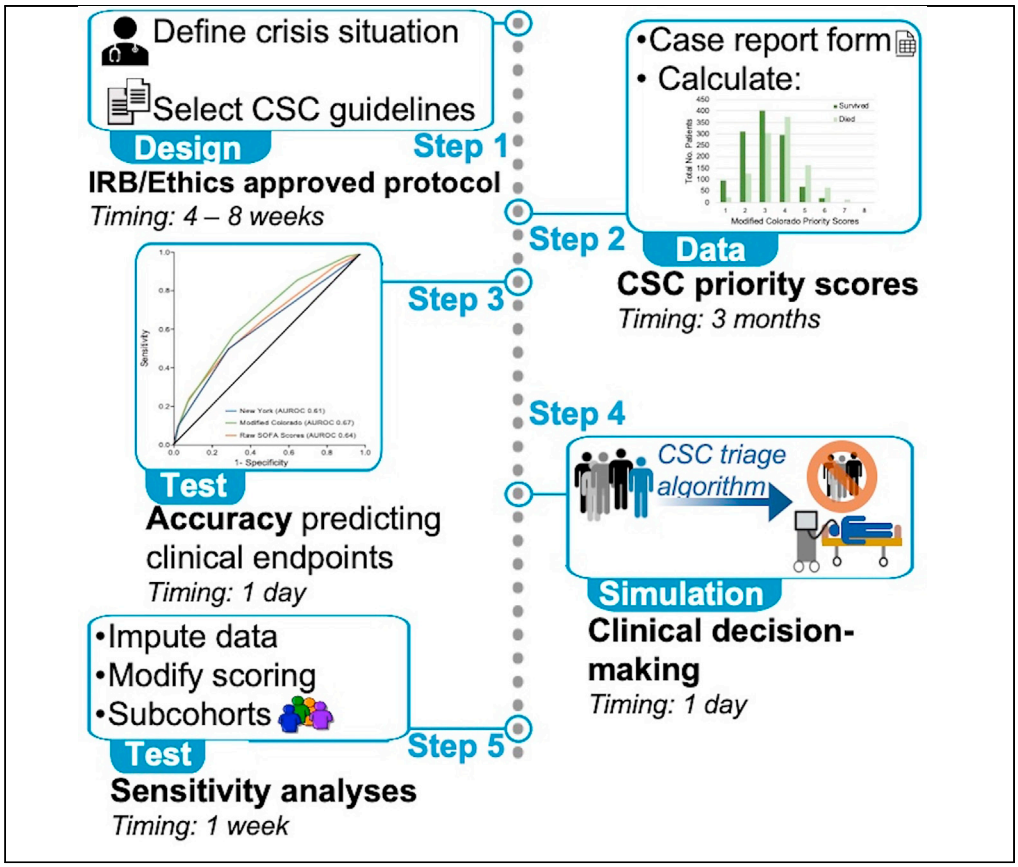
STAR Protocols

CellPress
OPEN ACCESS

Protocol

# Protocol to assess performance of crisis standards of care guidelines for clinical triage




During the COVID-19 pandemic, US states developed Crisis Standards of Care (CSC) algorithms to triage allocation of scarce resources to maximize population-wide benefit. While CSC algorithms were developed by ethical debate, this protocol guides their quantitative assessment. For CSC algorithms, this protocol addresses (1) adapting algorithms for empirical study, (2) quantifying predictive accuracy, and (3) simulating clinical decision-making. This protocol provides a framework for healthcare systems and governments to test the performance of CSC algorithms to ensure they meet their stated ethical goals.

Louis T. Merriam, Maheetha Bharadwaj, Julia L. Jezmir, David E. Leaf, Edy Y. Kim

cmerriam100@gmail.com (L.T.M.)
ekim11@bwh.harvard.edu (E.Y.K.)

## Highlights

Scoring with Crisis Standards of Care (CSC) triage algorithms

Assessing the predictive accuracy of triage algorithms

Simulating clinical decision-making by triage algorithms

Troubleshooting disease severity, comorbidity scoring, and ties

Merriam et al., STAR Protocols 2, 100943
December 17, 2021 
https://doi.org/10.1016/j.xpro.2021.100943

STAR Protocols

CellPress
OPEN ACCESS

Protocol

# Protocol to assess performance of crisis standards of care guidelines for clinical triage

Louis T. Merriam,[1,5,*] Maheetha Bharadwaj,[2] Julia L. Jezmir,[3] David E. Leaf,[2,4] and Edy Y. Kim[1,2,6,*]

[1]Division of Pulmonary and Critical Care Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA

[2]Harvard Medical School, Boston, MA 02115, USA

[3]Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA

[4]Division of Renal Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA

[5]Technical contact

[6]Lead contact

*Correspondence: cmerriam100@gmail.com (L.T.M.), ekim11@bwh.harvard.edu (E.Y.K.)
https://doi.org/10.1016/j.xpro.2021.100943

## SUMMARY

**During the COVID-19 pandemic, US states developed Crisis Standards of Care (CSC) algorithms to triage allocation of scarce resources to maximize population-wide benefit. While CSC algorithms were developed by ethical debate, this protocol guides their quantitative assessment. For CSC algorithms, this protocol addresses (1) adapting algorithms for empirical study, (2) quantifying predictive accuracy, and (3) simulating clinical decision-making. This protocol provides a framework for healthcare systems and governments to test the performance of CSC algorithms to ensure they meet their stated ethical goals. For complete details on the use and execution of this protocol, please refer to Jezmir et al. (2021).**

## BEFORE YOU BEGIN

Timing: 1–2 h

The protocol below focuses on triage of critically ill patients, and the example in Expected Outcomes uses a patient cohort of patients with acute respiratory distress syndrome (ARDS) due to coronavirus disease 2019 (COVID-19). However, we have also applied this protocol to critically ill patients with sepsis (Bharadwaj et al., 2021), and the protocol can be used for hospitalized patients with non-critical illness.

### Basics of Crisis Standards of Care (CSC) guidelines and triage algorithms

1. **The need for quantitative testing of CSC algorithms.** CSC guidelines were developed to triage patients in times when medical resources are scarce, such as in natural disasters, mass casualty events, and pandemics. These guidelines attempt to maximize the number of lives saved across an entire population in an ethical manner. Central to CSC are predictive triage algorithms that utilize physiologic and demographic criteria to prioritize those patients most likely to survive when allocated scarce resources, such as ventilators and intensive care unit (ICU) beds (Iacorossi et al., 2020). Ethicists have long debated the design of triage algorithms, their limitations in predictive power and the circumstances in which they should be used. Incomplete knowledge of the behavior of CSC algorithms in actual clinical cohorts hampers comparisons of their performance. During the COVID-19 pandemic, U.S. state governments published CSC algorithms with distinct approaches. Most of these algorithms use a measure of clinical disease severity, such as the

Sequential Organ Failure Assessment (SOFA) score. However, there is considerable state-to-state variation in whether prior medical comorbidities are incorporated and how priority points are assigned. Without empirical testing, it is unclear how these triage algorithms perform in practice, and whether they exacerbate racial or ethnic health disparities. Studies utilizing empiric testing by our group and others found that scoring systems representing commonly used approaches poorly predicted survival among critically ill patients with COVID-19 and frequently resulted in tied priority scores (Bharadwaj et al., 2021; Jezmir et al., 2021). The predictive accuracy of CSC algorithms for 28 day mortality were reduced in Black when compared to white patients, which suggested the need for further study of the potential of CSC to exacerbate health disparities. Therefore, the use of statistical and computational methods to empirically test the performance of triage algorithms may help guide their revision to better fulfill their stated ethical goals.

2. **Clinical endpoints.** The most common clinical endpoint discussed for triage algorithms is survival. Ethical principles would, in theory, focus on long-term timepoints like years. However, research studies are typically constrained to shorter timeframes, such as 28 days. Besides survival, other clinical endpoints can be considered. For example, randomized controlled trials (RCTs) often use endpoints such as organ support free days (e.g., number of days not receiving mechanical ventilation or renal replacement therapy), clinical remission or stability. To capture morbidity and optimize statistical power, 40% of COVID-19 phase III RCTs used an ordinal scale that assigned points based on a patient's clinical status (Desai and Gyawali, 2020). For example, in the ACTT RCTs, their 8 point scale includes categories such as 1 point for death, 3 points for hospitalized on high-flow nasal cannula or non-invasive ventilation, and 7 points for not hospitalized but with limitations on activities or on supplemental oxygen (Beigel et al., 2020).
3. **Sequential Organ Failure Assessment (SOFA) and Modified SOFA (MSOFA).** A basic understanding of how CSC triage algorithms are developed and scored is necessary before empirical testing. Most CSC guidelines incorporate a measure of clinical disease severity. Many scoring systems have been developed in clinical medicine to attempt to predict morbidity and mortality, such as APACHE I, II & III, SAPA, MODS, MPM, and TRIOS (Vincent and Moreno, 2010). SOFA was originally developed in 1994 to predict mortality among critically ill patients (Vincent et al., 1996). Its initial use was intended for patients suffering from sepsis; indeed, "S" originally stood for *sepsis*. Over time its application was broadened to other critical illnesses, including patient triage in times of crisis, and is now the primary tool for U.S. State CSC guidelines. The SOFA score is calculated based on data from six organ systems (Table 1). To simply SOFA, a modification was proposed in 2010 as MSOFA that eliminates several laboratory values (e.g., platelet count, serum bilirubin, partial pressure of arterial oxygen [$PaO_2$]) and adds bedside assessments of pulse oximetry and scleral icterus or jaundice (Grissom et al., 2010). SOFA and MSOFA are used by over 90% of states that triage patients based on mortality risk (Hantel et al., 2021; Piscitello et al., 2020).

For CSC algorithms, after initial calculation of SOFA or MSOFA for each patient, raw SOFA or MSOFA scores are grouped into ranges (Table 2). For example, New York's state government guidelines assign 1 priority point to SOFA scores less than 7, and two priority points for SOFA score ranging from 8 to 11 (https://www.health.ny.gov/regulations/task_force/reports_publications/docs/ventilator_guidelines.pdf, accessed June 20, 2020). To address the limitations inherent in disease severity scores like SOFA, some CSC guidelines also incorporate data on comorbidities. For example, Colorado's state government guidelines incorporates both SOFA score ranges and comorbidity scoring based on the Charlson comorbidity index (https://www.colorado.gov/pacific/sites/default/files/Crisis%20Standards%20of%20Care%20Triage%20Standards-April%202020.pdf, accessed June 20, 2020). Rather than comprehensively review CSC algorithms (Hantel et al., 2021), the focus of this protocol is to provide a method for empirical testing that can be applied to any existing or hypothetical triage algorithm.

4. **Clinical data used in triage algorithms.** Understanding triage algorithms necessitates an introduction to the physiologic and laboratory measurements that may be unfamiliar to non-clinicians

**Table 1. SOFA and MSOFA scoring**

| | | Score | | | | |
|---|---|---|---|---|---|---|
| System | Target | 0 | +1 | +2 | +3 | +4 |
| SOFA | | | | | | |
| Respiratory | Pa02/Fi02 | >400 | <400 | <315 | <235 | <150 |
| CNS | Glasgow Coma Scale | 15 | 13–14 | 10–12 | 6–9 | <6 |
| Cardio-vascular | MAP mmHg<br>Doses of drug mcg/kg/min | > 70 | > 70 | Dopamine < 5 **OR ANY** dobutamine | Dopamine > 5 Epinephrine **OR** Norepinephrine < .1 | Dopamine < 5 Epinephrine **OR** Norepinephrine < .1 |
| Liver | Bilirubin mg/dL | < 1.2 | 1.2–1.9 | 2.0–5.9 | 6.0–11.9 | > 12 |
| Coagulation | Platelets #/micrL | > 150 | < 150 | < 100 | < 50 | < 20 |
| Renal | Creatinine (mg/dL)<br>Urine output (cc/24h) | < 1.2 | 1.2–1.9 | 2.0–3.4 | 3.5–4.9<br>< 500 | > 5.0<br>< 200 |
| MSOFA | | | | | | |
| Respiratory | Sp02/Fi02 | | | | | |
| CNS | Glasgow Coma Scale | 15 | 13–14 | 10–12 | 6–9 | <6 |
| Cardiovascular | MAP (mmHg)<br>Doses of drug mcg/kg/min | > 70 | > 70 | Dopamine < 5 **OR ANY** dobutamine | Dopamine > 5 Epinephrine **OR** Norepinephrine < .1 | Dopamine < 5 Epinephrine **OR** Norepinephrine < .1 |
| Liver | Icterus **OR** Jaundice | None | | | Scleral Icterus **OR** Jaundice | |
| Renal | Creatinine, mg/dL | < 1.2 | 1.2–1.9 | 2.0–3.4 | 3.5–4.9 | > 5.0 |

who are stakeholders in creating CSC guidelines, such as ethicists, government representatives, and the public. Here, we focus on the components of SOFA and MSOFA scores:

a. **Respiratory.** $PaO_2$, or *partial pressure of arterial oxygen*, is the amount of oxygen (mmHg) in arterial blood and requires laboratory measurement. The $SpO_2$ is the *saturation of oxygen in arterial blood* (stated as percentage) measured continuously and non-invasively by a transcutaneous oximeter (i.e., pulse oximeter). Since laboratory measured $PaO_2$ may not be available in the medical record, studies benefit from estimating $PaO_2$ by imputation from $SpO_2$ measurements (Brown et al., 2017), as discussed in Table 1 of our prior STAR Protocol "Assessing and predicting acute respiratory decline in hospitalized patients" (Crowley et al., 2021). $FiO_2$ is the *percentage of oxygen in the air* inspired by the patient. Room air (i.e., no oxygen

**Table 2. Triage algorithm scoring schemes and example**

| Model | New York | Modified Colorado | Raw SOFA score |
|---|---|---|---|
| SOFA Priority Points | SOFA < 7: 1 point<br>SOFA 8–11: 2 points<br>SOFA > 11: 3 points | SOFA < 6: 1 point<br>SOFA 6–9: 2 points<br>SOFA 10–12: 3 points<br>SOFA > 12: 4 Points | Assign 1 priority point per SOFA score point (e.g., SOFA 1: 1 point; SOFA 2: 2 points) |
| Comorbidity priority points | None | Modified Charlson Comorbidity Index | None |
| Priority Number Calculation | SOFA Score | SOFA Prioritization + Charlson Comorbidity Index Score | SOFA Score |
| Priority Grouping based on Priority Number | High Priority: 1<br>Intermediate: 2<br>Low Priority: 3 | None | None |
| Tie Brakers | **1st Tie Breaker:** Children **2nd Tie Breaker**: Lottery | **1st Tie Breaker**: Children, Health Care Workers and/or First Responders **2nd Tie Breaker**: Age (years lived) Pregnancy, And/or sole caretaker for elderly **3rd Tie Breaker**: Lottery | **1st Tie Breaker**: Age **2nd Tie Breaker**: Lottery |

**New York's CSC algorithm** (https://www.health.ny.gov/regulations/task_force/reports_publications/docs/ventilator_guidelines.pdf, accessed June 20, 2020), the modified Colorado algorithm (https://www.colorado.gov/pacific/sites/default/files/Crisis%20Standards%20of%20Care%20Triage%20Standards-April%202020.pdf, accessed June 20, 2020), and a hypothetical raw SOFA score algorithm are shown.

**Table 3. Glasgow Coma scale**

| Patient behavior | Best response | Score |
|---|---|---|
| Eye opening | Spontaneously | 4 |
| | To Speech | 3 |
| | To Pain | 2 |
| | No Response | 1 |
| Best Verbal Response | Oriented to time, place, person | 5 |
| | Confused | 4 |
| | Inappropriate words | 3 |
| | Incomprehensible sounds | 2 |
| | No response | 1 |
| Best Motor Response | Obeys commands | 6 |
| | Moves to localized pain | 5 |
| | Flexion withdrawal from pain | 4 |
| | Abnormal flexion (decorticate" | 3 |
| | Abnormal extension (decerebrate) | 2 |
| | No response | 1 |
| Total Score | *Best response* | 15 |
| | *Comatose* | < 8 |
| | *Totally unresponsive* | < 3 |

supplementation) has a $FiO_2$ of 21%. Patients receiving supplemental oxygen while on mechanical ventilation, high flow nasal cannula and Venturi face mask have an $FiO_2$ directly selected by the clinician. For other oxygen devices, such as low-flow nasal cannula or simple facemasks, the clinician sets the oxygen *flow rate* but not the $FiO_2$. For these devices, $FiO_2$ can be estimated (Coudroy et al., 2020; Frat et al., 2015), as shown in Table 2 in our prior STAR Protocol (Crowley et al., 2021). Acute hypoxemic respiratory failure results in a gradient between the oxygen content of the inspired air ($FiO_2$) and the resulting oxygen content of the patient's blood ($PaO_2$). This "air-to-arterial blood" oxygen gradient is quantified as the ratio of $PaO_2$ / $FiO_2$, known as the P/F ratio. The P/F ratio allows comparison of patients treated with different concentrations of inspired oxygen.

b. **Central Nervous System (CNS):** Critically ill patients often develop brain dysfunction secondary to several etiologies, such as metabolic disorders (e.g., hepatic dysfunction, renal failure), infection, systemic inflammation, hypoxia, trauma and disordered sleep. The altered sensorium and cognition in critically ill patients are often diagnosed as "delirium due to toxic-metabolic causes." The Glasgow Coma Scale (GCS) is a common method to assess brain function and level of consciousness. The score is based on three areas: Eye opening, verbal response, and motor response (Table 3). The lower the GCS score, the greater the degree of brain injury.

c. **Cardiovascular (CV):** Critical illness can lead to low blood pressure (BP), or *hypotension*. Scoring in SOFA/MSOFA relies on the degree of hypotension as measured by the mean arterial pressure ("MAP") calculated from the systolic and diastolic blood pressure over the course of one heartbeat. MAP is considered superior to either systolic or diastolic blood pressure alone as a measure of end-organ perfusion. MAP is most accurately measured using invasive monitoring, such as an arterial catheter, but can be measured non-invasively by a sphygmomanometer (i.e., blood pressure cuff). The SOFA/MSOFA cardiovascular component also measures the use of i.v. vasopressor medications, such as norepinephrine, as a measure of the severity of underlying hypotension. The vasopressor medications are titrated for a MAP threshold, with 60 or 65 as common goals. A greater dosage of vasopressor reflects more severe hemodynamic compromise.

d. **Liver.** SOFA scoring incorporates the plasma bilirubin level. Over 80% of bilirubin comes from the degradation of hemoglobin from senescent or destroyed erythrocytes. Once released, most bilirubin is metabolized in the liver and excreted into the duodenum via the biliary tract. Liver dysfunction can lead to decreased bilirubin metabolism, excretion, and subsequently

result in elevated levels of circulating bilirubin. When circulating bilirubin exceeds 2.5mg/dl, as seen in liver dysfunction or failure, there is increased deposition of bilirubin in soft tissue that causes the characteristic yellow color of eyes (i.e., scleral icterus) and skin (i.e., jaundice). These signs are clinically assessed at the bedside and used as a proxy for liver dysfunction in MSOFA scoring.

e. **Renal.** Serum creatinine is the most common laboratory value used to measure renal function. Creatinine is a by-product of creatine phosphate metabolism and excreted primarily in the kidney. Urine output is also part of the SOFA score.

f. **Coagulation.** Blood clotting requires a complex system of tissue- based and circulating proteins, blood cells, and other molecules, such as fibrinogen. Thrombocytopenia is the decreased number of platelets, the key cell in coagulation.

5. **Comorbidities.** To help calculate the risk of long-term mortality, some triage algorithms consider patient comorbidity scores adapted from summary measures such as the Charlson and Elixhauser comorbidity indexes. The Charlson comorbidity index includes age, history of cardiovascular disease or stroke, dementia, chronic obstructive pulmonary disease (COPD), connective tissue disease, peptic ulcer disease, hepatic disease, diabetes mellitus complications, hemi- or paraplegia, renal disease, malignancy or acquired immunodeficiency syndrome (AIDS) (Charlson et al., 1987). The Charlson comorbidity index was initially validated for prediction of 1 and 10 year survival and can have good discrimination for in-hospital mortality (Hope et al., 2015; Sundararajan et al., 2004). For hospitalized patients, when the Charlson comorbidity index is combined with basic demographic data (e.g., age, sex, social factors) and surgical/medical status, its predictive accuracy for in-hospital, 30 day and 1 year mortality approaches that of disease severity scores (e.g., SAPS III and APACHE II) and can equal or exceed performance of disease severity scores if the presence of mechanical ventilation or dialysis is added (Christensen et al., 2011). Typically, triage algorithms use comorbidity scoring in conjunction with disease severity scores. While separately assessing comorbidities has been less widely applied in CSC guidelines, their use has proven helpful in improving the predictive power of CSC algorithms in some circumstances, such as candidemia (Asai et al., 2021). However, their use in this context has created concern regarding potential ethical bias against specific patient groups who are more vulnerable to developing comorbidities, such as racial minorities and people with disabilities.

6. **Race, ethnicity, socioeconomic status, age, and disability.** The COVID-19 pandemic highlighted the presence of significant disparities in healthcare in socioeconomic, racial, and ethnic groups (Cleveland Manchanda et al., 2020). An increased prevalence of comorbidities and decreased healthcare access for specific patient groups could inject bias in prognostication and prioritization by triage algorithms (Gershengorn et al., 2021). Evidence is limited regarding how these disparities might affect the performance of CSC algorithms, so empiric testing helps quantify potential bias. For example, we found that several triage algorithms performed worse in Black patients with COVID-19 when compared to white patients (Jezmir et al., 2021). Advanced age is well known to associate with increased mortality in critical illness; however, considerable ethical debate exists regarding the use of age in CSC algorithms (Altman, 2021; Rueda, 2021). Some experts feel that consideration of age is a civil rights violation. One argument is that the patient's overall physiological status, sometimes referred to as their "biological age," drives medical outcomes to a greater extent than chronological age (Hope et al., 2015). Alternatively, some argue for the limited use of age, such as in tie-breakers, since younger patients have not yet had the opportunity to experience all stages of life. The question of including disability in comorbidity indices has also raised ethical debate. For example, hemi- or paraplegia is part of the Charlson comorbidity index used by some CSC guidelines. Other considerations emerged in the tie-breaking criteria used by some CSC guidelines, such as favoring healthcare workers, first-responders, pregnant women, and essential workers. When testing triage algorithms, it is important to understand the civil rights landscape and patient groups that have experienced structural disadvantages.

7. **Electronic health records (EHR).** EHR archives the patient demographic information, laboratory values and physiologic data needed to calculate triage algorithm scores. Despite their ever-increasing use in research and healthcare, extracting and translating EHR data into usable information can be

challenging. Researchers should be aware of techniques for: avoiding error in manual chart review and/or accurate pre-processing of electronic queries; imputation of missing data; and trouble-shooting error in clinical measurements. These pitfalls are discussed in "troubleshooting" and in our prior STAR Protocol (Crowley et al., 2021). Researchers should familiarize themselves with best practices for reproducible clinical research developed from EHR, such as the step-by-step guidance and examples in the open-access textbook developed by the MIT Critical Data group (MIT Critical Data, 2016).

## KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Software and algorithms | | |
| REDCap | Project REDCap | https://www.project-redcap.org |
| SPSS Statistics 25 | IBM | https://ibm.com/products/spss-statistics |
| Prism software (version 9.2.0) | GraphPad | https://www.graphpad.com/scientific-software/prism |
| R software (version 3.6.1) | The R Project | https://www.r-project.org |
| Simulation of clinical decision-making (selecting from small groups of patients) | This paper | https://github.com/maheetha/CSC https://doi.org/10.5281/zenodo.5204454 |
| Experimental models: Organisms/strains | | |
| Human subjects. Age: > 19 yrs. Gender: Male, Female | STOP-COVID registry (Jezmir et al., 2021) | N/A |

## STEP-BY-STEP METHOD DETAILS

### Study design

**Timing: 3 days for protocol development. IRB/ethics review and revision can take 4–8 weeks.**

Empirical testing of CSC triage algorithms starts with careful consideration of the clinical data needed to capture the nature of the crisis. In depth review of existing algorithms, their applicability toward the specific crisis, and whether adequate patient data is available in health records are important steps when designing and testing triage algorithms. In addition, developing study parameters and selecting patient cohorts are necessary before beginning the data acquisition stage.

1. Define the crisis situation (e.g., mass casualty event, COVID-19 pandemic).
2. Select existing algorithms for study, such as those issued by U.S. states: Select representative approaches that test hypotheses and areas of interest, like comorbidities.
3. Define hypothetical triage algorithms for study:

Design de novo algorithms that test hypotheses and areas of interest, such as disease severity scoring systems, comorbidity scoring indices, or tie-breaker criteria.

⚠ CRITICAL: Selection of an initial triage algorithm should reflect the clinical nature of the crisis. We recommend having physician specialists with direct experience in this area to guide this process.

4. Select primary and secondary clinical endpoints—both the type of outcome (e.g., mortality, length of stay) and time point (e.g., 28 days).
5. Determine the clinical data needed to calculate triage algorithm scores and clinical endpoints.
6. Select the patient cohort for study: Common approaches are a retrospective cohort or cross-sectional analysis of a prospective cohort.

7. Assess whether the necessary clinical data are available in this patient cohort.

⚠ CRITICAL: To make the study feasible, the clinical endpoint(s) or choice of patient cohort may need to be changed or modified. A multi-disciplinary team of domain experts (e.g., clinician-scientists, data scientists, and biostatisticians) is essential to ensure the proposed study has clinical relevance and statistical rigor.

8. Adapt the triage algorithm scoring scheme to data available in the selected patient cohort.
9. Submit the proposed study to the IRB or ethics panel.

⚠ CRITICAL: IRB or ethics approval is required prior to further work.

### Calculating priority scores

Timing: 3 weeks to 3 months (varies by method of data acquisition)

Acquiring and recording patient data needed for triage scoring is an important process when testing algorithms. Viewing and recording health information requires a secure management system (i.e., software) that will protect patients privacy as well as maintain data integrity. The following section outlines the steps involved in developing and recording data into a case report form, searching data from health records, and using the case report form to calculate triage and priority scores from the designed algorithm.

10. Create a case report form with fields for all data required to calculate triage algorithm scores and clinical endpoints, as discussed in our prior STAR Protocol (Crowley et al., 2021).

Select a data management system. Key features include collaborative tools, security to protect patient health information, and back-up methods. Examples approved at our institution include REDCap (www.project-redcap.org), enterprise Dropbox, or enterprise Microsoft Teams.

⚠ CRITICAL: Data management systems must be approved by your institution's human subject research policies.

11. Select the method of data entry (e.g., manual chart review, electronic queries) and establish protocols for data quality checks and pre-processing.
12. Perform pilot data acquisition with the case report form. In an iterative fashion, refine protocols for data entry and pre-processing. Confirm the study's feasibility in this cohort.
13. After finalizing the data acquisition protocol, complete the case report form for all patients.
14. Apply the triage algorithms and calculate priority scores for each patient in the cohort.

⚠ CRITICAL: Case report forms should anticipate and include additional data fields that may be needed for sensitivity testing and algorithm adaptation.

### Testing algorithm accuracy

Timing: 1 day

After the design, data acquisition, and calculation of priority scores are completed, the next step is to test the algorithm's accuracy in predicting the defined outcome(s). Utilizing each patient's priority score, the study parameters, and the known outcomes obtained from the health record, the number of true and false positives within the patient cohort can be calculated. Using this data, accuracy can be calculated. We recommend using the AUROC method by DeLong et al. (1988).

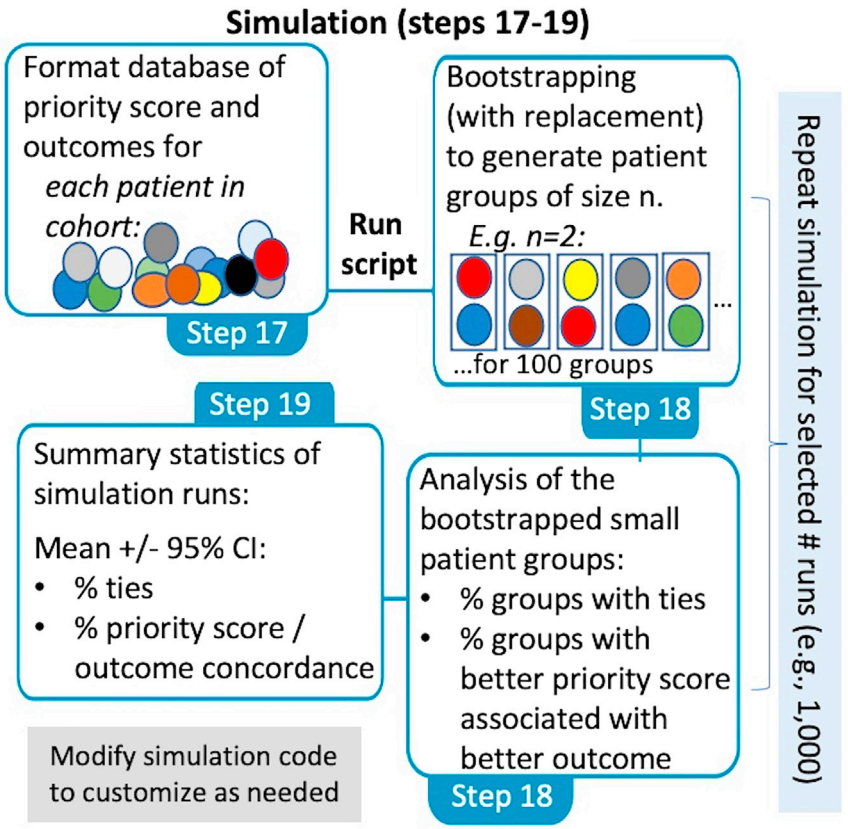



**Figure 1. Simulation of small patient group clinical decision-making (steps 17–19)**

In this illustration, investigators selected groups of two patients (n=2); generation of 100 patient groups per simulation run; and iteration of the simulation for 1,000 runs to generate summary statistics. The simulation code is open-access and can be modified as needed (e.g., to model patient groups of different sizes within the same simulation run). Reproduced from Jezmir et al. (2021)

15. Determine the accuracy of the priority scores for clinical endpoints that are binary values (e.g., survival) using the area under the receiver operating characteristic (AUROC) curve.

### Simulation of clinical decision-making: Selection from a smaller group

Timing: 1 day

Clinicians will make triage decisions in small groups of patients (e.g., 2 to 5 patients), rather than across a large population. To quantify performance in a simulation of this scenario, this protocol section uses a bootstrap method and then iterates this sampling to generate summary statistics for performance (Figure 1). The code and an example using sample data is provided at https://github.com/maheetha/csc.

16. Format input files by assigning a patient to each row and assigning these columns:
    a. Outcomes: 0 or 1
    b. Priority score calculated by triage algorithm 1
    c. Priority score calculated by triage algorithm 2
    d. Priority score calculated by triage algorithm 3
       Continue for all triage algorithms tested
    e. Sensitivity analyses (e.g., race)

17. Run CSC script: Rscript Groups_Analysis.R [input file] [group size] [number of iterations] [size of each bootstrap] [final output filename]:

The [group size] refers to the number of patients in each small patient group that one patient is selected from (e.g., 5 patients). The [size of each bootstrap] refers to the number of patient small groups randomly selected in each iteration.

⚠ CRITICAL: Ensure input file is formatted correctly in step 17.

18. Assess bootstrap analysis output, with sequential columns reporting:
    a. Triage algorithm
    b. Mean percentage of patient selections made without tied priority scores
    c. 95% confidence interval (CI)
    d. Mean percentage of decisions that selected the patient with the better outcome (among non-tied decisions)
    e. 95% CI for non-tied decisions that selected for the better outcome.
    f. Mean percentage of all decisions (i.e., tied or non-tied) that chose a surviving patient
    g. 95% CI for all decisions that chose a surviving patient.

### Sensitivity analyses

Timing: 1 week

Test the sensitivity of triage algorithm performance to key factors by repeating the analysis in steps 16–19 with different inputs.

19. Test the effect of data processing methods by repeating step 16–19 with different data protocols (e.g., different methods of imputing missing data values).
20. Test the effect of patient characteristics by repeating steps 16–19 in patient sub cohorts (e.g., subdivided by race),
21. Test the effect of triage algorithm components by repeating steps 16–19 with modified triage algorithms (e.g., separately scoring the disease severity and comorbidity components).

## EXPECTED OUTCOMES

The examples below demonstrate application of this protocol to 2,272 hospitalized patients from the STOP-COVID registry at the time of ICU transfer and intubation due to COVID-19 in our study (Jezmir et al., 2021).

### Method 1: Adapt the algorithm scoring scheme (step 8)

Colorado's CSC guidelines specify a triage algorithm incorporating both SOFA score ranges and the Charlson comorbidity index. In step 8, we adapt Colorado's algorithm to the data available in our data registry. For the cardiovascular component of the SOFA, we only have data on whether 0, 1 or 2 vasopressors were used. Since we lack MAP values and the identity and dosages of the vasopressors, we adapt scoring as: no vasopressor = 0 points; 1 vasopressor = 3 points; 2 vasopressor = 4 points. We assign 1 vasopressor as 3 points, rather than 2 points, because we assume that the first-line vasopressor is norepinephrine, following standard clinical practice. For the CNS component of the SOFA score, we lack data to calculate the GCS, so we adapt scoring as: no altered mental status = 0 points; altered mental status = 1 point, as in Table S2 of (Jezmir et al., 2021). To adapt the Charlson comorbidity index to our dataset, as in Table S3 of (Jezmir et al., 2021), we omit scoring of dementia, hemi- or paraplegia, AIDS, and metastatic solid tumor. All hepatic disease is scored as mild, and all diabetes mellitus is scored as with chronic complications.

**Table 4. Example of the calculation of priority points in a patient**

| Patient ID: #5 | PmHx: Diabetes | | 28 day outcome: Alive | |
|---|---|---|---|---|
| Category/system | Value | SOFA score | NY Algorithm | Colorado Algorithm |
| $Pa0_2/Fi0_2$ | 350 | 1 | 1 | 1 |
| Glasgow Coma Scale # | 14 | 1 | 1 | 1 |
| Liver: Bilirubin (mg/dl) | 0.2 | 0 | 0 | 0 |
| Coagulation: (Platelet # in 1000s) | 200 | 0 | 0 | 0 |
| Renal: Creatinine (mg/dl) | 1.2 | 1 | 1 | 1 |
| Comorbidity score (Charlson) | | no | No | 2 |
| **Total Score** | | **3** | **3** | **3** |
| **Priority Points** | | **3** | **1** | **5** |
| **Priority Group** | | **no** | **High** | **High** |

### Method 2: Calculate priority scores and test algorithm accuracy (steps 14–16)

**Calculation of priority scores.** In steps 14–15, we calculate priority scores following the New York (NY), Colorado and hypothetical raw SOFA score algorithms (Table 2). Clinical data at the time of ICU transfer and intubation are manually extracted from the EHR and placed into the case report form. Table 4 is an example of a case report form from a hypothetical surviving patient with diabetes. It demonstrates the calculation of the raw SOFA score and how triage priority points for the SOFA component are assigned following the NY and Colorado triage algorithms. For the hypothetical algorithm of raw SOFA, each SOFA score point equals one priority point. NY's algorithm groups raw SOFA scores into three ranges, with the lowest SOFA scores sorted into the "high priority" category that would be favored for scarce resources. Colorado's algorithm groups raw SOFA scores into four ranges. Unlike NY, Colorado also has a second priority scoring component based on comorbidities. Colorado's algorithm assigns priority points based on a scale adapted from the Charlson comorbidity index. For Colorado's algorithm, the priority points from the SOFA and comorbidity components are added together to yield the final priority score. In Colorado's algorithm, patients with the lowest numerical priority score are favored for scarce resources.

**Testing triage algorithm accuracy.** Triage algorithms favor allocation of scarce resources to patients with the fewest priority points (i.e., lowest scores for disease severity and comorbidities). The ethical justification assumes that higher priority scores predict poor outcomes. In our example, higher priority scores associated with greater 28d mortality for every triage algorithm tested (Figures 2A–2C). In step 16, to quantitatively compare the predictive accuracy for 28 day mortality, we calculate area under the receiver operating characteristic (AUROC) curve for each triage algorithm (Figures 2D and 2E).

### Method 3: Simulate clinical decision making (steps 17–19).

AUROC curve assess triage algorithm performance across a large population. However, in "real world" conditions, clinical triage teams decide prioritization among smaller groups of patients that need a scarce resource at the same time. Following steps 17–19, we implement a bootstrap method of random resampling to assess triage algorithm performance in small groups of patients. In this example, we randomly select 1,000 groups of five patients each. We exclude small groups in which all the patients have the same outcome (i.e., all survivors or all deceased at 28d). For each small group, we calculate the priority score for each patient for every triage algorithm. In each small group of patients, we assess whether a single "winner" (with the most favorable priority score) can be selected or whether two or more patients are tied for the most favorable priority score. If a single winner is selected, we assess whether the selected "winning" patient had the better outcome (i.e., survival). We then tabulate the number of small groups in which the triage algorithm avoided ties. We also count whether the triage algorithm "correctly" selected a patient with the better outcome. We then iterate this process 100 times to estimate summary statistics (Table 5A).

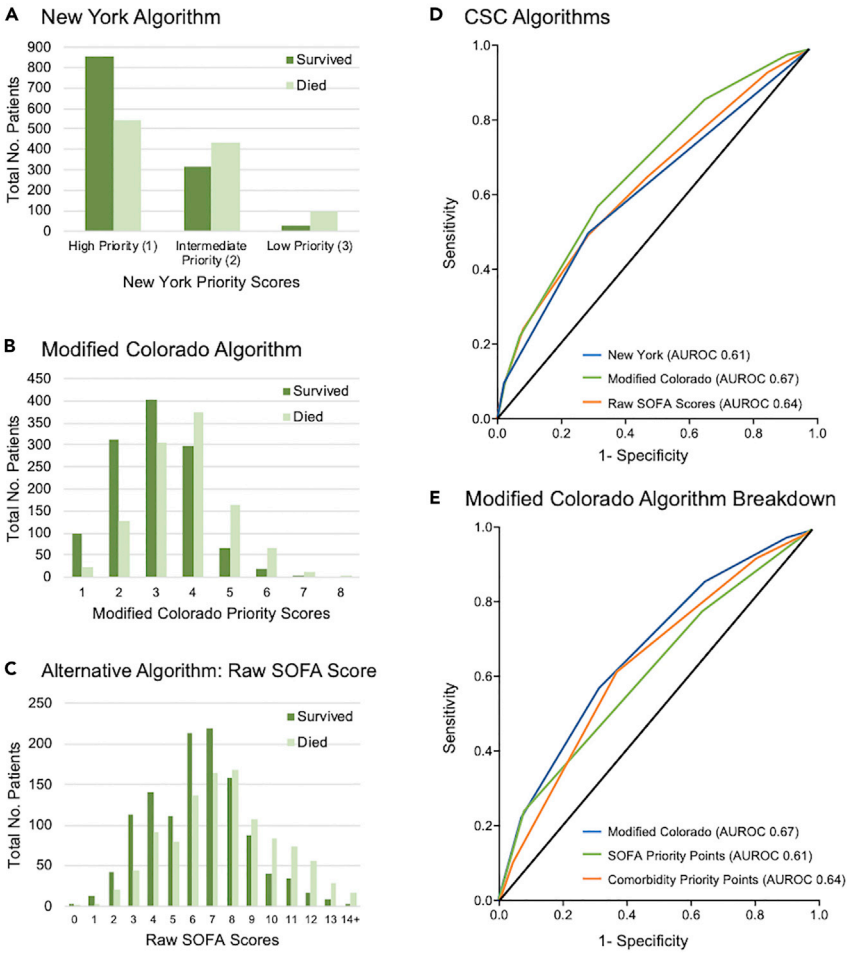



**Figure 2. Calculating priority scores and testing predictive accuracy**

(A–C) Number of surviving and deceased patients in each priority category of (A) New York's algorithm or in each priority point total for (B) Colorado's algorithm, or (C) a hypothetical raw SOFA score algorithm.

(D and E) AUROC curve shown for the accuracy of priority point totals in predicting 28d mortality for (D) three triage algorithms and (E) the component parts of Colorado's algorithm (SOFA priority points or comorbidity priority points) or Colorado's algorithm incorporating both SOFA and comorbidity scores. (Reproduced from Jezmir et al., 2021).

### Method 4: Sensitivity analyses (steps 20–22).

Sensitivity analyses help identify the determinants of triage algorithm performance by modifying the data inputs or the triage algorithms themselves. In an example of applying step 21 to the simulation of clinical decision-making (steps 17–19), we re-test the triage algorithms in patient sub cohorts separated by race. Colorado's algorithm had modestly worse performance for Black patients when compared to white patients (Tables 5B and 5C). In an example of applying step 22 to AUROC curve analysis (step 16), we modify Colorado's triage algorithm to test the SOFA component and comorbidity component separately (Figure 1E). Here, the original triage algorithm incorporating both SOFA and comorbidity components mildly outperforms each component used on its own. Our analysis in Jezmir et al. (2021) delineates other sensitivity analyses, such as the effect of excluding end-stage renal disease from the comorbidity index or adding age as a tie-breaker.

## QUANTIFICATION AND STATISTICAL ANALYSIS

The steps below detail key considerations in statistical analysis. This approach was used in our study of patients with critical illness due to COVID-19 detailed in "Expected outcomes."

**Table 5. A bootstrap method to assess a simulation of clinical decision-making**

| Triage algorithm | % Decisions not requiring a tie-breaker: | % Correct selections among decisions not requiring a tie-breaker | Overall performance: % Correct selections among all decisions |
|---|---|---|---|
| A. Full cohort | | | |
| New York | 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) |
| B. white subcohort | | | |
| New York | 6 (5–7) | 64 (51–75) | 61 (58–64) |
| Colorado | 58 (56–61) | 74 (70–77) | 71 (69–74) |
| Raw SOFA | 78 (76–81) | 66 (63–69) | 65 (62–69) |
| C. Black subcohort | | | |
| New York | 12 (10–14) | 63 (51–71) | 60 (57–63) |
| Colorado | 58 (55–61) | 66 (63–70) | 63 (60–65) |
| Raw SOFA | 81 (78–83) | 60 (57–63) | 60 (57–63) |

This example ran 100 iterations. In each iteration 1,000 randomly selected groups of five patients each were examined. (A) Full cohort. (B-C) Sensitivity analyses of subcohorts of (B) white or (C) Black patients. Mean (95% CI) shown. Adapted from Jezmir et al. (2021).Figure legends

1. To assess summary statistics of patient demographics and clinical characteristics:
   a. Assess whether the data have a normal distribution using the Shapiro-Wilk test, D'Agostino and Pearson test, or Kolmogorov-Smirnov test.
   b. For data with normal distribution, calculate mean with standard deviation. For non-normal data, use median values with interquartile ranges.
   c. For comparisons of continuous variables in two groups, use Student's t test (if normal) or Mann-Whitney U test (if non-normal).
   d. For comparisons of variables that are binary, use the chi-square or fisher exact test.
2. In step 16, to assess the predictive accuracy of triage algorithms, use the AUROC curve:
   a. Compare AUROC curves by the method of DeLong et al. (1988)
3. In steps 17–19, the summary statistics of the bootstrap method will converge to a normal distribution with sufficient iterations, such as the 1,000 iterations in our example, due to the central limit theorem.
4. In our study (Jezmir et al., 2021), statistical analysis was performed using SPSS, R, and custom-written code detailed in the key resources table. Prism can also be used for analyses.

## LIMITATIONS

### Adaptation of the algorithm scoring scheme (step 8)

Adaptation of the algorithm scoring scheme to available clinical data is necessary but can alter the performance of the algorithm. In our example in "Expected outcomes," we omitted assessment of dementia and metastatic cancer, as it was not recorded in the EHR and data registry. These comorbidities have a significant association with clinical outcomes and so our adaptation may have reduced the predictive accuracy of Colorado's algorithm. Some of this limitation could be tested by sensitivity analyses of similar adaptations.

### Selection of clinical outcome (steps 4–8) and testing algorithm accuracy (steps 14–16)

Testing the predictive accuracy of triage algorithms will require the selection of clinical outcomes that are feasible in the patient cohort, such as 28 day mortality. However, these selected outcomes will often fall short of ethical goals, such as longer-term outcomes over years or complicated outcomes like activities of daily living.

### Simulation of clinical decision making (steps 17–19)

Our reductive simulation of clinical decision making in small patient groups does not address the likely rapid change in numerical combinations of scarce resources and patients, and we note that

the code underlying the simulation is open-access for modification for future studies (Figure 1). For example, there may be two ICU beds for four waiting patients at one moment, and then one ICU bed for seven waiting patients at the next moment. A more complicated code could be written that randomly varied resource and patient numbers. In addition, the bootstrap method of random sampling across an entire patient cohort assumes that patients are independent of each other. However, there may be temporal associations among patients in a cohort. For example, in "Expected outcomes," the cohort spanned March to June 2020. It is possible that patients presenting at the same time in March 2020 differed in clinical characteristics from small patient groups later in June 2020. Some studies find different clinical outcomes by time of day, week (versus weekend), or year for particular clinical scenarios (Cavallazzi et al., 2010; Patel et al., 2016; Young et al., 2011). Some of this limitation could be tested by sensitivity analyses with patients in sub cohorts by date or time of presentation.

## TROUBLESHOOTING

### Problem 1

Retrospective cohort study (step 6) and missing or erroneous data in the EHR (steps 12–14)

All retrospective cohort studies are vulnerable to bias, such as selection bias. Clinical studies typically suffer from missing or erroneous data in the clinical record.

### Potential solution

A prospective cohort is ideal, but often impractical. For a retrospective cohort study, the inclusion criteria should be established prior to initiation of the study and optimized to reduce selection bias, such as by enrolling consecutive patients rather than a convenience cohort, if possible. Similarly, protocols for missing or erroneous data should be established prior to initiation of the study. Potential solutions are discussed in greater detail in our prior STAR protocol (Crowley et al., 2021). In brief, the first consideration is whether data are missing or erroneous at random. If "missing completely at random" (MCAR), "listwise deletion" can simply remove any subjects missing critical data. In "pairwise deletion," the subject is deleted from the part of the study affected by the missing data but is included in other analyses. If data are missing at random, the missing data can be imputed by methods ranging from the mean values in other subjects ("mean substitution") or linear regression models ("regression substation") to more complex methods, like the "multiple imputation" method, in which iterations of imputation with slightly different values give estimated standard error, and others (Crowley et al., 2021). A major challenge is that data can be missing or incomplete for non-random reasons. For example, the patients with the most severe disease or unstable clinical status on hospital admission may have the least detailed admission notes and abbreviated past medical history. This non-random missing data could introduce systematic bias of missing comorbidities in the patients that presented with more severe disease earlier in their hospital course. Sensitivity and other analyses can help address potential biases. In this example, the sub cohorts of patients with more severe or less severe disease early in their hospital course could be compared. Full discussion is beyond the scope of this protocol and requires reference to more in-depth biostatical resources (MIT Critical Data, 2016).

### Problem 2

Systematic errors in clinical data in disease severity scores (steps 10, 13–15)

Focusing on the example of SOFA and MSOFA scoring, each component can have systematic errors.

**Respiratory.** The measurement of oxygen saturation can have systematic bias. $SpO_2$ measured non-invasively by a pulse oximeter on the finger or toes can falsely underestimate oxygen saturation if there is reduced peripheral circulation due to vasopressor medications, hypothermia or comorbidities such as Raynaud's or scleroderma. Patients with shock or certain comorbidities could have systematically lower $SpO_2$ readings. At low oxygen saturation, $SpO_2$ can systematically overestimate

oxygen saturation in patients with dark skin pigment (Bickler et al., 2005; Jubran and Tobin, 1990; Sjoding et al., 2020). Clinicians or institutions can have clinical practice patterns that influence the respiratory score, since $PaO_2$ and P/F ratios can reflect clinical interventions rather than underlying pathophysiology. For example, the P/F ratio can change if the clinician adjusts the end-expiratory pressure (PEEP) settings for a patient on non-invasive or mechanical ventilation. The P/F ratio can vary rapidly from moment to moment due to the patient's posture, airway secretions, or synchrony with ventilation and other factors.

**Central nervous system.** Sedating and analgesic medications can depress CNS function and confound GCS scoring. Patients intubated and on mechanical ventilation often require deeper sedation that prevents assessment of underlying CNS function. Individual or institutional differences in clinical practices regarding sedation or other factors, such as the use of awake non-invasive mask ventilation versus intubated mechanical ventilation, could systematically affect CNS scoring.

**Cardiovascular.** SOFA scoring is not optimized for current practice patterns. For example, the scale implies that first-line vasopressor is dopamine, which is no longer standard of care. Common second-line vasopressors, such as vasopressin or phenylephrine, are not part of the scoring system at all. In addition, SOFA CV scoring does not distinguish among hypotension due to distributive shock in sepsis, cardiogenic shock due to heart failure, or hypotension due to sedating medications, which may all have different prognostic implications.

**Liver.** Elevated plasma bilirubin levels do not necessarily reflect hepatic dysfunction but can reflect excess release of bilirubin by hemolysis or biliary tract obstruction.

**Renal.** Patients with severe renal failure can undergo renal replacement therapy (RRT), also known as dialysis. RRT can normalize serum creatinine values, so assessment of creatinine values without knowledge of RRT status is misleading. In addition, some patients continue to make urine but require dialysis for defective renal filtration, so urine output is also not an entirely sensitive measurement of severe renal failure.

**Coagulation.** SOFA score uses the degree of thrombocytopenia in its calculation. Thrombocytopenia correlates with disease severity in sepsis and recall that SOFA was originally designed for septic patients. However, platelet counts have not been as closely associated with other critical illnesses. Therefore, some more recent disease severity scoring systems, such as MSOFA, do not include platelet count. The platelet count can also be affected by iatrogenic processes, such as medication-induced thrombocytopenia (e.g., due to heparin or vancomycin) or RRT (i.e., dialysis).

## Potential solution

Several approaches are required to assess and limit systematic error in clinical data.

**Examination of possible confounders and other sources of error.** The case report form and electronic data queries can be designed to test confounders. For example, race, Raynaud's disease, scleroderma, other known confounders of $SpO_2$ measurements can be included in the data query. These patients can be tested for possible systematic error in $SpO_2$ that might suggest reliance on $PaO_2$, measured by arterial blood gas, rather than $SpO_2$ for these patients. Certainly, this adaptation itself would need to be assessed for introduction of biases. Similar approaches can be considered for other scenarios, such as identifying patients with hemolysis, biliary obstruction, or cholangitis for closer examination of their bilirubin values.

**Sensitivity analyses.** Sensitivity analyses may help identify and quantify sources of error. For example, the performance of triage algorithms can be examined in sub cohorts of Black and white patients. If there are performance differences by race, further sensitivity analyses can be performed

to identify which aspect of the triage algorithms (e.g., respiratory score, comorbidity scoring) drives the difference in results and whether any factor is due to errors in clinical measurement. To refer back to the previous example, a sensitivity analysis could test excluding patients with the possible confounders of the bilirubin component of the SOFA score, such as hemolysis or cholangitis.

**Adaptation of existing triage algorithms to limit the effect of data with systematic error.** In some situations, it is preferable to adapt triage algorithms to avoid problematic data. In our example in "Expected outcomes," we simplified the CNS component of the SOFA score to present or absence of altered mental status, since further adjudication of mental status was unreliable due to poor clinical detail in the medical record, the confounding effect of medications, and possible institutional and individual differences in neurological assessment and sedation practices. In addition, using different measurements of neurological function, such as the FOUR score, can address some of the institutional variation in assessing non-verbal patients or those who are receiving mechanical ventilation (Foo et al., 2019). In addition, newer adaptations to the SOFA score, such as the QUICK SOFA (qSOFA) which measures only three categories (GCS, respiratory rate, and systolic blood pressure), limit the data needed from EHRs and therefore may improve its overall integrity. In another example, investigators should pay attention to whether a triage algorithm's SOFA/MSOFA-based scoring scheme accounts for RRT, which should merit the maximal score for renal disease despite normal serum creatinine values.

**Testing of hypothetical algorithms.** In step 3 we define hypothetical triage algorithms for comparison to existing triage algorithms issued by U.S. states and other entities. These hypothetical algorithms are also an opportunity to address issues inherent in clinical data. For instance, a hypothetical algorithm could be based on SOFA scoring but omit consideration of platelet count and modify the cardiovascular scoring to reflect current vasopressor practices and different types of shock. These possible solutions illustrate the benefit of a multi-disciplinary research groups that includes expertise in both clinical and biostatistical domains.

### Problem 3

Ethical concerns with age in algorithms

We and others have shown that using age in CSC algorithms can improve their predictive abilities (Jezmir et al., 2021). However, as advances in healthcare improve the health of many elderly people, triage decisions based on years lived has become more controversial. Investigators need to be aware of the ethical and civil rights concerns when studying an existing algorithm incorporating age or when proposing a hypothetical triage algorithm.

### Potential solution

Some favor using age in a limited fashion, such as a tie-breaker. An alternative solution is to substitute frailty for age, as frailty may be a proxy for "biological age." Geriatricians have employed the frailty index to quantify the 'frailty phenotype,' or the loss of physiological reserve with age. Frailty indices measure the *degree of functional disability* from aging and comorbidities. Frailty indices associate with morbidity, such as falls and hospitalization, and also mortality (Hanlon et al., 2018; Kojima et al., 2018). Two of the most commonly used indices are the Frailty Index (Mitnitski et al., 2001), developed using data from the Canadian Study of Health and Ageing, and the Fried Index (Fried et al., 2001) from the Cardiovascular Health study This approach may reduce ethical controversy by focusing on a patient's underlying physiological condition rather than chronological age.

### Problem 4

Adjudication of complex comorbidities

A key consideration in using comorbidities in triage algorithms is whether accurate and detailed medical information will be quickly available at the bedside. Full adjudication of many complex comorbidities will likely not be practical at the bedside. For example, thorough outpatient evaluation

of a patient with pulmonary fibrosis would include interdisciplinary discussion of pulmonary function testing, imaging, and any available pathology, laboratory testing, and functional testing, such as a six minute walk test. In another example, frailty indices, a possible solution to Problem 4, can be highly complex, as the initial version of the frailty index assesses 70 deficits. Further, frailty scales may incorporate the modified mini-mental state examination, patient histories of falls, cognitive impairment and activities of daily living and other medical history (Rockwood et al., 2005). In a further ethical consideration, if assessment of comorbidities and frailty depend on self-reporting by patients and their families, then patients could be "penalized" by more honest self-reporting.

### Potential solution

For some comorbidities, objective test values can be quickly interpretable and provide imperfect but helpful information on prognosis. For example, trends and percent predicted for pulmonary function tests, such as forced vital capacity (FVC) or diffusion capacity (DLCO) associate with disease progression in idiopathic pulmonary fibrosis (IPF) and could help refine comorbidity scoring (Jegal et al., 2005; Latsi et al., 2003). In heart failure with reduced ejection fraction, left ventricular ejection fraction associates with survival (Curtis et al., 2003). EHR could, in theory, be programmed to calculate these scores automatically. Frailty indices could also be simplified, as has been studied in surgical outcomes (Karam et al., 2013). However, some patients will certainly lack these test results, and not all comorbidities or indices would lend themselves to easily automated scoring.

### Problem 5

Assessing tie-breakers (steps 16–19)

Not infrequently, triage algorithms result in patients with tied priority point totals. Results that end in ties are not necessarily "good" or "bad" results, but simply reflect a performance characteristic of the algorithm. An important value of empirical testing is revealing the frequency of tie-breakers in the "real-world," so that ethicists, clinicians, legislators and other public stakeholders can decided whether this emphasis on (or lack of) tie-breaker use is what they intended. However, ties present a challenge for empirical analysis in multiple ways. First, while some guidelines use a lottery (i.e., "coin flip") as a tie-breaker, other guidelines often incorporate patient characteristics that are not well documented in the medical record, such as status as an essential worker. Second, some analytical approaches, such as the AUROC curves in step 16, only incompletely characterize the performance of algorithms with respect to ties.

### Potential solution

For information that is inconsistently documented in the medical record, such as employment as an essential worker, there is no easy work-around. In theory, the dataset could be supplemented by re-contacting a random sampling of the cohort. More preferably, a prospectively designed study would capture this information. Problem 5 also illustrates the advantage of the simulation of clinical decision making in steps 17–19, which estimates the frequency of tied priority scores and can be modified to test tie-breakers. For example, in Jezmir et al. (2021), we modified the code to test age as a tie-breaker. Unsurprisingly, we found that age as a tie-breaker greatly reduced the need for additional tie-breakers. Further, we found that the addition of age as a tie-breaker did not change the predictive accuracy of the triage algorithms for 28 day survival.

### Problem 6

Early practical considerations

In our study Jezmir et al. (2021), two key practical considerations were critical in the earliest stages of the study.:

a. Selection of triage algorithms to test: The selection of which triage algorithms to test can be difficult and depends on the nature of the crisis (i.e., pandemic, mass casualty). In addition, developing

triage scoring can quickly become complicated with multiple categories, comorbidity indices and ranking options.

b. Interdisciplinary challenges: Analysis of CSC triage algorithms has challenges across multiple domains. For example, EHRs can be difficult to navigate to obtain data needed for the specific categories in triage algorithms. As a second example, coding the simulations may be an iterative process, and the simulation itself may need to be modified..

### Potential solution

In our study Jezmir et al., (2021), it proved critical to implement these two solutions early in our study:

a. Selection or Deciding on which triage algorithms to test: When selecting which triage algorithms to study, using algorithms validated in a similar type of crisis situation will facilitate any adaptations made during sensitivity testing. For example, if the crisis situation is mass casualty related, starting with a trauma score (such as the Injury Severity Score) can facilitate testing. For algorithm analysis, we recommend starting with analysis of the most simple scoring system, which is an easier analysis in which to do the first round of data pre-processing and other adaptations as discussed in this troubleshooting section.

b. Interdisciplinary challenges: In particular, two types of domain experts are critical early in the implementation of the study. First, design of EHR data collection and adjudication of clinical data requires a clinician experienced in the areas of medicine concerned with the crisis. For example, the team should include a critical care physician or trauma physician for critical illness or mass casualty scenarios, respectively. Second, early involvement of experienced computer programmers and informaticians will make later modification of the simulation code much easier and more efficient, particularly if new simulations need to be created.

## RESOURCE AVAILABILITY

### Lead contact

Further information and requests should be directed to the lead contact, Edy Y. Kim (ekim11@bwh.harvard.edu).

### Materials availability

There were no materials generated for this study.

### Data and code availability

The patient datasets 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 per data use agreements with each of the 67 participating STOP-COVID institutions. Summary data from STOP-COVID are available. The code referenced in this protocol is available at https://github.com/maheetha/CSC (https://doi.org/10.5281/zxenodo.5204454) as in the key resources table.

## ACKNOWLEDGMENTS

We thank William B. Feldman (Brigham and Women's Hospital, Harvard Medical School) for his valuable input.

## AUTHOR CONTRIBUTIONS

M.B. and J.L.J. developed these protocols. M.B. wrote the code. L.T.M., D.E.L., and E.Y.K. prepared this manuscript. E.Y.K. supervised this work.

## DECLARATION OF INTERESTS

The authors have no disclosures or conflicts of interest relevant to this work. In unrelated disclosures, E.Y.K. is a co-investigator in NCT04389671 (Windtree Therapeutics) testing lucinactant (surfactant-like treatment) in COVID-19 patients. E.Y.K. received unrelated research funding from Bayer, the US National Institutes of Health, the American Heart Association, the American Lung Association, and the American Thoracic Society.

## REFERENCES

Altman, M.C. (2021). A consequentialist argument for considering age in triage decisions during the coronavirus pandemic. Bioethics *35*, 356–365.

Asai, N., Ohashi, W., Sakanashi, D., Suematsu, H., Kato, H., Hagihara, M., Watanabe, H., Shiota, A., Koizumi, Y., Yamagishi, Y., et al. (2021). Combination of sequential organ failure assessment (SOFA) score and Charlson comorbidity index (CCI) could predict the severity and prognosis of candidemia more accurately than the acute physiology, age, chronic health evaluation II (APACHE II) score. BMC Infect. Dis. *21*, 77.

Beigel, J.H., Tomashek, K.M., Dodd, L.E., Mehta, A.K., Zingman, B.S., Kalil, A.C., Hohmann, E., Chu, H.Y., Luetkemeyer, A., Kline, S., et al. (2020). Remdesivir for the treatment of covid-19 - final report. N. Engl. J. Med. *383*, 1813–1826.

Bharadwaj, M., Jezmir, J.L., Kishore, S.P., Winkler, M., Diephus, B., Haider, H., Crowley, C.P., Pinilla-Vera, M., Varon, J., Baron, R.M., et al. (2021). Empirical assessment of U.S. coronavirus disease 2019 crisis standards of care guidelines. Crit. Care Explor. *3*, e0496.

Bickler, P.E., Feiner, J.R., and Severinghaus, J.W. (2005). Effects of skin pigmentation on pulse oximeter accuracy at low saturation. Anesthesiology *102*, 715–719.

Brown, S.M., Duggal, A., Hou, P.C., Tidswell, M., Khan, A., Exline, M., Park, P.K., Schoenfeld, D.A., Liu, M., Grissom, C.K., et al. (2017). Nonlinear imputation of PaO2/FiO2 from SpO2/FiO2 among mechanically ventilated patients in the ICU: A prospective, observational study. Crit. Care Med. *45*, 1317–1324.

Cavallazzi, R., Marik, P.E., Hirani, A., Pachinburavan, M., Vasu, T.S., and Leiby, B.E. (2010). Association between time of admission to the ICU and mortality: a systematic review and metaanalysis. Chest *138*, 68–75.

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.

Christensen, S., Johansen, M.B., Christiansen, C.F., Jensen, R., and Lemeshow, S. (2011). Comparison of Charlson comorbidity index with SAPS and APACHE scores for prediction of mortality following intensive care. Clin. Epidemiol. *3*, 203–211.

Cleveland Manchanda, E., Couillard, C., and Sivashanker, K. (2020). Inequity in crisis standards of care. N. Engl. J. Med. *383*, e16.

Coudroy, R., Frat, J.-P., Girault, C., and Thille, A.W. (2020). Reliability of methods to estimate the fraction of inspired oxygen in patients with acute respiratory failure breathing through non-rebreather reservoir bag oxygen mask. Thorax *75*, 805–807.

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.

Curtis, J.P., Sokol, S.I., Wang, Y., Rathore, S.S., Ko, D.T., Jadbabaie, F., Portnay, E.L., Marshalko, S.J., Radford, M.J., and Krumholz, H.M. (2003). The association of left ventricular ejection fraction, mortality, and cause of death in stable outpatients with heart failure. J. Am. Coll. Cardiol. *42*, 736–742.

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.

Desai, A., and Gyawali, B. (2020). Endpoints used in phase III randomized controlled trials of treatment options for COVID-19. EClinicalMedicine *23*, 100403.

Foo, C.C., Loan, J.J.M., and Brennan, P.M. (2019). The relationship of the FOUR score to patient outcome: A systematic review. J. Neurotrauma *36*, 2469–2483.

Frat, J.-P., Thille, A.W., Mercat, A., Girault, C., Ragot, S., Perbet, S., Prat, G., Boulain, T., Morawiec, E., Cottereau, A., et al. (2015). High-flow oxygen through nasal cannula in acute hypoxemic respiratory failure. N. Engl. J. Med. *372*, 2185–2196.

Fried, L.P., Tangen, C.M., Walston, J., Newman, A.B., Hirsch, C., Gottdiener, J., Seeman, T., Tracy, R., Kop, W.J., Burke, G., et al. (2001). Frailty in older adults: evidence for a phenotype. J. Gerontol. A Biol. Sci. Med. Sci. *56*, 146–156.

Gershengorn, H.B., Holt, G.E., Rezk, A., Delgado, S., Shah, N., Arora, A., Colucci, L.B., Mora, B., Iyengar, R.S., Lopez, A., et al. (2021). Assessment of disparities associated with a crisis standards of care resource allocation algorithm for patients in 2 US hospitals during the COVID-19 pandemic. JAMA Netw. Open *4*, e214149.

Grissom, C.K., Brown, S.M., Kuttler, K.G., Boltax, J.P., Jones, J., Jephson, A.R., and Orme, J.F. (2010). A modified sequential organ failure assessment score for critical care triage. Disaster Med. Public Health Prep. *4*, 277–284.

Hanlon, P., Nicholl, B.I., Jani, B.D., Lee, D., McQueenie, R., and Mair, F.S. (2018). Frailty and pre-frailty in middle-aged and older adults and its association with multimorbidity and mortality: a prospective analysis of 493737UK Biobank participants. Lancet Public Health *3*, e323–e332.

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.

Hope, A.A., Gong, M.N., Guerra, C., and Wunsch, H. (2015). Frailty before critical illness and mortality for elderly medicare beneficiaries. J. Am. Geriatr. Soc. *63*, 1121–1128.

Iacorossi, L., Fauci, A.J., Napoletano, A., D'Angelo, D., Salomone, K., Latina, R., Coclite, D., and Iannone, P. (2020). Triage protocol for allocation of critical health resources during Covid-19 pandemic and public health emergencies. A narrative review. Acta Biomed. *91*, e2020162.

Jegal, Y., Kim, D.S., Shim, T.S., Lim, C.-M., Do Lee, S., Koh, Y., Kim, W.S., Kim, W.D., Lee, J.S., Travis, W.D., et al. (2005). Physiology is a stronger predictor of survival than pathology in fibrotic interstitial pneumonia. Am. J. Respir. Crit. Care Med. *171*, 639–644.

Jezmir, J.L., Bharadwaj, M., Chaitoff, A., Diephuis, B., Crowley, C.P., Kishore, S.P., Goralnick, E., Merriam, L.T., Milliken, A., Rhee, C., et al. (2021). Performance of crisis standards of care guidelines in a cohort of critically ill COVID-19 patients in the United States. Cell Rep. Med. *2*, 100376.

Jubran, A., and Tobin, M.J. (1990). Reliability of pulse oximetry in titrating supplemental oxygen therapy in ventilator-dependent patients. Chest *97*, 1420–1425.

Karam, J., Tsiouris, A., Shepard, A., Velanovich, V., and Rubinfeld, I. (2013). Simplified frailty index to predict adverse outcomes and mortality in vascular surgery patients. Ann. Vasc. Surg. *27*, 904–908.

Kojima, G., Iliffe, S., and Walters, K. (2018). Frailty index as a predictor of mortality: a systematic review and meta-analysis. Age Ageing *47*, 193–200.

Latsi, P.I., du Bois, R.M., Nicholson, A.G., Colby, T.V., Bisirtzoglou, D., Nikolakopoulou, A., Veeraraghavan, S., Hansell, D.M., and Wells, A.U. (2003). Fibrotic idiopathic interstitial pneumonia: the prognostic value of longitudinal functional trends. Am. J. Respir. Crit. Care Med. *168*, 531–537.

MIT Critical Data (2016). Secondary Analysis of Electronic Health Records (Springer).

Mitnitski, A.B., Mogilner, A.J., and Rockwood, K. (2001). Accumulation of deficits as a proxy measure of aging. Scientific World Journal *1*, 323–336.

Patel, R., Chesney, E., Cullen, A.E., Tulloch, A.D., Broadbent, M., Stewart, R., and McGuire, P. (2016). Clinical outcomes and mortality associated with weekend admission to psychiatric hospital. Br. J. Psychiatry *209*, 29–34.

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.

Rockwood, K., Song, X., MacKnight, C., Bergman, H., Hogan, D.B., McDowell, I., and Mitnitski, A. (2005). A global clinical measure of fitness and frailty in elderly people. Can. Med. Assoc. J. *173*, 489–495.

Rueda, J. (2021). Ageism in the COVID-19 pandemic: age-based discrimination in triage decisions and beyond. Hist. Philos. Life Sci. *43*, 91.

Sjoding, M.W., Dickson, R.P., Iwashyna, T.J., Gay, S.E., and Valley, T.S. (2020). Racial bias in pulse oximetry measurement. N. Engl. J. Med. *383*, 2477–2478.

Sundararajan, V., Henderson, T., Perry, C., Muggivan, A., Quan, H., and Ghali, W.A. (2004). New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. J. Clin. Epidemiol. *57*, 1288–1294.

Vincent, J.-L., and Moreno, R. (2010). Clinical review: scoring systems in the critically ill. Crit. Care *14*, 207.

Vincent, J.L., Moreno, R., Takala, J., Willatts, S., Mendona, A.D., 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. Intensive Care Med. *22*, 707–710.

Young, J.Q., Ranji, S.R., Wachter, R.M., Lee, C.M., Niehaus, B., and Auerbach, A.D. (2011). July effect": impact of the academic year-end changeover on patient outcomes: a systematic review. Ann. Intern. Med. *155*, 309–315.