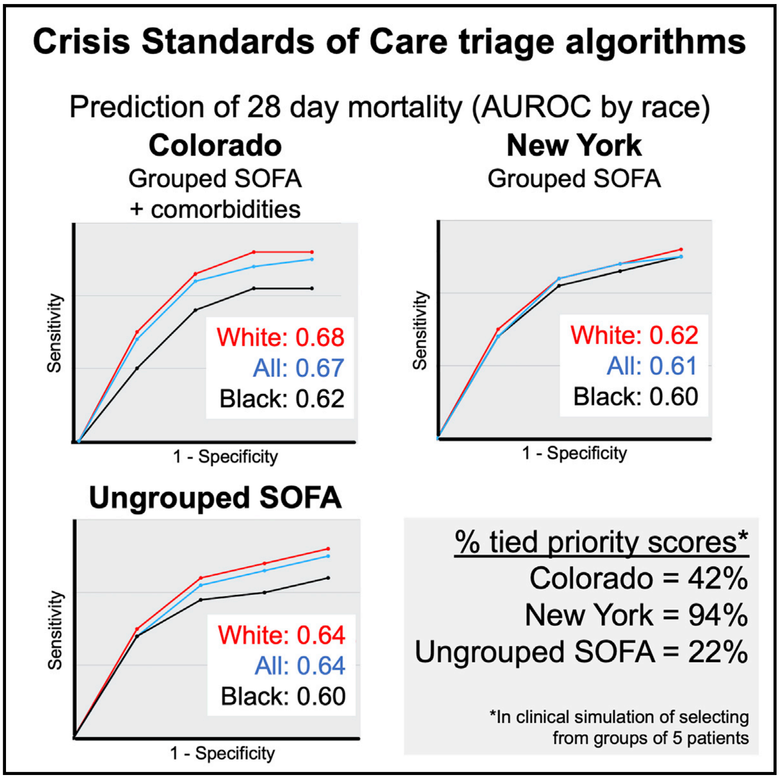
Cell Reports
Medicine

Report

Performance of crisis standards of care guidelines in a cohort of critically ill COVID-19 patients in the United States

Graphical abstract



Authors

Julia L. Jezmir, Maheetha Bharadwaj, Alexander Chaitoff, ..., William B. Feldman, Edy Y. Kim, the STOP-COVID Investigators

Correspondence

wbfeldman@bwh.harvard.edu (W.B.F.), ekim11@bwh.harvard.edu (E.Y.K.)

In brief

Jezmir et al. show that crisis standards of care (CSC) guidelines, used to allocate scarce medical resources, poorly discriminate 28-day mortality and result in frequently tied priority scores. The authors present a framework for testing CSC guidelines to ensure they meet their stated ethical goals.

Highlights

- Crisis standards of care (CSC) guidelines have poor prediction of 28-day mortality
- Consideration of comorbidities modestly improves guideline performance
- Simulation of clinical decision-making shows frequent ties in priority scores
- Using comorbidities in CSC has the potential to exacerbate racial inequities

Jezmir et al., 2021, Cell Reports Medicine 2, 100376
September 21, 2021 © 2021 The Author(s).
https://doi.org/10.1016/j.xcrm.2021.100376

Cell Reports Medicine

CellPress
OPEN ACCESS

Report

Performance of crisis standards of care guidelines in a cohort of critically ill COVID-19 patients in the United States

Julia L. Jezmir,[1,2,15] Maheetha Bharadwaj,[2,15] Alexander Chaitoff,[1,2] Bradford Diephuis,[1,2] Conor P. Crowley,[3] Sandeep P. Kishore,[1,2,4] Eric Goralnick,[2,5] Louis T. Merriam,[3] Aimee Milliken,[2,6] Chanu Rhee,[7,8] Nicholas Sadovnikoff,[2,9,13] Sejal B. Shah,[2,10] Shruti Gupta,[2,11] David E. Leaf,[2,11,14] William B. Feldman,[2,3,12,14,*] Edy Y. Kim,[2,3,14,16,*] and the STOP-COVID Investigators

[1]Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA
[2]Harvard Medical School, Boston, MA 02115, USA
[3]Division of Pulmonary and Critical Care Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA
[4]Department of Global Health & Health System Design, Icahn School of Medicine at Mount Sinai, New York NY 10029, USA
[5]Department of Emergency Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA
[6]Ethics Service, Brigham and Women's Hospital, Boston, MA 02115, USA
[7]Division of Infectious Diseases, Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA
[8]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA 02215, USA
[9]Department of Anesthesiology, Perioperative Medicine and Pain Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA
[10]Department of Psychiatry, Brigham and Women's Hospital, Boston, MA 02115, USA
[11]Division of Renal Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA
[12]Program On Regulation, Therapeutics, And Law (PORTAL), Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA
[13]Center for Bioethics, Harvard Medical School, Boston, MA 02115, USA
[14]These authors contributed equally
[15]These authors contributed equally
[16]Lead contact
*Correspondence: wbfeldman@bwh.harvard.edu (W.B.F.), ekim11@bwh.harvard.edu (E.Y.K.)

https://doi.org/10.1016/j.xcrm.2021.100376

SUMMARY

Many US states published crisis standards of care (CSC) guidelines for allocating scarce critical care resources during the COVID-19 pandemic. However, the performance of these guidelines in maximizing their population benefit has not been well tested. In 2,272 adults with COVID-19 requiring mechanical ventilation drawn from the Study of the Treatment and Outcomes in Critically Ill Patients with COVID-19 (STOP-COVID) multicenter cohort, we test the following three approaches to CSC algorithms: Sequential Organ Failure Assessment (SOFA) scores grouped into ranges, SOFA score ranges plus comorbidities, and a hypothetical approach using raw SOFA scores not grouped into ranges. We find that area under receiver operating characteristic (AUROC) curves for all three algorithms demonstrate only modest discrimination for 28-day mortality. Adding comorbidity scoring modestly improves algorithm performance over SOFA scores alone. The algorithm incorporating comorbidities has modestly worse predictive performance for Black compared to white patients. CSC algorithms should be empirically examined to refine approaches to the allocation of scarce resources during pandemics and to avoid potential exacerbation of racial inequities.

INTRODUCTION

During the COVID-19 pandemic, more than 30 US states developed crisis standards of care (CSC) guidelines.[2–4] These guidelines are designed to help hospitals allocate resources, such as ventilators, if they became scarce.[5] Unlike the "all-come, all-served" promise of hospital resources during non-crisis situations, CSC guidelines aim to maximize the population-wide benefit.[6,7] State guidelines generally describe ethical principles and outline triage algorithms for resource allocation.[2,3]

To maximize the population-wide benefit, these algorithms aim to identify patients most likely to survive if offered scarce resources. Nearly 90% of states with CSC triage algorithms adapted Sequential Organ Failure Assessment (SOFA) or Modified SOFA (MSOFA) scores to predict short-term prognosis (i.e., survival to hospital discharge) in an effort to maximize the number of lives saved.[2,3,8–11] States vary in their use SOFA/MSOFA scores—for example, by grouping scores in different ranges to assign priority points or by modifying scoring calculations.[2,3] Approximately 70% of states also incorporate measures of

Table 1. CSC algorithms

Algorithm component	New York model[a]	Modified Colorado model	Raw SOFA score model
SOFA priority points	SOFA < 7: 1 point SOFA 8–11: 2 points SOFA > 11: 3 points	SOFA < 6: 1 point SOFA 6-9: 2 points SOFA 10–12: 3 points SOFA > 12: 4 points	SOFA 1: 1 point SOFA 2: 2 points SOFA 3: 3 points SOFA score = priority points
Comorbidities priority points	None	Modified Charlson Comorbidity Index[b]	None
Priority score calculation	SOFA score	SOFA prioritization + Charlson Comorbidity Index Score	SOFA score
Priority grouping based on priority score	High priority: 1 Intermediate priority: 2 Low priority: 3	None	None
Tie breakers	1st tie breaker: children 2nd tie breaker: lottery	1st tie breaker: children, health care workers, and/or first responders 2nd tie breaker: life cycle (age)[c], pregnancy, and/or sole caretakers for elderly 3rd tie breaker: lottery	1st tie breaker: age 2nd tie breaker: lottery

[a] The New York Algorithm exclusion criteria include the following: (1) unwitnessed cardiac arrest, recurrent arrest without hemodynamic stability, arrest unresponsive to standard interventions and measures; trauma-related arrest; (2) irreversible age-specific hypotension unresponsive to fluid resuscitation and vasopressor therapy; (3) traumatic brain injury with no motor response to painful stimulus (i.e., best motor response = 1); (4) severe burns where predicted survival is ≤10% even with unlimited aggressive therapy; and (5) any other conditions resulting in immediate or near-immediate mortality even with aggressive therapy. None of the patients in this cohort fell into this exclusion criteria.

[b] Original and modified Comorbidity Index provided in Table S3.

[c] Life cycle groupings (age, years) for Colorado also used for Raw SOFA model: 0–49 = 1 (highest priority), 50–59 = 2, 60–69 = 3, 70–79 = 4, and 80+ = 5 (lowest priority).

comorbidities in priority scoring, which may affect both short- and long-term prognosis.[2,3] Additionally, some states use a multicomponent model that incorporates factors such as estimated survival or duration of benefit or need in addition to SOFA scores.[2]

Although ethicists have debated triage approaches, CSC algorithms have had limited empirical testing in the COVID-19 pandemic. In prior studies of non-COVID-19 ICU cohorts, different CSC algorithms yield different prioritization results.[12] In a recent study of 675 critically ill patients with COVID-19, raw SOFA scores alone calculated at the time of intubation had limited ability to predict mortality.[13] Given the poor discrimination by SOFA scores, we hypothesized that algorithms that incorporate comorbidities, in addition to SOFA scores, have superior discriminant ability compared to algorithms that use SOFA scores alone. Because comorbidities are associated with race and ethnicity,[14–19] we also hypothesized that incorporating comorbidities could alter the performance of CSC algorithms by race and ethnicity.

In a multicenter cohort study of critically ill patients with COVID-19 admitted to ICUs across the United States, we evaluated two representative CSC guidelines—New York's algorithm,[20] which relies exclusively on SOFA scores grouped into ranges, and a modified version of Colorado's algorithm,[21] which relies on SOFA score groupings plus comorbidities (Table 1). We tested the performance of these representative CSC algorithms in discriminating 28-day in-hospital mortality and in simulated clinical scenarios in which the algorithm selected one patient among a group of two to five patients. We focused on these two state guidelines because they represent two ends of the spectrum in considering comorbidities, with New York excluding consideration of comorbidities altogether and Colorado incorporating a broad range of pre-existing conditions. Most, if not all, state algorithms take one of these approaches, with many state algorithms using a narrower range of comorbidities than Colorado.[2–4] We also tested a hypothetical algorithm of raw SOFA scores (not grouped into ranges) to assess the impact of the SOFA score ranges used by most states.

RESULTS

CSC algorithms poorly discriminate 28-day mortality

We analyzed 2,722 patients who were intubated on the first day of ICU admission in the Study of the Treatment and Outcomes in Critically Ill Patients with COVID-19 (STOP-COVID) (Figure S1), a multicenter cohort study of adult patients at 68 hospitals across the United States (Table S1).[1] The mean age (SD) was 61 (14) (Table 2); 1,475 patients (65%) were male, 797 (35%) were female, 601 (26%) were Black, and 867 (38%) were white. A total of 1,073 (47%) patients died within 28 days of ICU admission.

CSC algorithms assign "priority points" to estimate the likelihood of survival. Patients with fewer priority points are estimated to have a greater chance of survival, so these patients with lower priority scores are offered scarce resources. Nearly all verifiable CSC algorithms use the SOFA score,[2,3,8–11] a metric that assesses the level of dysfunction of six organ systems at the time of calculation.[22] For our analysis, we adapted the SOFA score calculation[10] to accommodate the data available in the

Table 2. Population characteristics

Patient characteristics	All patients (N = 2,272)	White patients (n = 867)	Black patients (n = 603)	p value
Age, mean (SD), year	61.4 ± 14.1	62.8 ± 13.9	62.1 ± 13.1	0.2847
Male, n (%)	1,475 (64.9)	579 (66.8)	340 (56.4)	<0.001
Self-reported ethnicity, no. (%)				
Hispanic/Latino	588 (25.8)	260 (30.0)	13 (2.2)	<0.001
Non-Hispanic/non-Latino	1,368 (60.2)	555 (64.0)	551 (91.4)	
Not known	298 (13.1)	52 (5.9)	13 (2.1)	
Self-reported race, no. (%)				
White	867 (38.1)		Not applicable	
Black	601 (26.4)			
Asian	144 (6.3)			
American Indian/Alaska Native	11 (0.5)			
Native Hawaiian/Other Pacific Islander	15 (0.7)			
More than one race	28 (1.2)			
Unknown/unspecified	605 (26.6)			
SOFA scores, mean ± SD[a]				
New York priority group	1.4 ± 0.6	1.4 ± 0.5	1.6 ± 0.6	<0.001
Raw SOFA score	6.9 ± 2.7	6.6 ± 2.4	7.1 ± 2.7	<0.001
Colorado priority group	3.3 ± 1.2	3.3 ± 1.1	3.5 ± 1.2	<0.001
SOFA score componentsa, mean (SD)[b]				
Respiratory	3.0 (0.99)	2.96 (0.99)	3.02 (1.00)	0.2374
Coagulation	0.22 (0.54)	0.24 (0.58)	0.23 (0.54)	0.6619
Liver	0.16 (0.48)	0.14 (0.44)	0.17 (0.54)	0.3050
Cardiovascular	2.26 (1.54)	2.28 (1.53)	2.19 (1.57)	0.3038
Central nervous system	0.30 (0.46)	0.33 (0.47)	0.31 (0.46)	0.2251
Renal	0.95 (1.34)	0.75 (1.21)	1.44 (1.49)	<0.001
Comorbidities, n (%)[c]				
Congestive heart failure	204 (8.9)	76 (8.8)	83 (13.8)	0.003
Chronic pulmonary disease[d]	489 (21.5)	209 (24.1)	158 (26.2)	0.3942
Chronic renal disease[e]	407 (17.9)	120 (13.8)	134 (22.2)	<0.001
End-stage renal disease	75 (3.3)	26 (3.0)	31 (5.1)	0.051
Active malignancy	101 (4.4)	58 (6.7)	22 (3.7)	0.015
Diabetes with complications	357 (15.7)	109 (12.6)	132 (21.9)	<0.001
Chronic liver disease	71 (3.1)	28 (3.2)	16 (2.7)	0.6298
Death, n (%)	1,073 (47.2)	407 (46.9)	286 (47.5)	0.8494

[a] SOFA score components are rated on a scale of 0–4. Table S2 provides definitions for each component used in this paper.

[b] A total of 594 patients were excluded from the analysis if any of the components of the SOFA score were missing other than the cardiovascular component.

[c] Table S2 provides for definitions of comorbidities and highlights differences in comorbidities between the full Colorado and the modified Colorado model used in this paper.

[d] Chronic pulmonary disease as defined by chronic obstructive pulmonary disease or asthma.

[e] Chronic renal disease as defined by chronic kidney disease (estimated glomerular filtration rate (eGFR) < 60 on at least 2 consecutive values at least 12 weeks apart) or end-stage renal disease.

STOP-COVID database, modifying the cardiovascular and central nervous system components (Table S2). Most CSC algorithms do not assign priority points based on the raw SOFA score but rather assign priority points to SOFA scores grouped into ranges.[2,3] For example, New York's algorithm assigns 1 point to patients with SOFA scores of <7, 2 points for SOFA scores of 8–11, and 3 points for scores >11. When two patients receive the same number of priority points, New York uses a lottery to break the tie (Table 1).[20]

In contrast to New York, Colorado's algorithm assigns points to both SOFA score ranges and comorbidities.[21] Patients receive 1 priority point for SOFA scores of <6, 2 points for scores 6–9, 3 points for scores 10–12, and 4 points for scores of >12. Additional priority points are based on the Charlson Comorbidities

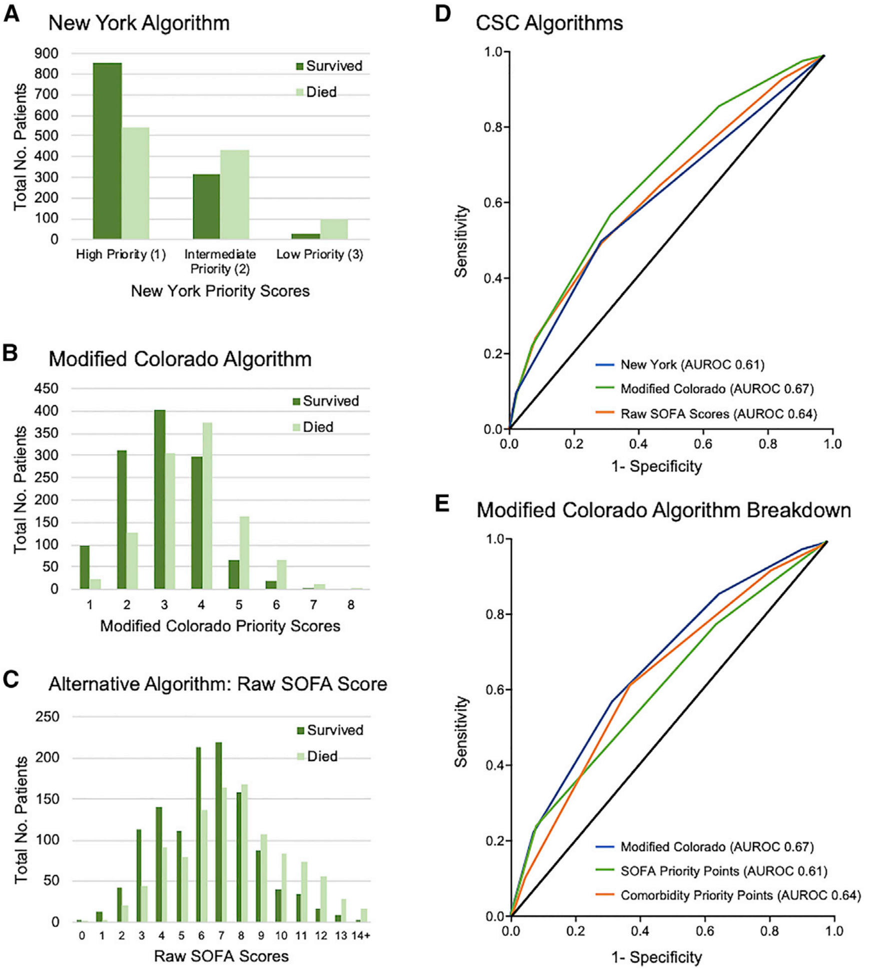


Figure 1. Association of priority score or category with 28-day mortality

(A–C) The number of patients who survived or died at 28 days after ICU admission and intubation are shown for each priority point value (or category) for each algorithm. (A) New York (SOFA score groups only). (B) Colorado (SOFA score groups and comorbidities). (C) Hypothetical algorithm of raw (ungrouped) SOFA score.

(D and E) AUROC curves for discrimination of 28-day mortality by priority scores are shown for the following algorithms: New York (SOFA score groups) (D), Colorado (SOFA score groups and comorbidities) (D), and raw SOFA scores. (E) Colorado SOFA score component, Colorado comorbidity scoring component, or full Colorado algorithm (SOFA and comorbidities).

Index.[23] Because we adapted the Charlson Comorbidity Index to available data (Table S3), we refer to a "modified" Colorado model. When two patients receive the same number of priority points, Colorado first prioritizes children, healthcare workers, and first responders. If still tied, Colorado prioritizes younger patients, pregnant patients, and caretakers for the elderly. If still tied, Colorado uses a lottery. As a third approach, we used a hypothetical algorithm of raw SOFA scores that are not collapsed into ranges (Table 1).

The primary outcome was 28-day in-hospital mortality, which we assessed for each patient subcohort defined by their CSC priority score at the time of ICU admission and intubation (Figures 1A–1C). All algorithms have an increased fraction of surviving patients in the "better" priority categories (i.e., lower priority point total) that would be prioritized for scarce resources. We next assessed the accuracy of each CSC algorithm in discriminating 28-day in-hospital mortality by the area under the receiver operating characteristic (AUROC) curve. AUROC was 0.61 (95% confidence interval [CI], 0.59–0.63) for New York (i.e., SOFA score ranges), 0.67 (95% CI, 0.65–0.69) for Colorado (i.e., SOFA score ranges and comorbidities), and 0.64 (95% CI, 0.62–0.66) for raw SOFA scores ($p < 0.001$; Figure 1D). In a sensitivity analysis, we imputed missing SOFA score components with normal (i.e., healthy) values to assess our exclusion of patients with these components missing in their clinical record. We included an additional 594 patients and found similar trends in AUROC, with 0.59 (95% CI, 0.57–0.61) for New York, 0.65 (95% CI, 0.64–0.67) for Colorado, and 0.61 (95% CI, 0.59–0.63) for raw SOFA scores (Figure S2A).

Comorbidity scores alone modestly outperformed SOFA score ranges alone

To investigate what may drive differences in performance among CSC algorithms, we conducted sensitivity analyses examining the effect of comorbidities and the effect of how SOFA scores are grouped into ranges. First, we assessed how each components of Colorado's algorithm—SOFA scores and comorbidity scores—perform on their own in discrimination of 28-day mortality (Figure 1E). The SOFA score component of Colorado's algorithm yielded an AUROC (95% CI) of 0.61 (0.59–0.63), and the comorbidity component alone yielded an AUROC of 0.64 (0.62–0.66) with $p < 0.001$, compared to 0.67 (0.65-0.69) for Colorado's complete algorithm with both SOFA and comorbidity components.

The elements of SOFA and comorbidities are not independent. For example, in Colorado's algorithm, chronic kidney disease (CKD) might be "counted twice" for some patients as a comorbidity in the Charlson Comorbidity Index and as a marker of organ failure in the SOFA score. In a sensitivity analysis, we excluded renal disease from the Charlson Comorbidity Index for the 332 patients with CKD or end-stage renal disease (ESRD). For 28-day mortality, the AUROC (95% CI) was 0.62 (0.55–0.68) if CKD/ESRD were excluded from comorbidity scoring, compared to 0.61 (0.54–0.67) with CKD/ESRD included

Table 3. CSC Algorithm performance in small group comparisons[a]

	(A) Decisions not requiring lottery tie-breaker		(B) Correct selections among decisions not requiring lottery		(C) Overall performance for correct selections	
Algorithm	%	95% CI	%	95% CI	%	95% CI
Groups of two						
New York[b]	52	48–55	72	66–77	61	57–65
Colorado	77	74–82	72	68–76	67	63–71
Raw SOFA	89	87–92	65	62–70	64	60–68
Algorithms + age as tie-breaker						
New York + age	90	87–93	70	66–75	68	65–72
Colorado + age[b]	93	91–96	69	65–73	68	65–72
Raw SOFA + age	98	97–99	66	62–70	66	62–69
Groups of five						
New York[b]	6	5–7	64	51–75	61	58–63
Colorado	58	56–61	74	70–77	70	67–72
Raw SOFA	78	76–81	66	63–69	64	62–67
Algorithms + age as tie-breaker						
New York + age	68	65–71	73	69–76	71	69–74
Colorado + age[b]	83	80–85	72	69–75	71	68–74
Raw SOFA + age	95	93–96	67	64–70	66	63–69

[a] Triage decisions by CSC algorithms in a simulation of 1,000 random groups of two or five patients. Column A, i.e., two or more patients not tied for the "best" (lowest) priority score. Column B, i.e., survival. Column C, i.e., selecting a surviving patient across all decisions (i.e., all decisions regardless whether selected by priority score or requiring a tie-breaking lottery. Unpaired t tests were conducted to compare all algorithms (with and without age as a tie-breaker) to each other. Nearly all comparisons were significant at $p < 0.01$. The only non-significant comparisons were New York versus Colorado for groups of two in column B, New York + age versus Colorado + age for groups of two and groups of five in Column B, and New York + age versus Colorado + age for groups of two and groups of five in Column C.

[b] Indicates the algorithm that is closest state guidelines. New York's algorithm as written in the state guidelines does not use a tie-breaker, whereas Colorado's algorithm does.

(Figure S2B). If CKD/ESRD were excluded from the scoring of comorbidities, the AUROC for the entire cohort of 2,272 patients was 0.67 (0.65–0.69), which was unchanged from the original analysis.

New York's and Colorado's algorithms differ in both their approach to comorbidities and how they group SOFA score into ranges. Grouping SOFA scores into ranges is a common feature of state CSC algorithms, but the effect of grouping schemes on performance has not been studied empirically. We performed a sensitivity analysis to assess the effect of SOFA score grouping schemes on algorithm performance, for example, for SOFA scores in increments of two, with SOFA scores of 1 or 2 receiving a priority score of 1; SOFA scores of 3 or 4 receiving a priority score of 2, and so forth. The predictive accuracy of SOFA scores as insensitive to grouping in ranges of 1, 2, or 3. For SOFA score increments of 1 (i.e., the ungrouped, raw SOFA algorithm), the AUROC was 0.61 (0.59–0.63); for increments of 2, 0.63 (0.61–0.65); and for increments of 3, 0.62 (0.60–0.64) (Figure S2C).

Age tie-breakers improve algorithm performance in small group comparisons

To examine how prioritization algorithms may function in clinical scenarios, we simulated selecting one patient to receive scarce resources out of groups of two or five patients by using a bootstrap method. For each group of patients, a "winner" with the "best" priority score (i.e., lowest priority point total) was selected, and the winner's 28-day outcome (survivor or deceased) was noted. The group was considered tied if two or more patients tied for the best (lowest) priority point total. We performed 100 iterations of a computational simulation in which we randomly selected 1,000 groups of 2 or 5 patients. We excluded patient groups in which all the patients had the same outcome (i.e., all survivors or all deceased) because we cannot assess if the algorithm correctly selects a patient with a better outcome if all the patients in that group shared the same outcome. For each simulation of 1,000 patient groups, we calculated the percentage of groups for which the algorithm chose a patient who survived, and we computed the percentage of groups in which the algorithm required a tie-breaking lottery. These simulations yielded distributions for the percentage of algorithm decisions that selected a survivor or required a lottery. The results suggested that algorithms struggle with selecting one patient from a larger group, as all algorithms had worse performance in the groups of five patients than that of groups of two patients. First, we examined the frequency of patient groups with tied priority scores that required a tie-breaker, such as a lottery (Table 3A). New York selected a patient without a lottery tie-breaker in 51% (95% CI, 47–55) of patient groups of 2 but selected a patient without a lottery tie-breaker in only 6% (4–7) of patient groups of 5. That is, when selecting among a group of five patients, New York is almost a pure lottery, as 94% of the groups have tied priority scores requiring a lottery. For Colorado, the percentage of decisions made without requiring a tie-breaking lottery was 77% (95% CI, 74–80) and 58% (56–61) for patient groups of 2 and 5, respectively. For our raw SOFA algorithm, the percentage of decisions made without lottery was 89% (95% CI, 87–92) and 78% (76–81) for groups of 2 and 5, respectively (Table 3A).

In this simulation, we further examined those decisions that did not require a lottery tie-breaker. Among decisions not requiring a tie-breaker, we assessed whether the algorithm made the "correct" choice by prioritizing a patient with the better outcome (i.e., survival) (Table 3B). New York chose a patient who survived in 72% (95% CI, 66–77) of patient groups of 2 and 64% (51–75) of groups of 5. Colorado selected a surviving patient in 72% (95% CI, 68–76) and 74% (70–77) of decisions, for patient groups of 2 and 5, respectively. The raw SOFA algorithm selected a surviving patient in 65% (62–70) and 66% (63–69) of decisions, for patient groups of 2 and 5, respectively. Thus, for groups of five patients, the algorithm incorporating comorbidities (Colorado) had superior accuracy in selecting a patient with the better outcome than the algorithm (New York) that only considered SOFA score ranges. We next calculated the percentage of correct decisions (i.e., selecting a surviving patient) across all decisions, whether the patient was selected by priority score or required a tie-breaking lottery (Table 3C). In patient groups of 2, Colorado (67% correct decisions) had better overall performance than New York (61% correct) because although Colorado and New York shared the same accuracy among non-lottery decisions, Colorado had fewer decisions go to lottery, which is only 50% by chance (Table 3C). For patient groups of 5, Colorado (70% correct) continued to outperform New York (61% correct) with a combination of fewer lotteries and better accuracy in non-lottery decisions.

Our simulations suggested that tie-breakers would be important and frequently used in clinical practice. Although no states use age as categorical exclusion criteria, many initial guidelines included age categories as tie-breaker criteria. Many states have since moved away from specifying age as the main tie-breaker based on US Department of Health and Human Services guidance, instead considering age as part of individualized assessments when addressing tie-breakers.[24] We applied the tie-breaker based on age categories in Colorado's guidelines to all three algorithms. Adding age as a tie-breaker improved algorithm performance (Table 3). With age as the tie-breaker, New York selected a patient without a lottery in 90% (95% CI, 87–93) of decisions in patient groups of 2 and 68% (65–71) of decisions in groups of 5, compared to 51% and 6%, respectively, without an age-based tiebreaker (Table 3A). Of the decisions that did not require a lottery, New York chose the surviving patient in 70% (66–75) of pairs and 73% (69–76) of the groups of 5, compared to 72% and 64%, respectively, without an age-based tie-breaker (Table 3B). With an age tie-breaker, Colorado selected a patient without lottery in 93% (95% CI, 91–96) and 83% (80–85) for patient groups of 2 and 5, respectively (increased from 77% and 58% without age tie-breakers) (Table 3A). In non-lottery decisions, Colorado had similar accuracy in choosing the surviving patient with or without age tie-breakers (Table 3B). Similar trends as those for Colorado were seen for the hypothetical raw SOFA score algorithm. For New York, the addition of age tie-breakers improved overall performance in selecting the correct patient from 61% to 71% (for patient groups of 5; Table 3C). For Colorado and raw SOFA, the overall performance was minimally improved by only 1%–2%. In summary, the key effect of an age tie-breaker is to increase the percentage of decisions made without lottery without altering the percentage of correct choices.

Incorporating comorbidities has the potential to worsen algorithm performance for Black patients, compared to white patients

To examine the differences in algorithm performance by self-reported race and ethnicity, 1,468 patients were included in the analysis, with 867 (59%) white and 601 (41%) Black patients. For ethnicity, 1,956 patients were included, with 588 (30%) Hispanic/Latino and 1,368 non-Hispanic/Latino (70%). For Colorado, the AUROC curve for predicting 28-day mortality was 0.62 (95% CI, 0.57–0.66) for Black patients and 0.68 (0.65–0.72) for white patients ($p < 0.03$). There were no statistically significant differences in AUROC curves for Colorado by ethnicity or New York by race or ethnicity (Figure S3). To assess the clinical meaning of Colorado's modest differences in performance, we turned to our simulation of selecting a patient out of a small group. Colorado's algorithm consistently performed better for the subcohort of white patients than for Black patients. In the simulation of selecting 1 patient from a group of 5 patients, Colorado selected the patient with the better outcome in 71% of groups in the white subcohort but only 63% in the Black subcohort ($p < 0.01$) (Table S4C). In contrast, New York selected the patient with the better outcome at similar rates for white and Black subcohorts (61% and 60%, respectively). Similar results were seen in groups of two and with an age tie-breaker. The hypothetical raw SOFA algorithm selected the correct patient more frequently for the white compared to the Black subcohort (65% and 60%, respectively). New York's SOFA ranges reduced the race-dependent effects seen with raw (ungrouped) SOFA scores. However, even though Colorado's SOFA score groupings resemble those of New York, Colorado had race-dependent differences in algorithm performance that exceeded what was seen with raw SOFA scores. These results suggest that including comorbidities has the potential to worsen the performance of the algorithm for Black, compared to white, patients. Further study into differences based on race in algorithm performance is necessary to better understand the factors that may be contributing to the difference in performance.

DISCUSSION

In this multicenter, nationally representative cohort study of critically ill patients with COVID-19, we found that both New York's (SOFA score groups) and Colorado's (SOFA score groups and comorbidities) algorithms had modest accuracy in discriminating 28-day mortality. In Colorado's algorithm, the addition of comorbidities modestly improved performance over the SOFA score alone. CSC algorithms greatly varied in the frequency of ties, which ranged from 11% to 94% depending on the scenario and algorithm (Table 3). Frequent ties are not necessarily a positive or negative feature. Some ethicists deem lotteries as the "most fair" method. However, when a state is selecting a CSC algorithm, the frequency of tied priority scores is an important performance characteristic to assess.

Our results for the SOFA score component alone fall within the broad range seen in studies of SOFA scores for triage of critically ill patients without COVID-19. Prior studies had AUROC curves of 0.55–0.88 for the prediction of outcomes by SOFA scores.[3,25–29] Differences among studies were likely driven both by cohort characteristics and the design of triage algorithms. However, states have little empirical guidance on how to design algorithms. For example, a key design decision is how to group SOFA scores in ranges, if at all. In our sensitivity analysis, SOFA score grouping had little difference on predictive accuracy for 28-day mortality, although using raw (ungrouped) SOFA scores reduced ties. Further study is needed to better understand grouping strategies in patient cohorts with different distributions of SOFA scores and outcomes. We also found that the average SOFA scores for both New York's and Colorado's algorithms fall within the highest priority category (indicating the highest likelihood of survival), which may contribute to the modest performance of algorithms in discriminating outcomes. Our study and prior literature raise the question of whether the use of SOFA scores in CSC guidelines should be reconsidered.[10,25,27] Possible substitutes for SOFA scores may include blood laboratory values associated with outcomes in COVID-19 like C-reactive protein (CRP) or lactate levels.[30–32] A challenge is finding metrics that work for a mixed population with diagnoses ranging from sepsis to respiratory failure.

If activated in this study cohort, the CSC algorithms would have denied scarce resources to many patients who would have survived and allocated resources to many who would have died. In the clinical scenario of selecting from small groups of patients, CSC algorithms correctly selected the patient with the better outcome in 65% to 74% of decisions. Whether 70% is an acceptable success rate depends on a variety of ethical and practical considerations. A state may decide against incorporating comorbidities, despite modestly worse overall performance, for simplicity or to avoid the potential for exacerbating racial disparities, although the possible relationship of race, ethnicity and CSC algorithm performance requires further study.[33–37] However, another state may include comorbidities because a modest improvement in performance may result in a meaningful number of lives saved when applied to many. It is vital that states empirically test triage algorithms to quantify whether an algorithm fulfills the ethical principle of maximizing lives saved and reaches acceptable thresholds for "real world" performance set by medical, lay, and other communities. We have offered a framework for conducting such tests to ensure that CSC algorithms achieve the ethical principles they are designed to operationalize.

LIMITATIONS OF THE STUDY

Our study has several limitations. First, the study is limited to patients with COVID-19. CSC algorithms may be more or less predictive of other diseases, thus systematically advantaging or disadvantaging those with COVID-19. Second, the study may not generalize to patients who were intubated several days after ICU admission or the less common situation of intubation greater than 24 h before ICU admission. More generally, this study did not distinguish patients who deteriorated early in their hospital course (e.g., hospital day 1) and those patients who deteriorated later in their hospital course after several days of non-critical illness. Third, due to data limitations, we used a modified version of the SOFA score, approximating two of the components (cardiovascular and central nervous system). Fourth, although CSC were not activated at our study sites, the study cannot account for how the severity of the COVID-19 pandemic and individual illnesses may have influenced the decisions of individual clinicians regarding intubation and ICU triage, nor can our study account for changes to clinical practice, such as more intensive palliative care consultation, during different phases of the COVID-19 pandemic. Fifth, it is possible that scoring systems may perform differently now than in the spring of 2020 after the introduction of new therapies and improvement in outcomes.[38] Sixth, we examined two state guidelines representing the most common elements in state algorithms, but there are differences by state that may affect performance. Finally, we did not assess outcomes beyond 28 days, although our prior study found that the vast majority of deaths that occur among critically ill patients with COVID-19 occur in the first 28 days following ICU admission.[1]

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - ○ Study Design and Population
  - ○ Data Collection
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2021.100376.

ACKNOWLEDGMENTS

W.B.F. received support from the U.S. National Institutes of Health (N.I.H.) (5T32HL007633-35). The complete author list for the STOP-COVID Investigators and the partcipating sites are included in Table S1.

AUTHOR CONTRIBUTIONS

Conceptualization and methodology, J.L.J., M.B., A.C., B.D., C.P.C., S.P.K., C.R., S.B.S., S.G., D.E.L., W.B.F., and E.Y.K.; formal analysis and investigation, J.L.J., M.B., A.C., and B.D.; data curation, J.L.J., M.B., S.B.S., S.G.,

and D.E.L.; writing - original draft, J.L.J., M.B., and A.C.; writing - review and editing, J.L.J., M.B., A.C., E.G., L.T.M., A.M., C.R., N.S., S.B.S., S.G., W.B.F., and E.Y.K.; supervision, W.B.F. and E.Y.K.

DECLARATION OF INTERESTS

The authors have no conflicts of interest relevant to this work. For unrelated work, the authors have the following disclosures: S.P.K. served as a consultant for Resolve to Save Lives and has previously led a partnership on multiple chronic conditions at the Icahn School of Medicine at Mount Sinai with Teva Pharmaceuticals. W.B.F. serves as a consultant for Alosa Health and Aetion. He also received an honorarium for a presentation to Blue Cross Blue Shield of Massachusetts. E.Y.K. is a co-investigator in NCT04389671 (Windtree Therapeutics) testing lucinactant (surfactant-like treatment) in COVID-19 patients and received unrelated research funding from Bayer AG, N.I.H., the American Heart Association, the American Lung Association, and the American Thoracic Society.

Received: March 15, 2021
Revised: June 8, 2021
Accepted: July 21, 2021
Published: July 28, 2021

REFERENCES

1. Gupta, S., Hayek, S.S., Wang, W., Chan, L., Mathews, K.S., Melamed, M.L., Brenner, S.K., Leonberg-Yoo, A., Schenck, E.J., Radbel, J., et al.; STOP-COVID Investigators (2020). Factors associated with death in critically ill patients with coronavirus disease 2019 in the US. JAMA Intern. Med. *180*, 1436–1447.
2. Piscitello, G.M., Kapania, E.M., Miller, W.D., Rojas, J.C., Siegler, M., and Parker, W.F. (2020). Variation in ventilator allocation guidelines by US state during the coronavirus disease 2019 pandemic: a systematic review. JAMA Netw. Open *3*, e2012606.
3. Hantel, A., Marron, J.M., Casey, M., Kurtz, S., Magnavita, E., and Abel, G.A. (2021). US State Government Crisis Standards of Care Guidelines: Implications for Patients With Cancer. JAMA Oncol. *7*, 199–205.
4. Romney, D., Fox, H., Carlson, S., Bachmann, D., O'Mathuna, D., and Kman, N. (2020). Allocation of Scarce Resources in a Pandemic: A Systematic Review of US State Crisis Standards of Care Documents. Disaster Med. Public Health Prep. *14*, 677–683.
5. Hertelendy, A.J., Ciottone, G.R., Mitchell, C.L., Gutberg, J., and Burkle, F.M. (2021). Crisis standards of care in a pandemic: navigating the ethical, clinical, psychological and policy-making maelstrom. Int. J. Qual. Health Care *33*, mzaa094.
6. Emanuel, E.J., Persad, G., Upshur, R., Thome, B., Parker, M., Glickman, A., Zhang, C., Boyle, C., Smith, M., and Phillips, J.P. (2020). Fair Allocation of Scarce Medical Resources in the Time of Covid-19. N. Engl. J. Med. *382*, 2049–2055.
7. White, D.B., and Lo, B. (2020). A Framework for Rationing Ventilators and Critical Care Beds During the COVID-19 Pandemic. JAMA *323*, 1773–1774.
8. Sprung, C.L., Joynt, G.M., Christian, M.D., Truog, R.D., Rello, J., and Nates, J.L. (2020). Adult ICU Triage During the Coronavirus Disease 2019 Pandemic: Who Will Live and Who Will Die? Recommendations to Improve Survival. Crit. Care Med. *48*, 1196–1202.
9. Baker, M., and Fink, S.. At the Top of the Covid-19 Curve, How Do Hospitals Decide Who Gets Treatment? *The New York Times*, March 13, 2020.https://www.nytimes.com/2020/03/31/us/coronavirus-covid-triage-rationing-ventilators.html.
10. Hayek, S.S., Brenner, S.K., Azam, T.U., Shadid, H.R., Anderson, E., Berlin, H., Pan, M., Meloche, C., Feroz, R., O'Hayer, P., and Kaakati, R. (2020). In-hospital cardiac arrest in critically ill patients with covid-19: multicenter cohort study. BMJ *30*, 371.
11. Truog, R.D., Mitchell, C., and Daley, G.Q. (2020). The toughest triage—allocating ventilators in a pandemic. N. Engl. J. Med. *382*, 1973–1975.
12. Wunsch, H., Hill, A.D., Bosch, N., Adhikari, N.K.J., Rubenfeld, G., Walkey, A., Ferreyro, B.L., Tillmann, B.W., Amaral, A.C.K.B., Scales, D.C., et al. (2020). Comparison of 2 Triage Scoring Guidelines for Allocation of Mechanical Ventilators. JAMA Netw. Open *3*, e2029250.
13. Raschke, R.A., Agarwal, S., Rangan, P., Heise, C.W., and Curry, S.C. (2021). Discriminant Accuracy of the SOFA Score for Determining the Probable Mortality of Patients With COVID-19 Pneumonia Requiring Mechanical Ventilation. JAMA *325*, 1469–1470.
14. Williams, D.R., Mohammed, S.A., Leavell, J., and Collins, C. (2010). Race, socioeconomic status, and health: complexities, ongoing challenges, and research opportunities. Ann. N Y Acad. Sci. *1186*, 69–101.
15. Chen, J., Vargas-Bustamante, A., Mortensen, K., and Ortega, A.N. (2016). Racial and ethnic disparities in health care access and utilization under the Affordable Care Act. Med. Care *54*, 140–146.
16. Hardeman, R.R., Medina, E.M., and Kozhimannil, K.B. (2016). Structural Racism and Supporting Black Lives - The Role of Health Professionals. N. Engl. J. Med. *375*, 2113–2115.
17. Spencer Bonilla, G., Rodriguez-Gutierrez, R., and Montori, V.M. (2016). What We Don't Talk About When We Talk About Preventing Type 2 Diabetes-Addressing Socioeconomic Disadvantage. JAMA Intern. Med. *176*, 1053–1054.
18. Lackland, D.T. (2014). Racial differences in hypertension: implications for high blood pressure management. Am. J. Med. Sci. *348*, 135–138.
19. Ortiz, A. (2019). Burden, access and disparities in kidney disease: chronic kidney disease hotspots and progress one step at a time. Clin. Kidney J. *12*, 157–159.
20. NY State. NY guidelines for ventilators.https://www.health.ny.gov/regulations/task_force/reports_publications/docs/ventilator_guidelines.pdf.
21. Colorado State. Colorado guidelines for scarce resources. April 4, 2020.https://www.colorado.gov/pacific/sites/default/files/Crisis%20Standards%20of%20Care%20Triage%20Standards-April%202020.pdf.
22. Vincent, J.L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C.K., Suter, P.M., and Thijs, L.G. (1996). 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. *22*, 707–710.
23. Charlson, M.E., Pompei, P., Ales, K.L., and MacKenzie, C.R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J. Chronic Dis. *40*, 373–383.
24. US Health and Human Services. Civil Rights and COVID-19 Guidelines. July 26, 2021.https://www.hhs.gov/civil-rights/for-providers/civil-rights-covid19/index.html.
25. Minne, L., Abu-Hanna, A., and de Jonge, E. (2008). Evaluation of SOFA-based models for predicting mortality in the ICU: A systematic review. Crit. Care *12*, R161.
26. Ferreira, F.L., Bota, D.P., Bross, A., Mélot, C., and Vincent, J.L. (2001). Serial evaluation of the SOFA score to predict outcome in critically ill patients. JAMA *286*, 1754–1758.
27. Martinez, A.C., Dewaswala, N., Tuarez, F.R., Pino, J., Chait, R., Chen, K., Reddy, R., Abdallah, A., Al Abbasi, B., Torres, P., et al. (2020). Validation of SOFA Score in Critically Ill Patients with COVID-19. Chest *158*, A613.
28. Khan, Z., Hulme, J., and Sherwood, N. (2009). An assessment of the validity of SOFA score based triage in H1N1 critically ill patients during an influenza pandemic. Anaesthesia *64*, 1283–1288.
29. Christian, M.D., Hamielec, C., Lazar, N.M., Wax, R.S., Griffith, L., Herridge, M.S., Lee, D., and Cook, D.J. (2009). A retrospective cohort pilot study to evaluate a triage tool for use in a pandemic. Crit. Care *13*, R170.
30. Izcovich, A., Ragusa, M.A., Tortosa, F., Lavena Marzio, M.A., Agnoletti, C., Bengolea, A., Ceirano, A., Espinosa, F., Saavedra, E., Sanguine, V., et al.

(2020). Prognostic factors for severity and mortality in patients infected with COVID-19: A systematic review. PLoS One *15*, e0241955.

31. Mueller, A.A., Tamura, T., Crowley, C.P., DeGrado, J.R., Haider, H., Jezmir, J.L., Keras, G., Penn, E.H., Massaro, A.F., and Kim, E.Y. (2020). Inflammatory Biomarker Trends Predict Respiratory Decline in COVID-19 Patients. Cell Rep Med *1*, 100144.

32. Crowley, C.P., Merriam, L.T., Mueller, A.A., Tamura, T., DeGrado, J.R., Haider, H., Salciccioli, J.D., and Kim, E.Y. (2021). Protocol for assessing and predicting acute respiratory decline in hospitalized patients. STAR Protoc *2*, 100545.

33. Manchanda, E.C., Sanky, C., and Appel, J.M. (2021). Crisis standards of care in the USA: a systematic review and implications for equity amidst COVID-19. J. Racial Ethn. Health Disparities *8*, 824–836.

34. White, D.B., and Lo, B. (2020). Mitigating Inequities and Saving Lives with ICU Triage During the COVID-19 Pandemic. Am. J. Respir. Crit. Care Med. *203*, 287–295.

35. Chomilo, N.T., Heard-Garris, N., DeSilva, M., Blackstock, U. The Harm Of A Colorblind Allocation Of Scarce Resources. Health Affairs Blog, April 30, 2020. https://doi.org/10.1377/hblog20200428.904804.

36. Mello, M.M., Persad, G., and White, D.B. (2020). Respecting Disability Rights—Toward Improved Crisis Standards of Care. N. Engl. J. Med. *383*, e26.

37. Cleveland Manchanda, E., Couillard, C., and Sivashanker, K. (2020). Inequity in Crisis Standards of Care. N. Engl. J. Med. *383*, e16.

38. Dennis, J.M., McGovern, A.P., Vollmer, S.J., and Mateen, B.A. (2021). Improving Survival of Critical Care Patients With Coronavirus Disease 2019 in England: A National Cohort Study, March to June 2020. Crit. Care Med. *49*, 209–214.

39. DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics *44*, 837–845.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
SPSS Statistics Version 25	IBM	https://www.ibm.com/analytics/spss-statistics-software
R version 3.6.1	The R Project	https://www.r-project.org
Simulation of clinical decision-making (selecting from small groups of patients): Groups analysis		https://github.com/maheetha/CSC

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to and will be fulfilled by Lead Contact, Dr. Edy Kim (ekim11@bwh.harvard.edu), Division of Pulmonary and Critical Care Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, 02115 USA.

Materials availability

No new reagents or materials were generated as part of this study.

Data and code availability

Patient data reviewed in this study are not publicly available due to restrictions on patient privacy and data sharing. Individual, patient level data are not currently available because there are individual data use agreements with each of the 67 participating STOP-COVID institutions that do not permit sharing of individual patient data with outside entities. Summary data from STOP-COVID are publicly available in the prior publications, such as Gupta et al.[1]

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Approval for this study was obtained from the Mass General Brigham Institutional Review Board. Demographic details of the study population, including age, gender, race, and ethnicity are provided in Table 2. Race and ethnicity were self reported. Race was reported as white, Black, Asian, American Indian / Alaska Native, Native Hawaiian / Other Pacific Islander, More than One Race, and Unknown / Unspecified. Ethnicity was reported as Hispanic / Latino, Non-Hispanic / Non-Latino, and Unknown.

METHOD DETAILS

Study Design and Population

This is a multicenter, retrospective cohort study, utilizing the previously published cohort the Study of the Treatment and Outcomes in Critically Ill Patients with COVID-19 (STOP-COVID), with inclusion, exclusion and data collection previously described in detail.[1] This study enrolled 4,717 consecutive adult patients with laboratory-confirmed COVID-19 admitted to ICUs at 68 hospitals across the United States from March 4 to June 17, 2020 (Table S1).[1] Inclusion criteria for the current manuscript were intubation on ICU day 1 and availability of data required to calculate SOFA scores within the first 48 hours of ICU admission.

The SOFA score is a tool to assess the level of dysfunction of six organ systems, including respiratory function (ratio of the partial pressure of arterial oxygen to the fraction of inspired oxygen [PaO2 / FiO2]), coagulation (platelet count), liver function (total bilirubin), neurological function (Glasgow Coma Scale), cardiovascular function (number and dose of vasopressors), and renal function (serum creatinine and urine output). Colorado's algorithm altered the SOFA respiratory score to either pulse oximetry measurement of percent oxygen saturation (SpO_2) or the standard arterial blood gas measurement of percent arterial oxygen saturation (PaO_2). Of the STOP-COVID cohort, a total of 2,445 patients were excluded for lack of intubation, intubation later than ICU day 1, or lack of data to calculate SOFA scores (Figure S1). Of 2,866 (20% of original cohort) patients intubated on ICU day 1, 594 patients were excluded due to insufficient clinical data to calculate SOFA score (Table S2). The analysis by race and ethnicity was restricted to patients who self-identified as Black or white, as other self-identified categories had low numbers of patients (i.e., Asian, American Indian / Alaska Native, Native Hawaiian / Other Pacific Islander, More than One Race).

For each patient, CSC priority points were calculated according to two state algorithms (New York and Colorado) and a hypothetical algorithm of raw SOFA scores not grouped into ranges (Table 1). New York's algorithm grouped raw SOFA scores into three groups of ranges. Colorado's algorithm incorporated two components: raw SOFA scores grouped into four groups of ranges and comorbidities according to the Charlson Comorbidity Index. For this study, we adapted the Charlson Comorbidity Index to comorbidity data available in the STOP-COVID database (Tables S3 and S4), which we refer to as the ''modified'' Colorado model. The algorithms are described further in the Results section and in Table 1. The primary outcome was 28-day in-hospital mortality. Patients discharged alive from the hospital prior to 28 days were considered to be alive at 28 days. The validity of this assumption was verified in a subset of patients, as described elsewhere.[1]

Data Collection

Data for the STOP-COVID cohort were collected by manual review of electronic health records as described previously.[1] Demographic data collected included age, gender, self-reported race and ethnicity, and comorbidities. Clinical data were collected at the time of ICU admission and included measurements of hemodynamics and oxygenation, respiratory and vasopressor support, and laboratory values.

SOFA scores were calculated using data from ICU Day 1. Each ICU day is defined as a 24-hour period, from midnight to midnight. ICU Day 1 refers to the 24-hour period from the midnight prior to ICU admission to the midnight after ICU admission. If more than one lab value was available, the first value (i.e., first value recorded after midnight) was taken as the value for the 24-hour time period. If unavailable, data from ICU Day 2 were used. If no value was available on either ICU days 1 or 2, the following approach to missing data was followed: Patients were excluded from the analysis if they had missing data for the following components: PaO2 (161 patients), FiO2 (86 patients), platelets (37 patients), bilirubin (176 patients), altered mental status (310). Some patients had multiple missing values. No patients were excluded for a missing creatinine value. A total of 594 out of 2866 patients (20%) intubated on Day 1 of ICU admission were excluded based on lack of data availability. The SOFA score[10] was adapted to accommodate the STOP-COVID database (Table S2). The scoring of the SOFA cardiovascular component was adapted to the STOP-COVID registry which did not collect data on vasopressor dosage, only the number of vasopressors/inotropes administered each day. U.S. intensivists typically choose norepinephrine as the first vasopressor, so initiation of a vasopressor was scored as 3, corresponding to the scoring of norepinephrine initiation in standard SOFA scoring. The addition of a second vasopressor was scored as 4, since a second vasopressor is typically added only when norepinephrine dosage > 0.1. These adaptations eliminated the cardiovascular score of 1 (mean arterial pressure < 70), and we cannot exclude exceptions to the most common clinical practice in study sites. For the central nervous system (CNS) component, the Glasgow Coma Scale was approximated based on whether ''altered mental status'' (AMS) was indicated on the most recent physical exam prior to intubation. A score of ''1'' indicates that the patient had AMS, while a score of ''0'' indicates that the patient did not have AMS. A total of 310 patients who were marked as ''data not available'' were excluded. This adaptation lacks the range of CNS scoring in standard SOFA scoring.

QUANTIFICATION AND STATISTICAL ANALYSIS

Normality was assessed using the Shapiro-Wilk test. Descriptive statistics were reported as mean (standard deviation) for normal distributions or median (interquartile range) for non-normal distributions. Standard error was calculated using the method described by DeLong et al.,[19] and confidence intervals were calculated with the exact binomial test. For continuous variables, unpaired Student's t tests (normal distribution) or Mann-Whitney U tests (non-normal distribution) were used for two-group comparisons. Area under the receiver operating characteristic (AUROC) curves were calculated to assess the accuracy of each CSC algorithm in discriminating 28-day in-hospital mortality. AUROCs were compared according to the method of DeLong et al.[39]

To simulate a clinical scenario, we analyzed algorithm performance in small groups of two or five patients drawn at random from either the entire cohort (Table 3) or subcohorts defined by race (Table S5). We performed 100 iterations of a computational simulation in which we randomly selected 1,000 groups of two or five patients. Patient groups were excluded in which all the patients had the same outcome (i.e., all survivors or all deceased), since we cannot assess if the algorithm correctly selects a patient with a better outcome if all the patients in that group shared the same outcome. Table 3 column A**:** For each simulation of 1,000 patient groups, we calculated the percent of groups for which the algorithm had a single patient with the ''best'' (lowest) priority score, and so a tie-breaker, such as a lottery, was not required**.** Table 3 **column B:** Among the patient groups that did not require a tie-breaker, we assessed whether the algorithm made a ''correct decision,'' as defined by the selection of a surviving patient. Table 3 column C**:** We calculated algorithm performance in making ''correct decisions'' (i.e., selecting a surviving patient) across all groups, that is the groups in column B (no tie-breaker needed) and the groups that required a tie-breaker. We further examined the effect of adding age as the 1st tie-breaker before lottery. Each simulation of 1,000 patient groups was iterated 100 times to generate a distribution. An unpaired t test was used to calculate significant differences between the distributions. Statistical analysis was conducted in SPSS Statistics Version 25 (IBM) and R Version 3.6.1 (The R Project).