ORIGINAL ARTICLE

Equitably Allocating Resources during Crises
Racial Differences in Mortality Prediction Models

Deepshikha Charan Ashana[1,2], George L. Anesi[2,3,4], Vincent X. Liu[5], Gabriel J. Escobar[5], Christopher Chesley[2,3,4], Nwamaka D. Eneanya[2,4,6], Gary E. Weissman[2,3,4], William Dwight Miller[7], Michael O. Harhay[2,3,4,8], and Scott D. Halpern[2,3,4,8,9]

[1]Division of Pulmonary, Allergy, and Critical Care Medicine, Department of Medicine, Duke University, Durham, North Carolina; [2]Palliative and Advanced Illness Research Center, [3]Division of Pulmonary, Allergy, and Critical Care Medicine, Department of Medicine, [4]Leonard Davis Institute of Health Economics, [6]Renal-Electrolyte and Hypertension Division, [8]Department of Biostatistics, Epidemiology, and Informatics, and [9]Department of Medical Ethics and Health Policy, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania; [5]Division of Research, Kaiser Permanente, Oakland, California; and [7]Section of Pulmonary and Critical Care Medicine, Department of Medicine, University of Chicago, Chicago, Illinois

ORCID IDs: 0000-0003-2107-2110 (D.C.A.); 0000-0003-4585-0714 (G.L.A.); 0000-0001-9588-3819 (G.E.W.); 0000-0002-0553-674X (M.O.H.).

Abstract

Rationale: Crisis standards of care (CSCs) guide critical care resource allocation during crises. Most recommend ranking patients on the basis of their expected in-hospital mortality using the Sequential Organ Failure Assessment (SOFA) score, but it is unknown how SOFA or other acuity scores perform among patients of different races.

Objectives: To test the prognostic accuracy of the SOFA score and version 2 of the Laboratory-based Acute Physiology Score (LAPS2) among Black and white patients.

Methods: We included Black and white patients admitted for sepsis or acute respiratory failure at 27 hospitals. We calculated the discrimination and calibration for in-hospital mortality of SOFA, LAPS2, and modified versions of each, including categorical SOFA groups recommended in a popular CSC and a SOFA score without creatinine to reduce the influence of race.

Measurements and Main Results: Of 113,158 patients, 27,644 (24.4%) identified as Black. The LAPS2 demonstrated higher discrimination (area under the receiver operating characteristic curve [AUC], 0.76; 95% confidence interval [CI], 0.76–0.77) than the SOFA score (AUC, 0.68; 95% CI, 0.68–0.69). The LAPS2 was also better calibrated than the SOFA score, but both underestimated in-hospital mortality for white patients and overestimated in-hospital mortality for Black patients. Thus, in a simulation using observed mortality, 81.6% of Black patients were included in lower-priority CSC categories, and 9.4% of all Black patients were erroneously excluded from receiving the highest prioritization. The SOFA score without creatinine reduced racial miscalibration.

Conclusions: Using SOFA in CSCs may lead to racial disparities in resource allocation. More equitable mortality prediction scores are needed.

Keywords: critical care; triage; sepsis; acute respiratory failure; disaster planning

At a Glance Commentary

Scientific Knowledge on the Subject: Crisis standards of care have been developed to guide fair allocation of scarce critical care resources during the coronavirus disease (COVID-19) pandemic and other crises. Most recommend using the Sequential Organ Failure Assessment (SOFA) score to prioritize patients with the highest chances of short-term survival to receive scarce resources. However, it is unknown how SOFA performs among patients of different races.

What This Study Adds to the Field: In a cohort of Black and white patients with sepsis and acute respiratory failure, we found that the SOFA score is miscalibrated in a way that would systematically divert critical care resources away from Black patients. Therefore, using the SOFA score in crisis standards of care may lead to racial disparities in resource allocation.

(Received in original form December 9, 2020; accepted in final form March 22, 2021)

Supported by National Heart, Lung, and Blood Institute grant R01HL136719 (S.D.H.) and NIH grant R35GM128672 (V.X.L.).

Author Contributions: All authors fulfill International Committee of Medical Journal Editors criteria for authorship.

Correspondence and requests for reprints should be addressed to Deepshikha Charan Ashana, M.D., M.S., M.B.A., Duke University, Hanes House, 315 Trent Drive, Box 102352, Durham, NC 27710. E-mail: deepshikha.ashana@duke.edu.

This article has a related editorial.

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

Am J Respir Crit Care Med Vol 204, Iss 2, pp 178–186, Jul 15, 2021

Originally Published in Press as DOI: 10.1164/rccm.202012-4383OC on March 22, 2021
Internet address: www.atsjournals.org

The coronavirus disease (COVID-19) pandemic has resulted in several surges of patients with serious illness, raising concerns about overwhelming local supplies of critical care resources (1). In such situations, crisis standards of care (CSCs) are used to guide fair allocation of scarce resources (2, 3). CSCs promote the ethical goals of maximizing benefits for populations, which is often defined as saving the most lives or life-years, and of distributing these benefits fairly among groups (4, 5).

To achieve these goals, more than 75% of CSCs currently in use in the United States recommend ranking patients according to their likelihood of surviving their hospital stay using the Sequential Organ Failure Assessment (SOFA) score (2). A widely adopted CSC describes dividing the SOFA score into four categories by assigning 1 point to patients in the lowest SOFA category, indicating the highest likelihood of survival to hospital discharge, and an additional point for each subsequent category. After adding additional points for patients with reduced chances for near-term survival beyond the hospitalization, giving consideration to subtracting points for essential workers, and using younger age as a tiebreaker, patients with the fewest points are given the highest priority for critical care resources (6). Thus, patients with greater chances of in-hospital and near-term survival are most strongly prioritized.

However, the SOFA score may not be well suited for predicting mortality because it was developed to describe sepsis-related organ dysfunction (7, 8), and its developers cautioned that it "is designed not to predict outcome but to describe a sequence of complications in the critically ill" (9). Nevertheless, as organ failure often leads to mortality, subsequent studies have confirmed the prognostic accuracy of the SOFA score for in-hospital mortality, most commonly among critically ill patients with sepsis (10–12).

A more recently levied concern with the SOFA score is that it was derived in a predominantly European population, and its prognostic accuracy among Black and white patients in the United States is unknown. Indeed, there is growing concern about whether creatinine, a component of the SOFA renal subscore, should be adjusted for race (13, 14). As Black patients were found to have higher creatinine than white patients with the same glomerular filtration rate in prior studies, some caution that the lack of race-based modification may lead to systematically higher SOFA scores and lower CSC priority scores among Black patients (15, 16). Others believe that race-based modifications are invalid, in part because of the greater genetic differences within rather than between races, a difficulty in classifying patients according to race, and the risk of conflating race and racism (17–19). Yet another source of concern is that because treatments (e.g., respiratory support device use) affect patients' SOFA scores, the racial differences in access to, preferences for, or physician provision of these therapies may inappropriately influence the scores (20–23).

To promote racial equity in CSCs, we compared the prognostic accuracy of the SOFA score and version 2 of the Laboratory-based Acute Physiology Score (LAPS2) (24, 25), which is mentioned as an alternative to SOFA in a widely adopted CSC (26). We assessed the performance of these scores among Black and white patients admitted through the emergency department (ED) with sepsis or acute respiratory failure (ARF). We focused on this population because most CSC-based triage decisions would be made in the ED and because when patients with COVID-19 require critical care, this requirement is most commonly due to sepsis or ARF (27, 28).

Methods

The University of Pennsylvania and Kaiser Permanente institutional review boards approved the study. We followed the Strengthening the Reporting of Observational Studies in Epidemiology guidelines (29).

Study Population and Data Sources

We conducted a retrospective cohort study of Black and white patients who received care at 27 Kaiser Permanente Northern California and Penn Medicine hospitals between 2013 and 2018. We used electronic health record data to identify patients $\geq$18 years of age with sepsis at all sites and ARF at Penn Medicine sites who were admitted from the ED to an inpatient location (i.e., a ward, a step-down unit, or an ICU). Sepsis and ARF definitions in this cohort have previously been published (30, 31). In brief, sepsis or ARF must have been diagnosed in patients while they were in the ED on the basis of the adaptations of the Third International Consensus Definitions for Sepsis and Septic Shock criteria and physiologic and clinical indicators of respiratory failure (32). We excluded patients who had a code status other than a full code status, as this may impact both their SOFA pulmonary subscore, which includes the use of respiratory support devices (9), and their likelihood of dying in the hospital. We also excluded the 26% of patients who identified as being Asian, Native American, Hawaiian, a Pacific Islander, or of mixed race or who identified their race as "other" because our hypotheses related to differences between white and Black patients.

Study Variables

The SOFA renal subscore was calculated using creatinine alone, as urine output was not reliably recorded in the ED. We used the highest value for each SOFA subscore during the ED stay in our calculation of the total SOFA score (9).

The primary comparator with the SOFA score was the LAPS2, which includes more physiologic and laboratory data (including creatinine) than SOFA but includes no treatment variables (6, 24). We also created and tested several modifications of the SOFA score and the LAPS2. For the SOFA score, we tested the following four specifications: *1*) its original form as a continuous variable from 0 to 24 points; *2*) division into four categories ($<$6, 6–8, 9–11, and $\geq$12) as proposed in a commonly used CSC to facilitate use at the point of care (6); *3*) a partial modification to the renal subscore, in which we subtracted one-half of a point (equal to the difference between mean renal subscores between Black and white patients in this cohort) from the renal subscore for Black patients whose raw renal subscore was $>$0; and *4*) a SOFA score in which we eliminated the renal subscore entirely. For the LAPS2, we tested the following three specifications: *1*) its original form as a continuous variable from 0 to 414 points (scores $>$200 being uncommon), *2*) a continuous LAPS2 divided into eight equal categories (i.e., on the basis of the range of LAPS2s), and *3*) a continuous LAPS2 divided into four equal categories. Modifying the creatinine component of the LAPS2 was not possible because it is a two-stage prediction model that does not merely sum scores on different variables as SOFA does.

The primary outcome was in-hospital mortality, defined as death during the hospital stay or discharge to hospice.

Statistical Analyses

Analyses were performed with Stata/IC 14.2 (StataCorp LLC) and R (R Core Team, R Foundation for Statistical Computing). We compared patient characteristics using the chi-

square and *t* tests for categorical and continuous data, respectively.

To promote ease of use and prevent bias due to age, comorbidities, or other patient-level factors, most CSCs calculate a priority score using an unadjusted SOFA score or another score (6, 33, 34). Under an ideal CSC, Black and white patients with the same priority score (or in the same score category) would have equal likelihoods of dying in the hospital (35, 36). Therefore, to test the independent association of race with in-hospital mortality for each specification of the unadjusted SOFA score and LAPS2, we fit logistic regression models including race as the independent variable, in-hospital death as the dependent variable, and the mortality prediction score as a covariate.

The prognostic accuracy of all seven specifications of the unadjusted mortality prediction scores for in-hospital mortality was assessed on the basis of model discrimination and calibration (37). We first fit baseline logistic regression models using only in-hospital death and mortality prediction scores, as the SOFA score and LAPS2 will not generally be adjusted for other covariates in triage situations. For each model, we then calculated the area under the receiver operating characteristic curve (AUC) for all patients, Black patients, and white patients. Using a prior framework, we considered an AUC below 0.7 to be poor, an AUC of 0.7–0.8 to be acceptable, an AUC of 0.8–0.9 to be excellent, and an AUC higher than 0.9 to be outstanding (38). We compared the equality of AUCs using the DeLong test (39). We performed two sensitivity analyses. First, we repeated discrimination analyses including the center (i.e., hospital) as a random effect in the baseline logistic regression model, as triage officers will compare mortality prediction scores for patients within a hospital. Second, we repeated discrimination analyses after redefining the outcome to only include patients who died during their hospital stay.

To assess calibration, we calculated probabilities for in-hospital mortality predicted by each model, followed by the creation of calibration belts for all patients, Black patients, and white patients to compare predicted and observed mortality. These enabled visualization of the range and type of miscalibration and provided a statistical assessment of significant deviations from the bisector (i.e., the line of perfect calibration in which predicted and observed outcomes are equivalent) on the basis of the likelihood-ratio test (40). Two-sided P values ≤ 0.05 were considered to indicate significance.

We also derived a numerical measure of miscalibration, the integrated calibration index (ICI), which is the average of the absolute difference between observed and predicted probabilities weighted by the density of predicted probabilities (41). A perfectly calibrated model would have an ICI of zero. Because ICIs measure the magnitude but not the direction of miscalibration, they can be helpful in comparisons across models but not for comparisons between racial groups, in which the directions of errors strongly influence considerations of distributional equity.

To quantify the impact of model miscalibration, we calculated the number of Black patients who were inappropriately excluded from the highest-priority category (SOFA score < 6) on the basis of in-hospital mortality risk. Black and white patients in each of the four SOFA categories proposed in the most common CSC ought to have the same average in-hospital mortality (6). Therefore, we sequentially moved Black patients whose SOFA scores were marginally outside the highest-priority category (i.e., those with SOFA scores of 6, then 7, and so on) into the highest-priority category (which already included all Black and white patients with SOFA scores <6) until the average mortality for Black patients in this category approximated but did not exceed that of white patients in this category. We reasoned that the proportion of Black patients with SOFA scores >5 who would be reclassified as having the highest priority under an equitable model provides an estimate of the impact of the racial miscalibration of the SOFA score.

Finally, to determine whether adjustment for other variables that might plausibly be included in triage models improved performance, we conducted secondary analyses assessing discrimination and calibration after adjusting models for age, sex, and comorbidities using version 2 of the Comorbidity Point Score (COPS2). COPS2 is a measure of chronic illness and includes diagnoses from the electronic health record for the 12 months preceding patients' ED encounters (42).

Results

The final study sample of 113,158 patients included 75,942 with sepsis, 10,840 with ARF, and 26,376 with both. Overall, 24.4% of patients were Black, and 1.33% were Hispanic. Compared with white patients, Black patients were younger (mean age, 61.7 yr vs. 67.7 yr)

Table 1. Patient Characteristics

Characteristic	White ($n = 85,514$)*	Black ($n = 27,644$)*	*P* Value
Age, mean (SD)	67.7 (15.2)	61.7 (16.6)	<0.001
Sex, F, %	45.9	51.8	<0.001
SOFA score, mean (SD)	2.9 (1.8)	3.1 (2.1)	<0.001
SOFA subscores, mean (SD)	—	—	<0.001
Respiratory	0.5 (1.0)	0.4 (0.9)	<0.001
Coagulation	0.4 (0.8)	0.3 (0.7)	<0.001
Hepatic	0.3 (0.7)	0.3 (0.7)	<0.001
Cardiovascular	0.5 (0.5)	0.4 (0.5)	<0.001
Central nervous system	0.3 (0.7)	0.4 (1.0)	<0.001
Renal	0.8 (1.1)	1.3 (1.3)	<0.001
LAPS2, mean (SD)	103.1 (36.7)	102.2 (38.4)	<0.001
In-hospital mortality, %	8.6	7.5	<0.001

Definition of abbreviations: LAPS2 = version 2 of the Laboratory-based Acute Physiology Score; SOFA = Sequential Organ Failure Assessment.
*Data were missing for age ($n = 2,344$), the female sex ($n = 2,345$), LAPS2 ($n = 5$), and in-hospital mortality ($n = 1,144$).

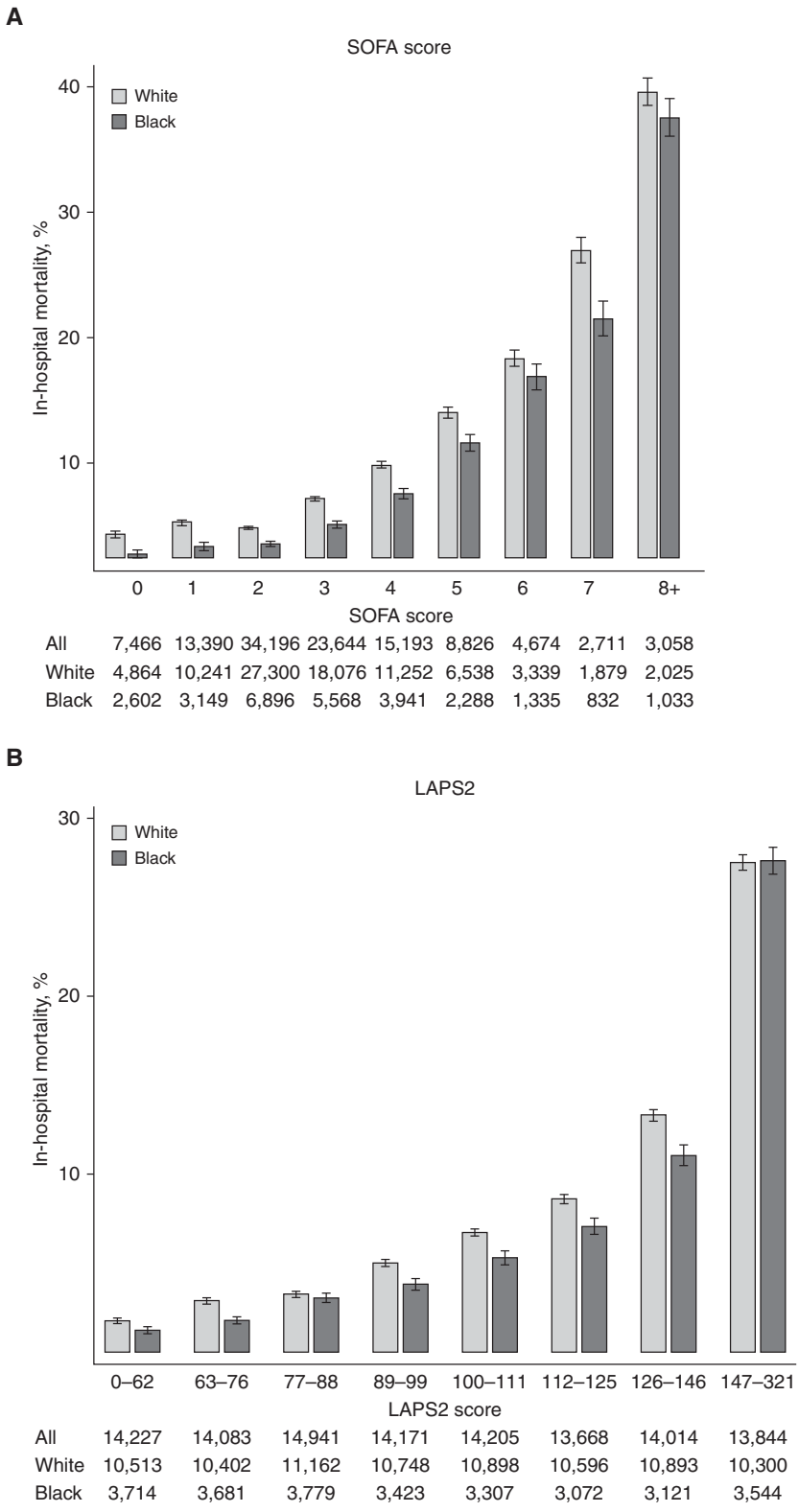

Figure 1. In-hospital mortality among white and Black patients by a mortality prediction score. (*A*) Sequential Organ Failure Assessment (SOFA) score. Those with a SOFA score of 8 or greater were included in one category, as few patients had very high SOFA scores. (*B*) LAPS2. In-hospital mortality was calculated for white and Black patients in each mortality prediction score category. The table beneath each graph demonstrates the total number of patients in each category. LAPS2 = version 2 of the Laboratory-based Acute Physiology Score.

Table 2. AUC for Mortality Prediction Scores among all Patients, White Patients, and Black Patients

Mortality Prediction Score	AUC (95% CI)		
	All Patients	White Patients	Black Patients
Original SOFA score	0.68 (0.68–0.69)	0.67 (0.66–0.68)	0.72 (0.71–0.73)*
CSC SOFA score categories†	0.61 (0.60–0.61)‡	0.60 (0.59–0.60)	0.65 (0.64–0.66)*
SOFA score with creatinine modification§	0.69 (0.69–0.69)‡	0.67 (0.66–0.68)	0.72 (0.71–0.74)*
SOFA score without creatinine	0.68 (0.67–0.69)	0.66 (0.66–0.67)	0.73 (0.72–0.74)*
Original LAPS2	0.76 (0.76–0.77)‡	0.75 (0.75–0.76)	0.79 (0.77–0.80)*
Eight-category LAPS2‖	0.74 (0.74–0.75)‡	0.73 (0.73–0.74)	0.77 (0.76–0.78)*
Four-category LAPS2¶	0.69 (0.69–0.70)‡	0.68 (0.68–0.69)	0.72 (0.71–0.73)*

Definition of abbreviations: AUC = area under the receiver operating characteristic curve; CI = confidence interval; CSC = crisis standard of care; LAPS2 = version 2 of the Laboratory-based Acute Physiology Score; SOFA = Sequential Organ Failure Assessment.
*Significant difference ($P < 0.05$) between white and Black patients for a given mortality prediction score.
†SOFA score divided into four categories: <6, 6–8, 9–11, and ⩾12.
‡Significant difference ($P < 0.05$) compared with SOFA score among all patients.
§One-half of a point subtracted from SOFA score if renal component of the SOFA score is >0 among Black patients.
‖LAPS2 divided into eight categories: <40, 40–79.9, 80–119.9, 120–159.9, 160–199.9, 200–239.9, 240–279.9, and >280.
¶LAPS2 divided into four categories: <80, 80–159.9, 160–239.9, and >240.

and more likely to be female (51.8% vs. 45.9%) ($P < 0.001$ for both) (Table 1). On average, Black patients had higher overall SOFA scores (3.1 vs. 2.9), primarily driven by their higher renal subscore (1.3 vs. 0.8) and slightly higher central nervous system subscore (0.4 vs. 0.3) ($P < 0.001$ for all). In contrast, Black patients had lower mean LAPS2s (102.2 vs. 103.1) and were less likely to die in the hospital (7.5% vs. 8.6%) than white patients.

Comparing patients with the same SOFA score or LAPS2 or range of scores, Black patients had lower in-hospital mortality than white patients in most categories (Figure 1). Consistent with this result, Black race was associated with significantly lower in-hospital mortality compared with white race after adjustment for each of the seven specifications of the SOFA score and LAPS2 in separate regression models (*see* Table E1 in the online supplement).

Discrimination

The SOFA score had poor discrimination for in-hospital mortality in this cohort (AUC, 0.68; 95% confidence interval [CI], 0.68–0.69), with even poorer discrimination when using the CSC SOFA score categories (AUC, 0.61; 95% CI, 0.60–0.61; $P<0.05$). Compared with the original SOFA score, the two modifications to the SOFA renal subscore did not result in meaningfully different AUCs (AUC, 0.69; 95% CI, 0.69–0.69 and AUC, 0.68; 95% CI, 0.67–0.69 for the SOFA score with creatinine modification and the SOFA score without creatinine, respectively) (Table 2). The LAPS2 had acceptable discrimination, which was significantly greater than that of the SOFA score (AUC, 0.76; 95% CI, 0.76–0.77; $P < 0.05$). Categorizing the LAPS2 into fewer categories resulted in incrementally lower discrimination (AUC, 0.74; 95% CI, 0.74–0.75 and AUC, 0.69; 95% CI, 0.69–0.70 for the eight-category and four-category LAPS2s, respectively; $P < 0.05$, comparing either score with the original LAPS2). All specifications of the mortality prediction scores had higher discrimination among Black patients than among white patients. In sensitivity analyses, adjusting for the center or excluding patients who were discharged to hospice from the outcome definition did not meaningfully change discrimination for the SOFA score or LAPS2, nor did it attenuate differences in discrimination between white and Black patients (Tables E2 and E3).

Calibration

Figures 2 and E1 show calibration belts for *1*) the SOFA score, *2*) the LAPS2, and *3*) the SOFA score without creatinine, as this model resulted in the best calibration among the SOFA score modifications. The SOFA score was significantly miscalibrated for all patients, Black patients, and white patients. The LAPS2 was perfectly calibrated for all patients but underestimated in-hospital mortality for white patients whose expected mortality was 2–16% and overestimated in-hospital mortality for Black patients whose expected mortality was 1–27%. The LAPS2 had the lowest ICI (best calibration) of the models tested among all patients and among each racial subgroup (Table E4). The SOFA score excluding creatinine showed the narrowest range of miscalibration (ranges of risk in which the confidence limits excluded perfect calibration) across patient groups. The remaining models had worse calibration than this modified SOFA score (Figure E1).

Recategorizing In-Hospital Mortality

Compared with white patients, Black patients in the highest-priority CSC category (SOFA score < 6) had lower in-hospital mortality (5.3% vs. 6.9%). Reclassifying Black patients with SOFA scores between 6 and 8 ($N = 2{,}611$; 9.4% of all Black patients and 81.6% of Black patients with SOFA scores >5) into the highest-priority category resulted in similar in-hospital mortality for Black and white patients in this category (6.7% vs. 6.9%) (Figure 3).

Adjusting for Age, Sex, and Comorbidities

Compared with the results of our unadjusted analyses, adjusting for age, sex, and COPS2 improved discrimination for all models. The magnitude of change was greater for the SOFA score and its modifications than for the LAPS2 and its modifications (Table E5). However, all adjusted models were miscalibrated. In-hospital mortality among Black patients was still overestimated in most models and never underestimated. Among white patients, calibration errors were present in both directions. Nearly all models continued to underestimate in-hospital mortality for white patients in the lower ranges of predicted mortality, representing the majority of white patients in the sample (Table E5).

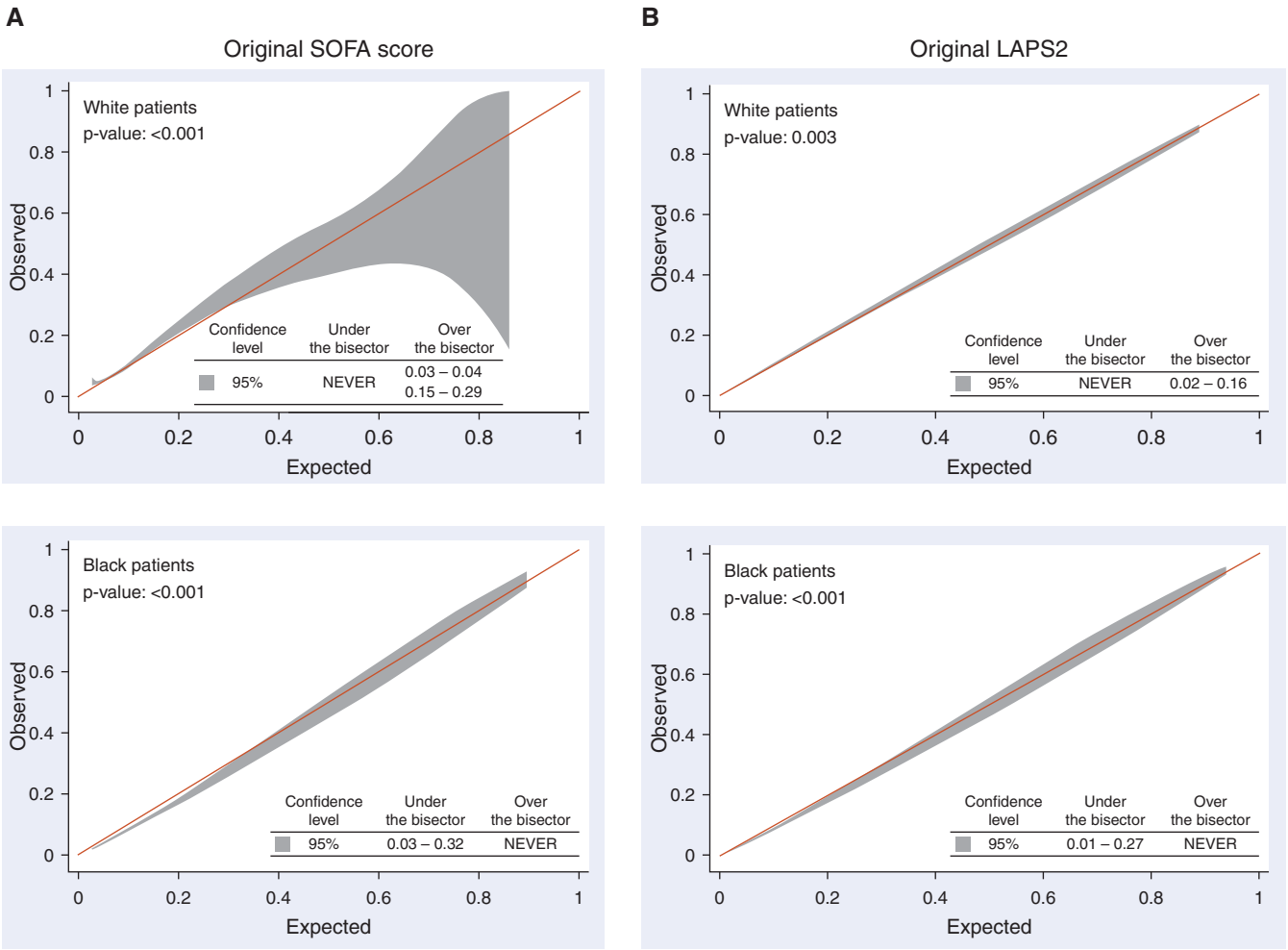

Figure 2. Calibration belts for selected mortality prediction scores among white and Black patients. (*A*) Original Sequential Organ Failure Assessment (SOFA) score. (*B*) Original LAPS2. (*C*) SOFA score without creatinine. A *P* value < 0.05 indicates miscalibration. Reported at the bottom of each graph are values of expected mortality for which observed values are significantly under (i.e., the model overestimates mortality) or over (i.e., the model underestimates mortality) the bisector using the 95% confidence interval (shaded area of the graph). LAPS2 = version 2 of the Laboratory-based Acute Physiology Score.

Discussion

Among patients with sepsis and ARF admitted to 27 U.S. hospitals, we found that the SOFA score had poor prognostic accuracy for in-hospital mortality. This might be expected, given the original intent of the SOFA score, but its poor performance is notable because of its widespread use in CSCs. Second, the LAPS2 had better overall discrimination than the SOFA score and had the best calibration among all patient groups of the models we tested. However, both scores were miscalibrated within racial subgroups in ways that could systematically divert critical care resources away from Black patients if used in CSCs. Third, compared with the original SOFA score, eliminating the renal subscore improved calibration without changing discrimination. Finally, consolidating continuous mortality prediction scores into few categories, as is advocated for the SOFA score in some CSCs, worsened discrimination without meaningfully changing calibration among racial subgroups.

Prior evaluations of the SOFA score have demonstrated higher discrimination for in-hospital mortality risk than were found in this study (10–12). However, the present cohort includes a more diverse group of patients defined by clinical criteria gleaned from the electronic health record, all of whom were initially treated in the ED. Differences in cohort inclusion criteria can influence measures of prognostic accuracy, highlighting the importance of testing the external validity of prediction models in new populations.

We found that the SOFA score underestimated in-hospital mortality risk for white patients and overestimated it for Black patients. Because CSCs prioritize patients with lower predicted mortality risk, these errors in mortality estimation could systematically divert critical care resources from Black patients to white patients, despite no true mortality risk differences. Importantly, the miscalibration occurred among patients with an expected mortality rate of less than 30% who are expected to derive the greatest benefit from having access to such resources. Similar calibration errors persisted after adjusting models for age, sex, and comorbidities, suggesting that other factors, which may be

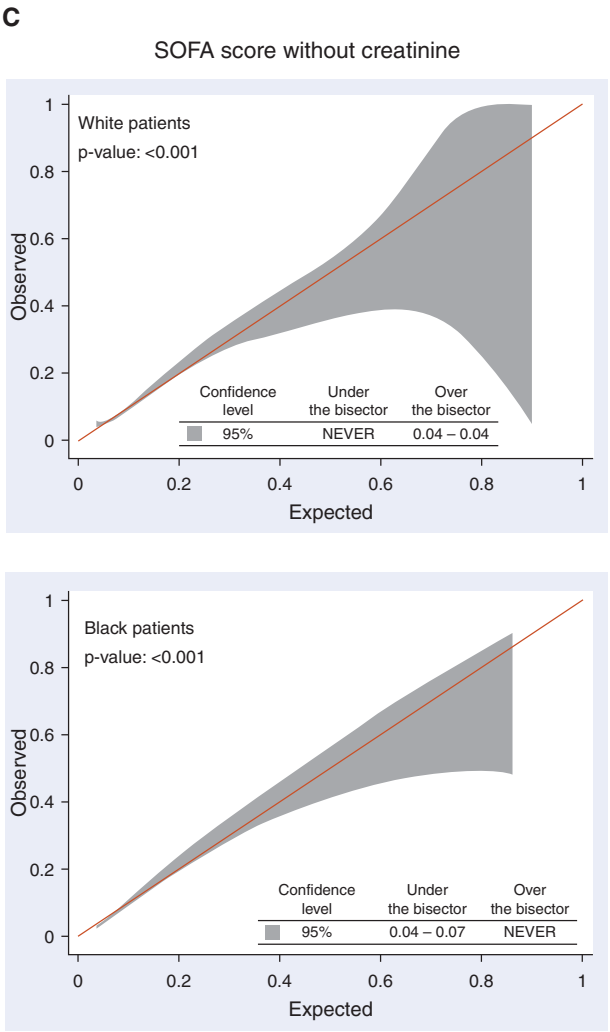

Figure 2. (*Continued*).

more difficult to include in CSCs (e.g., cumulative health effects of structural racism), differentially contribute to in-hospital mortality risk in Black and white patients. Using the SOFA categories proposed in a widely distributed CSC (6), we found that the impact of this miscalibration would be significant, as it would inappropriately prevent 9.4% of all Black patients and 81.6% of Black patients who were not originally in the highest-priority category from receiving the highest priority for critical care resources. Given the SOFA score's poor overall discrimination and disparate miscalibration among racial subgroups, we caution against its use in CSCs.

We evaluated two modifications to the renal component of the SOFA score. The SOFA score without creatinine resulted in better model performance than the SOFA score with the renal component modified. This result is consistent with evidence that creatinine is only a moderately accurate measure of the glomerular filtrate rate in racially diverse populations (43) and supports recent calls to move away from racialized medicine, as race is a social construct and is thus a poor proxy for biological measures (13, 18). Removing creatinine from mortality prediction models or replacing it with accurate measures of renal function that do not perform differently across races, such as cystatin C, may improve model performance but may be difficult to implement, given the acute need for CSCs (13, 44).

We found that the LAPS2 had discrimination superior to that of the SOFA score and had the best calibration among all models tested, perhaps because of its greater scale, inclusion of additional physiologic variables, and exclusion of treatment variables (45). However, the differential miscalibration within racial subgroups that we observed with the SOFA score persisted with the LAPS2. Unfortunately, because the LAPS2 is a two-stage prediction model, there is no straightforward method for removing creatinine from the LAPS2 calculation as was done with the SOFA score. Thus, although the LAPS2 might be preferable to the SOFA score in CSCs, future studies are needed to determine whether a new LAPS2 model that does not incorporate creatinine would enable more equitable calibration without worsening model discrimination.

Authors of CSCs should carefully weigh the harms and benefits of collapsing continuous mortality prediction scores into fewer categories. Although grouping patients into categories may facilitate time-sensitive application under crisis and reduce the

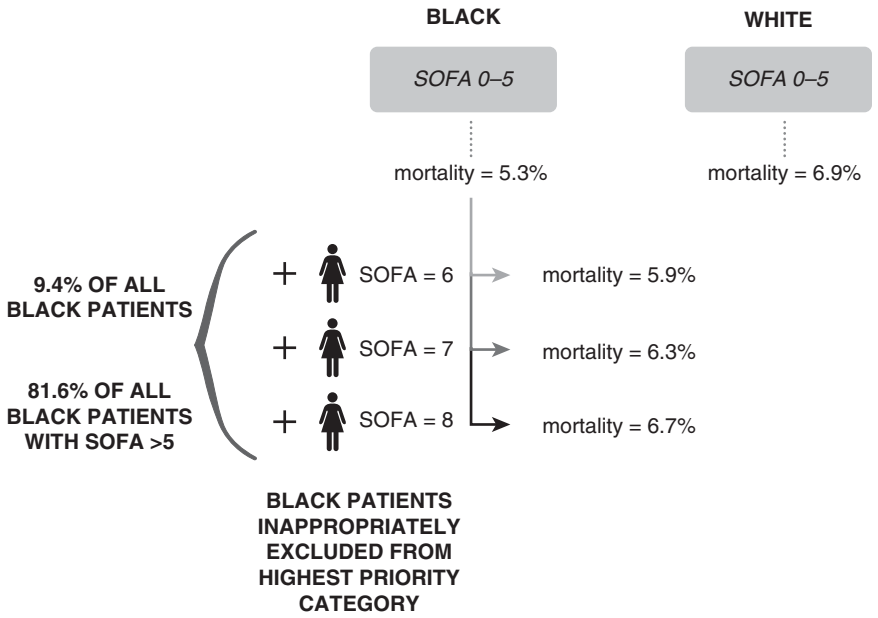

Figure 3. Black patients are inappropriately excluded from receiving the highest priority on the basis of in-hospital mortality in a commonly used crisis standard of care. In this simulated scenario, the Black patients with SOFA scores between 6 and 8 were sequentially reclassified as being the highest priority until the overall mortality risk for Black patients in the highest-priority category approximated, but did not exceed, that of white patients in this category. SOFA = Sequential Organ Failure Assessment.

chances that small risk differences could change determinations of priority, this study quantifies the extent to which categorizing continuous mortality risk scores reduces predictive use. Real-world or simulated comparisons are needed to determine how many lives are saved across and within racial groups when categorized versus continuous mortality prediction models are used in CSCs. Finally, without correcting the effects of structural racism in the United States, it may be that all prediction models display a degree of racial bias in calibration. Therefore, although recalibration prediction models to mitigate racial bias may be helpful, concurrent consideration and empirical evaluation of alternate remedies will also be needed (46).

This study has limitations. First, our cohort predates COVID-19. However, our sample reflects many of the patients who would be considered for critical care even during a COVID-19 surge. Second, our calculation of the SOFA renal subscore was based solely on creatinine and did not include urine output, potentially reducing prognostic accuracy. However, because urine output is inaccurately recorded in many triage settings, it is often excluded from mortality prediction scores in the context of CSCs. Third, we only included Black and white patients on the basis of our hypothesized differences in model performance between these two races. Given the high prevalence of COVID-19 among all racial and ethnic minority communities (47), future studies should test the prognostic accuracy of mortality prediction scores among Asian patients, Latinx patients, and patients of other races or ethnicities.

In conclusion, we found that, compared with the SOFA score, the LAPS2 had superior discrimination and calibration for in-hospital mortality among inpatients admitted from the ED with sepsis or ARF. However, differences in calibration among Black and white patients using either score may cause disparities in resource allocation. Future studies should examine the racial impacts of a new LAPS2 without creatinine. However, because all acuity scores are likely to display racial bias, more research is needed to understand how to improve the application of these scores in promoting equitable resource allocation. ■

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

Acknowledgment: The authors thank Ms. Nidhi Charan for creating Figure 3.

References

1. Ranney ML, Griffeth V, Jha AK. Critical supply shortages: the need for ventilators and personal protective equipment during the COVID-19 pandemic. *N Engl J Med* 2020;382:e41.
2. Antommaria AHM, Gibb TS, McGuire AL, Wolpe PR, Wynia MK, Applewhite MK, *et al.* Ventilator triage policies during the COVID-19 pandemic at U.S. hospitals associated with members of the association of bioethics program directors. *Ann Intern Med* 2020;173:188–194.
3. Committee on Guidance for Establishing Crisis Standards of Care for Use in Disaster Situations, Institute of Medicine. Crisis standards of care: a systems framework for catastrophic disaster response. Washington, DC: National Academies Press; 2012.
4. Emanuel EJ, Persad G, Upshur R, Thome B, Parker M, Glickman A, *et al.* Fair allocation of scarce medical resources in the time of COVID-19. *N Engl J Med* 2020;382:2049–2055.
5. White DB, Lo B. A framework for rationing ventilators and critical care beds during the COVID-19 pandemic. *JAMA* 2020;323:1773–1774.

6. White DB. Allocation of scarce critical care resources during a public health emergency. Pittsburgh, PA: University of Pittsburgh; 2020 [accessed 2020 Apr 15]. Available from: https://ccm.pitt.edu/?q=content/model-hospital-policy-allocating-scarce-critical-care-resources-available-online-now.
7. Vincent JL, de Mendonça A, Cantraine F, Moreno R, Takala J, Suter PM, *et al.* Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study: working group on "sepsis-related problems" of the European Society of Intensive Care Medicine. *Crit Care Med* 1998;26:1793–1800.
8. Schuler A, Wulf DA, Lu Y, Iwashyna TJ, Escobar GJ, Shah NH, *et al.* The impact of acute organ dysfunction on long-term survival in sepsis. *Crit Care Med* 2018;46:843–849.
9. 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.
10. Raith EP, Udy AA, Bailey M, McGloughlin S, MacIsaac C, Bellomo R, *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.
11. Ferreira FL, Bota DP, Bross A, Mélot C, Vincent JL. Serial evaluation of the SOFA score to predict outcome in critically ill patients. *JAMA* 2001;286:1754–1758.
12. Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, *et al.* Assessment of clinical criteria for sepsis: for the third international consensus definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* 2016;315:762–774.
13. Eneanya ND, Yang W, Reese PP. Reconsidering the consequences of using race to estimate kidney function. *JAMA* 2019;322:113–114.
14. Powe NR. Black kidney function matters: use or misuse of race? *JAMA* 2020;324:737–738.
15. Cleveland Manchanda E, Couillard C, Sivashanker K. Inequity in crisis standards of care. *N Engl J Med* 2020;383:e16.
16. Levey AS, Titan SM, Powe NR, Coresh J, Inker LA. Kidney disease, race, and GFR estimation. *Clin J Am Soc Nephrol* 2020;15:1203–1212.
17. Yudell M, Roberts D, DeSalle R, Tishkoff S. Science and society: taking race out of human genetics. *Science* 2016;351:564–565.
18. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight - reconsidering the use of race correction in clinical algorithms. *N Engl J Med* 2020;383:874–882.
19. Eneanya ND, Kostelanetz S, Mendu ML. Race-free biomarkers to quantify kidney function: health equity lessons learned from population-based research. *Am J Kidney Dis* [online ahead of print] 11 Feb 2021; DOI: 10.1053/j.ajkd.2020.12.001.
20. Barnato AE, Chang C-CH, Lave JR, Angus DC. The paradox of end-of-life hospital treatment intensity among black patients: a retrospective cohort study. *J Palliat Med* 2017;21:69–77.
21. Barnato AE, Chang C-CH, Saynina O, Garber AM. Influence of race on inpatient treatment intensity at the end of life. *J Gen Intern Med* 2007;22:338–345.
22. Johnson RW, Newby LK, Granger CB, Cook WA, Peterson ED, Echols M, *et al.* Differences in level of care at the end of life according to race. *Am J Crit Care* 2010;19:335–343; quiz 344.
23. Barnato AE, Berhane Z, Weissfeld LA, Chang CC, Linde-Zwirble WT, Angus DC; Robert Wood Johnson Foundation ICU End-of-Life Peer Group. Racial variation in end-of-life intensive care use: a race or hospital effect? *Health Serv Res* 2006;41:2219–2237.
24. Escobar GJ, Greene JD, Scheirer P, Gardner MN, Draper D, Kipnis P. Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. *Med Care* 2008;46:232–239.
25. Pimentel MA, Redfern OC, Malycha J, Meredith P, Prytherch D, Briggs J, *et al.* Detecting deteriorating patients in hospital: development and validation of a novel scoring system. *Am J Respir Crit Care Med* [online ahead of print] 1 Feb 2021; DOI: 10.1164/rccm.202007-2700OC.
26. White DB, Halpern SD. Allocation of scarce critical care resources during a public health emergency. Pittsburgh, PA: University of Pittsburgh; 2020 [accessed 2021 Jul 7]. Available from: https://ccm.pitt.edu/sites/default/files/UnivPittsburgh_ModelHospitalResourcePolicy_2020_04_15.pdf.
27. Bhatraju PK, Ghassemieh BJ, Nichols M, Kim R, Jerome KR, Nalla AK, *et al.* COVID-19 in critically ill patients in the Seattle region - case series. *N Engl J Med* 2020;382:2012–2022.
28. Grasselli G, Pesenti A, Cecconi M. Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: early experience and forecast during an emergency response. *JAMA* 2020;323:1545–1546.
29. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; Iniciativa STROBE. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies [in Spanish]. *Rev Esp Salud Publica* 2008;82:251–259.
30. Anesi GL, Chowdhury M, Small DS, Delgado MK, Kohn R, Bayes B, *et al.* Association of a novel index of hospital capacity strain with admission to intensive care units. *Ann Am Thorac Soc* 2020;17:1440–1447.
31. Anesi GL, Chelluri J, Qasim ZA, Chowdhury M, Kohn R, Weissman GE, *et al.* Association of an emergency department-embedded critical care unit with hospital outcomes and intensive care unit use. *Ann Am Thorac Soc* 2020;17:1599–1609.
32. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, *et al.* The third international consensus definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* 2016;315:801–810.
33. Auriemma CL, Molinero AM, Houtrow AJ, Persad G, White DB, Halpern SD. Eliminating categorical exclusion criteria in crisis standards of care frameworks. *Am J Bioeth* 2020;20:28–36.
34. Oxman D. The crisis in crisis standards of care. *Ann Am Thorac Soc* [online head of print] 5 Feb 2021; DOI: 10.1513/AnnalsATS.202012-1527VP.
35. Jones JM, Fingar KR, Miller MA, Coffey R, Barrett M, Flottemesch T, *et al.* Racial disparities in sepsis-related in-hospital mortality: using a broad case capture method and multivariate controls for clinical and hospital variables, 2004-2013. *Crit Care Med* 2017;45:e1209–e1217.
36. Barnato AE, Alexander SL, Linde-Zwirble WT, Angus DC. Racial variation in the incidence, care, and outcomes of severe sepsis: analysis of population, patient, and hospital characteristics. *Am J Respir Crit Care Med* 2008;177:279–284.
37. Leisman DE, Harhay MO, Lederer DJ, Abramson M, Adjei AA, Bakker J, *et al.* Development and reporting of prediction models: guidance for authors from editors of respiratory, sleep, and critical care journals. *Crit Care Med* 2020;48:623–633.
38. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol* 2010;5:1315–1316.
39. Cleves MA. From the help desk: comparing areas under receiver operating characteristic curves from two or more probit or logit models. *Stata J* 2002;2:301–313.
40. 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.
41. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Stat Med* 2019;38:4051–4065.
42. Escobar GJ, Gardner MN, Greene JD, Draper D, Kipnis P. Risk-adjusting hospital mortality using a comprehensive electronic record in an integrated health care delivery system. *Med Care* 2013;51:446–453.
43. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF III, Feldman HI, *et al.*; CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration). A new equation to estimate glomerular filtration rate. *Ann Intern Med* 2009;150:604–612.
44. Grubbs V. Precision in GFR reporting: let's stop playing the race card. *Clin J Am Soc Nephrol* 2020;15:1201–1202.
45. Lagu T, Pekow PS, Shieh MS, Stefan M, Pack QR, Kashef MA, *et al.* Validation and comparison of seven mortality prediction models for hospitalized patients with acute decompensated heart failure. *Circ Heart Fail* 2016;9:e002912.
46. White DB, Lo B. Mitigating inequities and saving lives with ICU triage during the COVID-19 pandemic. *Am J Respir Crit Care Med* 2021;203:287–295.
47. Escobar GJ, Adams AS, Liu VX, Soltesz L, Chen YI, Parodi SM, *et al.* Racial disparities in COVID-19 testing and outcomes : retrospective cohort study in an integrated health system. *Ann Intern Med* [online ahead of print] 9 Feb 2021; DOI: 10.7326/M20-6979.