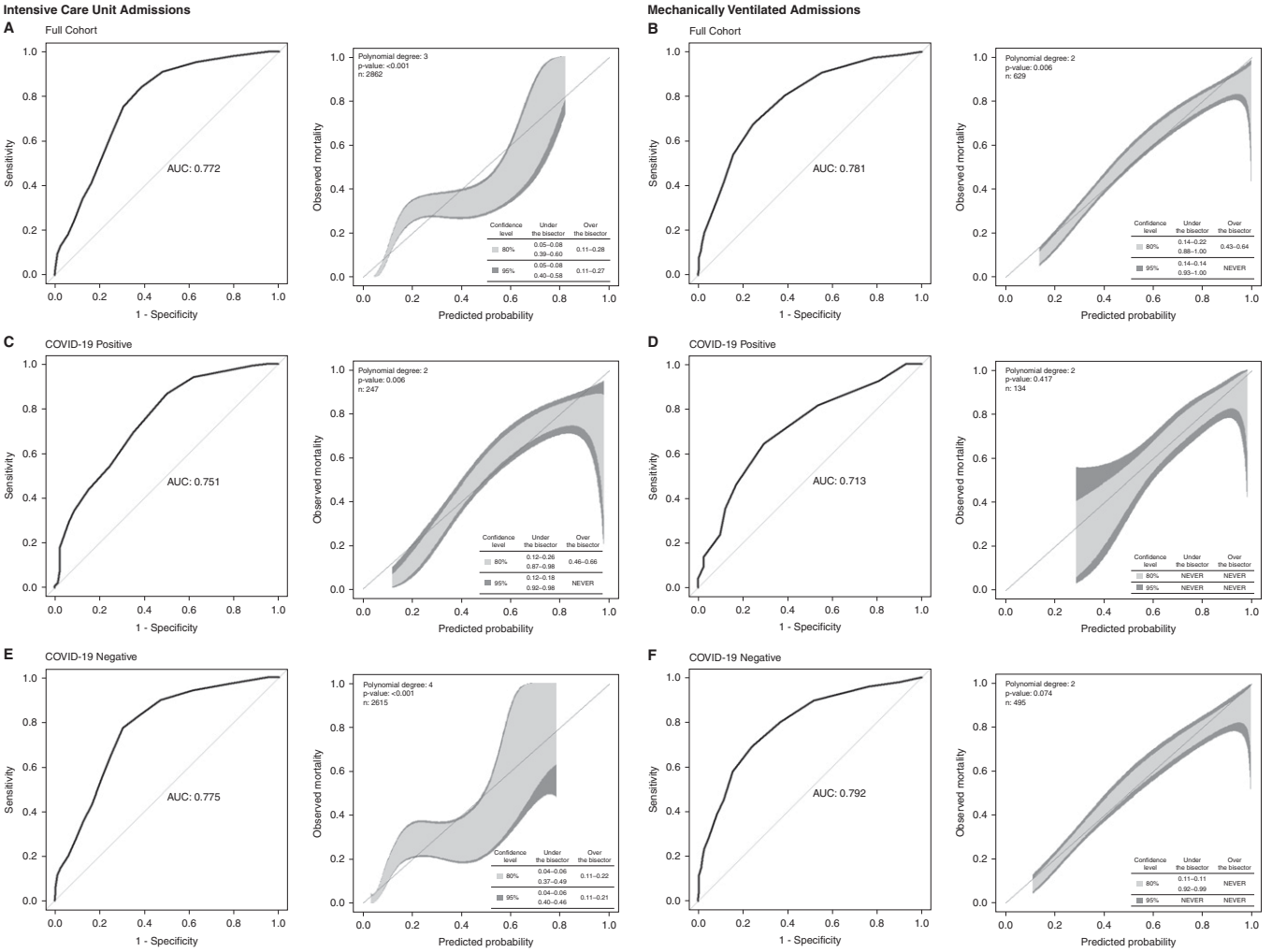

Figure 3. (*A–F*) Discrimination and calibration of maximum sequential organ failure assessment score for ICU (*A*, *C*, and *E*) and mechanically ventilated (*B*, *D*, and *F*) admissions stratified by coronavirus disease (COVID-19) status using Delong testing to compare discrimination for COVID-19–positive versus COVID-19–negative admissions. ICU admissions: $P = 0.46$; mechanically ventilated admissions: $P = 0.13$. AUC = area under the curve.

the first 48 hours following admission. It is not surprising that SOFA defined in this manner has enhanced accuracy as it incorporates more data and may include high scores occurring just prior to deaths. Whether predictions based on maximum SOFA defined in this way might differentially impact admissions based on COVID-19 status is unknown.

CSC policies must not create new or enhance existing biases against minority individuals. As such, SOFA accuracy across different racial and ethnic groups is of particular concern. For use in the current pandemic specifically, SOFA scores must be unbiased for both patients with and without COVID-19 as both populations may be "at risk" for triage away from potentially lifesaving resources if supply were limited. To our knowledge, ours is the first study to compare and find similar SOFA accuracy and calibration across racial and ethnic groups of COVID-19–positive hospital admissions. A recent study of patients negative for COVID-19 with sepsis or respiratory failure found that maximum SOFA in the emergency department was more accurate for Black (AUROC, 0.72) versus White (AUROC, 0.67; $P < 0.05$) patients and was notably differentially miscalibrated, overestimating mortality for Black and underestimating it for White patients (10). These findings contrast with ours of excellent SOFA accuracy for all racial and ethnic subgroups without COVID-19 infection and a qualitatively similar underestimation (for low-risk patients) and overestimation (for more moderate-risk patients) of mortality for both White and Black patients. Although cohort inclusion criteria and exposure definitions differed between the studies, it is not clear whether these factors are sufficient to explain the disparate results. Rather, the discrepancies suggest more study is required to understand how SOFA behaves in our typical hospitalized patient. Also in need of confirmatory study are our novel findings

that: among patients with COVID-19, SOFA is as accurate for Black and/or Hispanic individuals as it is for non-Hispanic White patients; among patients without COVID-19, calibration is better for Hispanic than non-Hispanic individuals; and, among all patients irrespective of COVID-19 status, accuracy of SOFA at the time of critical illness onset is lower for racial and ethnic minority patients than it is for non-Hispanic White patients.

The main strength of our study stems from our diverse cohort inclusive of admissions both with and without COVID-19, which *1*) mimics the population that would be exposed to a CSC policy for resource allocation; and *2*) allows for direct comparison of SOFA predictive value across subgroups. Limitations arise, however, from several areas. First, our cohort is confined to admissions within a single healthcare system in a uniquely diverse region of the United States, potentially limiting generalizability. Specifically, it is not known whether the experiences of Black and/or Hispanic patients in the South Florida area differ from those of similar patients in other parts of the country (e.g., South Florida has an abundance of Spanish-speaking clinicians, which may mitigate some aspects of disparate care). Moreover, COVID-19–related practices (e.g., use of high-flow nasal cannula or noninvasive positive pressure ventilation) likely vary substantially between hospitals. Second, although we never instituted our CSC policy, our hospitals were under strain to varying degrees throughout the period of study; if and how such strain affected care and might confound our results is unknown. Third, although rich in ethnic and Black/White racial diversity, our cohort consisted of few individuals of other racial minorities, making evaluation of SOFA accuracy in these groups impossible. Fourth, differential mortality rates may help explain the differential calibration observed across racial and ethnic groups. However, such differences in mortality will likely exist in the real-world settings in which SOFA may be applied as part of CSC.

Finally, SOFA and critical illnesses are both dynamic. Although our results were robust to different timeframes for defining maximum SOFA after hospitalization, the degrading accuracy of SOFA assessed at the onset of critical illness, when decisions about resource allocation may be required during CSC, is concerning. Moreover, patient subgroups may experience different disease courses (e.g., time to mortality may differ by COVID-19 status [Figure E10]). How consideration of critical illness dynamicity and SOFA trends over the course of illness might impact the predictive value of SOFA is unclear but is of great import if it remains integral to many CSC policies, especially if SOFA trajectory has a differential impact across patient subgroups.

Conclusions

Our findings add to a growing literature showing that SOFA may perform differently in predicting short-term mortality, specifically owing to its variable calibration, across patient subgroups (e.g., by disease type or race and ethnicity). SOFA was developed in 1996 with the express purpose of understanding the "natural history of organ dysfunction" and the "effects of new therapies" and, as noted specifically, "not to predict outcome" (1). In the intervening decades, SOFA has been widely used in research to account for illness severity and, more recently, as the cornerstone of resource allocation for many CSC policies. Prediction tools are best if they are accurate. However, perhaps more importantly, if they are to underpin life-or-death decisions, they must also be precise; similar performance across all patient groups is imperative because real-world resource allocation will never be limited to isolated subgroups but, instead, will be considered for all patients at once. CSC policies aim to ensure fair and equitable resource allocation in times of shortage, yet reliance on an imprecise predictor of short-term mortality may undermine this mission. Whether a single predictor (e.g., SOFA) can achieve this goal or if a tool comprised of different predictors for different subgroups is required remains to be determined. ■

Author disclosures are available with the text of this article at www.atsjournals.org.

References

1 Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, *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 Cleveland Manchanda EC, Sanky C, Appel JM. Crisis standards of care in the USA: a systematic review and implications for equity amidst COVID-19. *J Racial Ethn Health Disparities* 2021;8:824–836.

3 Gershengorn HB, Holt GE, Rezk A, Delgado S, Shah N, Arora 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.

4 Wunsch H, Hill AD, Bosch N, Adhikari NKJ, Rubenfeld G, Walkey A, *et al.* Comparison of 2 triage scoring guidelines for allocation of mechanical ventilators. *JAMA Netw Open* 2020;3:e2029250.

5 Ma K, Xia Y, Hu B, Hu B, Zhang Y, Xu X, Zhang N, *et al.* Development and validation of a new prognostic scoring system for COVID-19. *Jpn J Infect Dis* 2021;74:359–366.

6 Liu S, Yao N, Qiu Y, He C. Predictive performance of SOFA and qSOFA for in-hospital mortality in severe novel coronavirus disease. *Am J Emerg Med* 2020;38:2074–2080.

7 Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol* 2010;5:1315–1316.

8 Sottile PD, Albers D, DeWitt PE, Russell S, Stroh JN, Kao DP, *et al.* Real-time electronic health record mortality prediction during the COVID-19 pandemic: a prospective cohort study. *J Am Med Inform Assoc* 2021;28:2354–2365.

9 Sarkar R, Martin C, Mattie H, Gichoya JW, Stone DJ, Celi LA. 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.

10 Ashana DC, Anesi GL, Liu VX, Escobar GJ, Chesley C, Eneanya ND, *et al.* Equitably allocating resources during crises: racial differences in mortality prediction models. *Am J Respir Crit Care Med* 2021;204:178–186.

11 Rice TW, Wheeler AP, Bernard GR, Hayden DL, Schoenfeld DA, Ware LB; National Institutes of Health, National Heart, Lung, and Blood Institute ARDS Network. Comparison of the SpO2/FIO2 ratio and the PaO2/FIO2 ratio in patients with acute lung injury or ARDS. *Chest* 2007;132:410–417.

12 Raschke RA, Agarwal S, Rangan P, Heise CW, Curry SC. Discriminant accuracy of the SOFA score for determining the probable mortality of patients with COVID-19 pneumonia requiring mechanical ventilation. *JAMA* 2021;325:1469–1470.

13 White DB, Katz M, Luce J, *et al.* Allocation of scarce critical care resources during a public health emergency. Pittsburgh, PA: University of Pittsburgh; 2020.

14 Nattino G, Lemeshow S, Phillips G, Finazzi S, Bertolini G. Assessing the calibration of dichotomous outcome models with the calibration belt. *Stata J* 2017;17:1003–1014.