OPEN

Overt and Occult Hypoxemia in Patients Hospitalized With COVID-19

Shrirang M. Gadrey, MBBS, MPH[1]
Piyus Mohanty, MS[3]
Sean P. Haughey, MD[1]
Beck A. Jacobsen, MD[1]
Kira J. Dubester, MD[1]
Katherine M. Webb, MD[1]
Rebecca L. Kowalski, MD[1]
Jessica J. Dreicer, MD[1]
Robert T. Andris, MS[1,2]
Matthew T. Clark, PhD[2,4]
Christopher C. Moore, MD[1,2]
Andre Holder, MD, MSc[3]
Rishi Kamaleswaran, PhD[3]
Sarah J. Ratcliffe, PhD[1,2]
J. Randall Moorman, MD[1,2]

IMPORTANCE: Progressive hypoxemia is the predominant mode of deterioration in COVID-19. Among hypoxemia measures, the ratio of the $\mathrm{Pa}{\scriptstyle\mathrm{O}}_2$ to the $\mathrm{F}{\scriptstyle\mathrm{IO}}_2$ (P/F ratio) has optimal construct validity but poor availability because it requires arterial blood sampling. Pulse oximetry reports oxygenation continuously (ratio of the $\mathrm{Sp}{\scriptstyle\mathrm{O}}_2$ to the $\mathrm{F}{\scriptstyle\mathrm{IO}}_2$ [S/F ratio]), but it is affected by skin color and occult hypoxemia can occur in Black patients. Oxygen dissociation curves allow noninvasive estimation of P/F ratios (ePFRs) but remain unproven.

OBJECTIVES: Measure overt and occult hypoxemia using ePFR.

DESIGN, SETTING, AND PARTICIPANTS: We retrospectively studied COVID-19 hospital encounters ($n = 5{,}319$) at two academic centers (University of Virginia [UVA] and Emory University).

MAIN OUTCOMES AND MEASURES: We measured primary outcomes (death or ICU transfer within 24 hr), ePFR, conventional hypoxemia measures, baseline predictors (age, sex, race, comorbidity), and acute predictors (National Early Warning Score [NEWS] and Sequential Organ Failure Assessment [SOFA]). We updated predictors every 15 minutes. We assessed predictive validity using adjusted odds ratios (AORs) and area under the receiver operating characteristic curves (AUROCs). We quantified disparities (Black vs non-Black) in empirical cumulative distributions using the Kolmogorov-Smirnov (K-S) two-sample test.

RESULTS: Overt hypoxemia (low ePFR) predicted bad outcomes (AOR for a 100-point ePFR drop: 2.7 [UVA]; 1.7 [Emory]; $p < 0.01$) with better discrimination (AUROC: 0.76 [UVA]; 0.71 [Emory]) than NEWS (0.70 [both sites]) or SOFA (0.68 [UVA]; 0.65 [Emory]) and similar to S/F ratio (0.76 [UVA]; 0.70 [Emory]). We found racial differences consistent with occult hypoxemia. Black patients had better apparent oxygenation (K-S distance: 0.17 [both sites]; $p < 0.01$) but, for comparable ePFRs, worse outcomes than other patients (AOR: 2.2 [UVA]; 1.2 [Emory]; $p < 0.01$).

CONCLUSIONS AND RELEVANCE: The ePFR was a valid measure of overt hypoxemia. In COVID-19, it may outperform multi-organ dysfunction models. By accounting for biased oximetry as well as clinicians' real-time responses to it (supplemental oxygen adjustment), ePFRs may reveal racial disparities attributable to occult hypoxemia.

KEY WORDS: COVID-19; hospital mortality; organ dysfunction scores; prognosis; respiratory failure

Modeling the risk of adverse outcomes from COVID-19 has been an area of intense investigation. Two recent systematic reviews identified over 200 new models, nearly half of which modeled risk of adverse outcomes (clinical deterioration, critical illness, or mortality) (1, 2). We reviewed the predictors that were reported as being useful in these reviews and seven subsequent studies (3–9). Since progressive hypoxemia is the predominant mode of deterioration in COVID-19, we expected hypoxemia markers to

DOI: 10.1097/CCE.0000000000000825

KEY POINTS

Question: Can we improve on the standard Pao$_2$ to Fio$_2$ ratio (P/F ratio) for oximetry-based detection of hypoxemia in COVID-19, especially in Black patients?

Findings: In this multicenter retrospective cohort study of 5,319 hospital encounters for COVID-19, we found that a new, simple algorithm for noninvasive, oximetry-based estimation of the P/F ratio was superior to other operational markers of hypoxemia in at least one domain of performance (availability, construct validity, predictive validity, and ability to characterize racial disparities) and was noninferior in all other domains.

Meaning: The P/F ratio estimated using the oxygen dissociation curve is an improved operational marker of hypoxemia for applications like clinical research, real-time predictive modeling and post-marketing surveillance for bias in pulse oximetry devices.

be the strongest predictors. Hypoxemia markers, however, predicted outcomes in only seven models (< 10%) (3, 6, 8, 10–13). This points to an opportunity to improve the hypoxemia markers used in clinical practice and research.

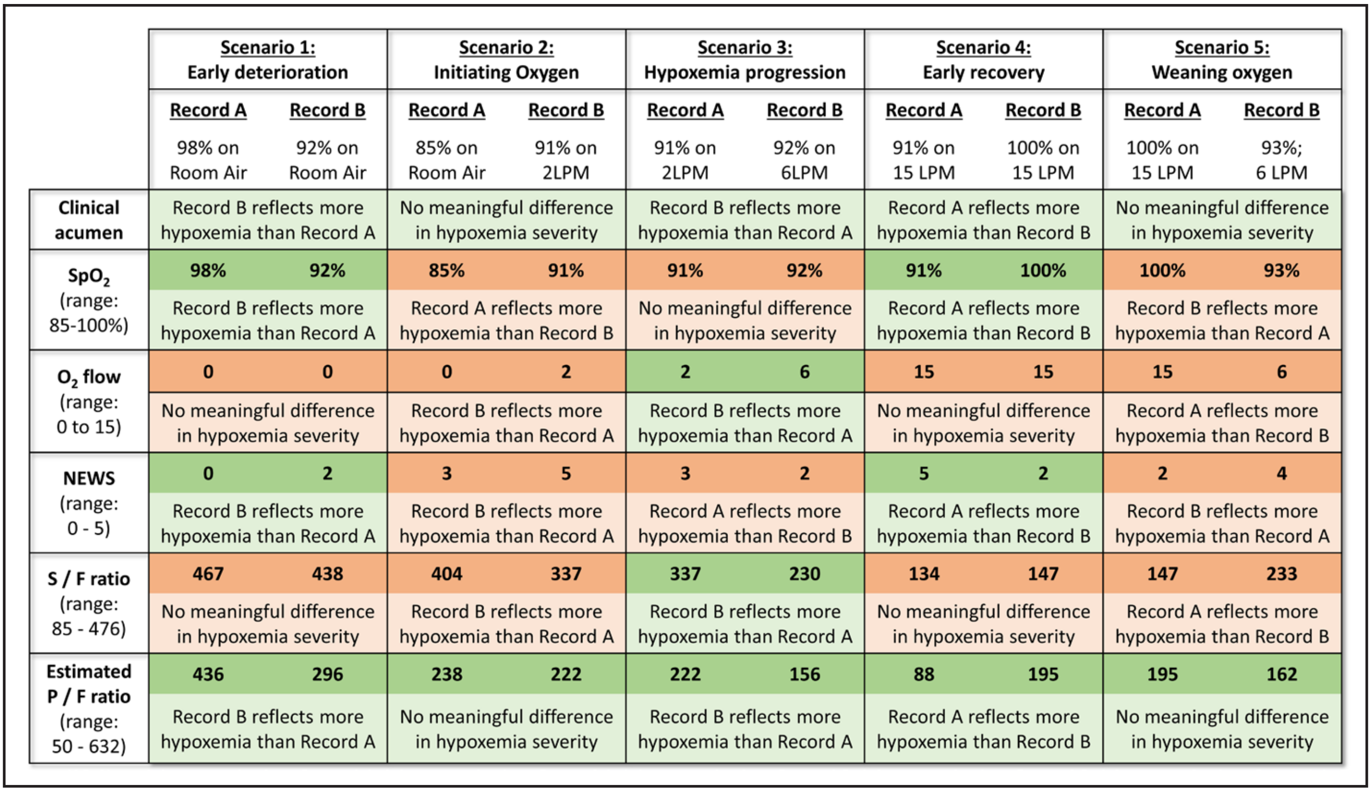
The most commonly featured hypoxemia markers were the oxygen saturation of binding sites of hemoglobin from pulse oximetry (Spo$_2$, %) and the oxygen flow rate (L/min). Most models only used Spo$_2$, without regard to oxygen supplementation (3, 10–12, 14). This approach loses power when patients with differing oxygen supplementation levels are compared (**Fig. 1**; scenarios 2, 3, 5). It is also affected by practice patterns like Spo$_2$ targets and promptness of weaning supplemental oxygen. The National Early Warning Score (NEWS) models include oxygen supplementation, but in a binary form where two points are assigned for supplemental oxygen use, regardless of the flow rate. The resulting scores do not always reflect severity of hypoxemia (Fig. 1; scenarios 2, 3, and 5).

The ratio of the Pao$_2$ (mm Hg) to the Fio$_2$ (no units) (P/F ratio) does not suffer from these drawbacks. We found only two models that include it—Sepsis-3 and Toward a COVID-19 Score (TACS) (13, 17). In both cases, Pao$_2$ is measured on arterial blood gas (ABG) samples. When Pao$_2$ was unavailable, the Sepsis-3 researchers used multiple imputation with chained equations and the TACS researchers imputed a P/F ratio of 381 (assuming Pao$_2$ at 80 and Fio$_2$ at 0.21 [room air]). However, as the proportion of missing data increases, these imputation methods become increasingly more unreliable (18). Outside the ICU, ABGs are missing in over 75% of cases (19, 20). It is not surprising, therefore, that the only models that used the measured P/F ratio were derived in the ICU.

The ratio of the Spo$_2$ to the Fio$_2$ (S/F ratio) has been used (7), but its construct validity is limited. The Spo$_2$ range (typically 85–100%) is narrower than the corresponding Pao$_2$ range (50–130 mm Hg). Thus, Fio$_2$ settings play a larger role in the S/F ratio than in the P/F ratio (Fig. 1: rows 3 and 5 agree in all scenarios, rows 3 and 6 do not). Additionally, the S/F ratio sidesteps the fact that the relationship between Pao$_2$ and Spo$_2$ is not a straight line. Judging hypoxemia severity using S/F ratios can, therefore, be misleading (Fig. 1; scenarios 1, 2, 4, and 5).

To allow noninvasive estimation of P/F ratios, we derived a new oxygen dissociation curve model from a cohort of hospitalized, nonintubated patients with simultaneous ABG and pulse oximetry recordings (21). Older models were derived from laboratory solutions of hemoglobin (22, 23) or whole blood specimens of a few young, healthy males (24). They underestimated the severity of hypoxemia (i.e., overestimated Pao$_2$) when applied to a hospitalized patient population, which is much older, more diversity (age, sex), and has higher comorbidity burden than average. In these high-risk patients, underestimation of hypoxemia can be a catastrophic mistake since early warnings can often trigger potentially life-saving responses. The newer model provided better estimates of hypoxemia in hospitalized patients (21). The P/F ratios estimated using this model (ePFRs) have high construct validity in all scenarios (Fig. 1). We hypothesized that ePFRs are a valid measure of overt hypoxemia. If so, clinicians might use the ubiquitous Spo$_2$ to monitor ePFRs continuously without being limited by arterial blood draws.

The relationship between Spo$_2$ readings and arterial oxygen saturation (Sao$_2$) is complicated. Pulse oximetry often overestimates arterial oxygenation, especially in darker-skinned individuals (25–33). One study showed a racial bias in pulse oximetry readings,



	Scenario 1: Early deterioration		Scenario 2: Initiating Oxygen		Scenario 3: Hypoxemia progression		Scenario 4: Early recovery		Scenario 5: Weaning oxygen	
	Record A	Record B	Record A	Record B	Record A	Record B	Record A	Record B	Record A	Record B
	98% on Room Air	92% on Room Air	85% on Room Air	91% on 2LPM	91% on 2LPM	92% on 6LPM	91% on 15 LPM	100% on 15 LPM	100% on 15 LPM	93%; 6 LPM
Clinical acumen	Record B reflects more hypoxemia than Record A		No meaningful difference in hypoxemia severity		Record B reflects more hypoxemia than Record A		Record A reflects more hypoxemia than Record B		No meaningful difference in hypoxemia severity	
SpO_2 (range: 85-100%)	98%	92%	85%	91%	91%	92%	91%	100%	100%	93%
	Record B reflects more hypoxemia than Record A		Record A reflects more hypoxemia than Record B		No meaningful difference in hypoxemia severity		Record A reflects more hypoxemia than Record B		Record B reflects more hypoxemia than Record A	
O_2 flow (range: 0 to 15)	0	0	0	2	2	6	15	15	15	6
	No meaningful difference in hypoxemia severity		Record B reflects more hypoxemia than Record A		Record B reflects more hypoxemia than Record A		No meaningful difference in hypoxemia severity		Record A reflects more hypoxemia than Record B	
NEWS (range: 0 - 5)	0	2	3	5	3	2	5	2	2	4
	Record B reflects more hypoxemia than Record A		Record B reflects more hypoxemia than Record A		Record A reflects more hypoxemia than Record B		Record A reflects more hypoxemia than Record B		Record B reflects more hypoxemia than Record A	
S / F ratio (range: 85 - 476)	467	438	404	337	337	230	134	147	147	233
	No meaningful difference in hypoxemia severity		Record B reflects more hypoxemia than Record A		Record B reflects more hypoxemia than Record A		No meaningful difference in hypoxemia severity		Record A reflects more hypoxemia than Record B	
Estimated P / F ratio (range: 50 - 632)	436	296	238	222	222	156	88	195	195	162
	Record B reflects more hypoxemia than Record A		No meaningful difference in hypoxemia severity		Record B reflects more hypoxemia than Record A		Record A reflects more hypoxemia than Record B		No meaningful difference in hypoxemia severity	

Figure 1. Evaluation of the construct validity of operational markers of hypoxemia in hypothetical clinical scenarios. Construct validity of any marker of hypoxemia is the extent to which that marker accurately reflects the clinical construct of hypoxemia. This figure examines the construct validity of five operational markers of hypoxemia (*rows*) in common clinical scenarios (*columns*). In each scenario (*column*), two records of a patient's oxygenation are compared (record A on *left*, record B on *right*). The first *row* titled "clinical acumen" describes a clinically sensible conclusion that a clinician might draw by comparing the two records. For example, in scenario 2, a clinician will likely conclude that the two records do not represent any meaningful change in the severity of hypoxemic respiratory failure (*row 1*, *column 2*). Rather, record B (oxygen saturation from pulse oximetry [Spo_2] of 91% on 2 L/min [LPM] of oxygen) might simply reflect the fact that a clinician initiated supplemental oxygen in response to record A (Spo_2 of 85% on room air). Each of the subsequent *rows* describes the conclusion based solely on comparing a particular marker of hypoxemia. For example, if one solely compared Spo_2 in scenario 2 (*row* 2, *column* 2), the conclusion would be that record A reflects significantly more severe hypoxemia than record B (Spo_2 of 85% vs 91%). Considering the varying range of each marker, we used the following cutoffs to determine a "significantly more/less hypoxemia": any difference greater than or equal to 1 for National Early Warning Score (NEWS) (range, 0–5), any difference greater than or equal to 2 for Spo_2 (range, 85–100) and supplemental oxygen flow rate (range, 0–15 LPM), and any difference greater than or equal to 50 for ratio of Spo_2/Fio_2 (S/F ratio) (range, 85–476) and ratio of Pao_2/Fio_2 (P/F ratio) (range, 50–632). A cell is *shaded green* when there is agreement between the marker of hypoxemia and clinical acumen, and it is *shaded red* when there is disagreement. This figure illustrates the advantages of estimated P/F ratios over other markers—it is the only marker to agree with clinical acumen in all scenarios. We were unable to conceptualize any scenario where P/F ratio would be inferior to other markers.

which led to "occult hypoxemia" (undiagnosed arterial desaturation) at three times the frequency in Black patients compared with White patients (27). Another study showed that even in the absence of bias, occult hypoxemia was more frequent among darker-skinned individuals due to a lower precision of oximetry readings (28). Occult hypoxemia may have deleterious effects on outcomes of darker-skinned individuals.

The ideal method to model the impact of occult hypoxemia on outcomes is unclear. Comparisons between simultaneously recorded Spo_2 (pulse oximetry) and Sao_2 (ABG) are limited by their exclusion of the majority of patients in whom arterial blood sampling is unavailable. Studying the population-wide distributions of Spo_2 may not be an appropriate alternative because these distributions are influenced by clinicians' real-time efforts to maintain Spo_2 in a particular range (typically 90–94%) by adjusting patients' supplemental oxygen settings. The ePFR overcomes this barrier by simultaneously accounting for any falsely reassuring pulse oximetry readings (the corresponding Pao_2 estimate) as well as clinicians' real-time responses to that false reassurance (lower F_{IO_2} setting). We therefore hypothesized that comparing population-wide distributions of ePFR by race would reveal occult hypoxemia and allow better modeling of its impact on clinical outcomes.

MATERIALS AND METHODS

Data Collection

We identified a retrospective cohort of adults (age ≥ 18 yr) with hospital encounters (emergency department [ED] visit and/or hospital admission) for acute COVID-19 at the University of Virginia (UVA) Medical Center, an academic tertiary-care center. We identified 1,172 instances where the first positive severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) test occurred in the context of a hospital encounter. Only the first positive test was used. We excluded: 1) nine encounters that lacked any vitals, tests, or notes; 2) 17 encounters where chart reviews showed that the timing of the SARS-CoV-2 infection did not match the hospital encounter (usually patients whose first positive test in our record was deemed to be a persistently positive test after a resolved infection at another facility); and 3) 46 encounters where the ICU admission and/or mortality occurred within 4 hours of encounter start time (which was necessary in the primary analysis because we censored data 4 hr prior to time of outcome). The final cohort consisted of 1,100 encounters in the first year of UVA's pandemic experience (March 2020 to February 2021).

To ensure reproducibility of findings in diverse populations, we studied similar encounters (March 2020 to December 2021) at two hospitals affiliated with the Emory University: the Emory University Hospital (EUH) and Emory University Hospital Midtown (EUH-M). While UVA serves a rural and predominantly White population, the Emory sites serve an urban and predominantly Black population. While UVA and EUH are university hospitals, EUH-M is a community-based academic hospital. The Emory sites had 12,784 COVID-19 hospital encounters by December 2021. We randomly sampled a third of these encounters ($n = 4{,}219$). This ensured that the Emory dataset represented more phases of the pandemic than the UVA dataset.

At UVA, we manually reviewed all charts to: 1) confirm acute COVID-19; 2) separate preinfection baseline Sequential Organ Failure Assessment (SOFA) from acute SOFA (**eTable 1**, http://links.lww.com/CCX/B110); and 3) ascertain the Charlson Comorbidity Index (CCI) (**eTable 2**, http://links.lww.com/CCX/B110). Seven of the authors (S.M.G., S.P.H., B.A.J., K.J.D., K.M.W., J.J.D., R.K.) were the reviewers. This procedure was not repeated in the Emory data. We queried the data warehouse to record: 1) baseline risk predictors (age, sex, race, height, weight, CCI); 2) all components of ePFR, S/F ratio, SOFA score, and NEWS (**eTable 3**, http://links.lww.com/CCX/B110; **eFig. 1**, http://links.lww.com/CCX/B110); and 3) the time of transfer to ICU and/or death. We used Admit-Discharge-Transfer patient location data to determine the time of transfer to the ICU. We included only the first transfer to ICU in patients who had multiple ward-to-ICU transfers.

We used the following oxygen-hemoglobin dissociation model to calculate ePFRs:

$$\mathrm{PaO_2} = \left(\frac{23{,}400}{\frac{1}{\mathrm{SpO_2}} - 0.99} \right)^{\frac{1}{3}}$$

To ensure adequate inter-rater reliability (IRR) of manually abstracted variables, we followed best practices including clear operational definitions and standardized abstraction forms (34). For data entry, we used Research Electronic Data Capture hosted at the UVA (35, 36). To measure IRR, we randomly sampled 10% of each reviewer's charts and conducted blinded second reviews on those charts. Our IRR metrics were percent agreement and Krippendorf's alpha (37). We prespecified adequate reliability as a) alpha greater than or equal to 0.8 or b) 0.8 greater than alpha greater than or equal to 0.67 with agreement greater than or equal to 90%.

Characterizing the Risk of Clinical Deterioration Associated With Overt Hypoxemia

The primary outcome of interest was clinical deterioration, defined as transfer to an ICU or in-hospital mortality. We validated the ePFR as a measure of overt hypoxemia in two ways. First, we calculated adjusted odds ratio (AOR) to determine the extent to which ePFR was independently associated with clinical deterioration. Second, we measured the rise in model discrimination when the ePFR was added to a baseline risk model. The baseline risk model included age, sex, race, and CCI. At UVA, the baseline risk model additionally included the baseline SOFA from chart reviews. For comparison, we measured the rise in model discrimination associated with addition of conventional hypoxemia measures (Spo_2, oxygen flow rate, and S/F ratio) and multiple system organ dysfunction scores (NEWS and SOFA) to the same baseline model.

Characterizing Racial Disparities Attributable to Occult Hypoxemia From Pulse Oximetry

To characterize the influence of skin color on predictive validity of pulse oximetry-based hypoxemia measures (ePFR, S/F ratio, and Spo_2), we used race as a surrogate for skin color and compared patients whose medical records indicate their race to be Black with all other patients (25, 27). We computed empirical cumulative distribution functions (ECDFs) for each measure and quantified racial differences (Black vs non-Black). We also visualized, by race, the relationship between the hypoxemia measure and the risk of imminent clinical deterioration. We used AOR to quantify the influence of race on this relationship.

Statistical Analyses

To calculate AOR for ePFR, we used logistic regression and adjusted for all nonhypoxemia components of the NEWS and SOFA models (temperature, heart rate, respiratory rate, mean arterial pressure, Glasgow Coma Scale, creatinine, platelet count, total bilirubin). In this model, we used all explanatory variables in a continuous form, rather than the categorical form prescribed in NEWS and SOFA. To study the rise in model discrimination, we compared area under the receiver operating characteristic curves (AUROCs) using DeLong test (38). To assess differences in ECDFs, we used the Kolmogorov-Smirnov (K-S) two-sample test. To calculate the AOR for race, we used logistic regression.

Each variable was assumed to be at the preinfection baseline for all rows until the first available value of that variable. The preinfection baseline was determined manually in the UVA cohort (eTable 1, http://links.lww.com/CCX/B110) and assumed to be normal in the Emory cohort. We updated predictors every 15 minutes from encounter start time. In the absence of new data, nursing flow sheet variables (e.g., vital signs, mental state assessments, and supplemental oxygen settings) were carried forward for 12 hours (the typical nursing shift), and laboratory values were carried forward for 24 hours (the typical frequency of phlebotomy in acute illness). For any data that were still missing, we used complete records in primary analysis and median imputation as a secondary sensitivity analysis. We censored data 4 hours prior to the time of outcome. In the primary analysis, all regression models were trained to predict occurrence of primary outcome within 24 hours. The regression models did not use data that were recorded either after the outcome or less than 4 hours before the outcome (i.e., the time of censoring). We used the Huber-White method for robust SE to correct for correlation from repeated measures. We specified statistical significance as *p* value of less than 0.05.

Sensitivity Analyses in UVA Data and Other Details

We tested the impact of restricted cubic splines (3 knots) for temperature, heart rate, respiratory rate, and mean arterial pressure, since either extreme of these vital signs are associated with clinical deterioration. We assessed the impact on the estimated predictive validity of the ePFR of excluding the patients who: 1) died without transfer to ICU; and 2) were discharged without outcome in less than 24 hours (most likely to be ED visits and brief observation stays). For our primary analysis, we used a prediction horizon of 24 hours. We repeated the analysis for 3-, 5-, 7-, and 14-day horizons. We varied the censoring from 4 hours before to the time of outcome in a secondary analysis. We also varied our missing data handling strategy (median imputation instead of complete cases). We compared Black to White patients (as opposed to Black vs non-Black comparison in primary analysis).

We used R Version 3.5.1 (The R Foundation for Statistical Computing, Vienna, Austria) to perform all analyses (39). The UVA and Emory Institutional Review Boards approved the study (Protocol 22246 at UVA [March 20, 2020]; Study-00000302 at Emory [March 23, 2021]). All study procedures were in accordance with the ethical standards of these boards and with the Helsinki Declaration of 1975.

RESULTS

Cohort Characteristics and Inter-Rater Reliability of Chart Reviews

At UVA, we analyzed 399,797 every 15-minute rows (1,100 individuals) and the primary outcome occurred in 177 patients (17%). At Emory, we analyzed 1,510,070 every 15-minute rows (4,219 individuals) and the primary outcome occurred in 791 patients (19%). The probability that a random row was followed by the outcome within 24 hours was 1.9% at UVA and 2.9% at Emory. The demographic and clinical cohort characteristics are outlined in **Table 1**. Most noteworthy differences were seen in the racial composition (higher proportion of Black patients at Emory) and comorbidity (higher CCI at Emory).

Of the manually abstracted data, agreement was 79% for CCI and 95% for baseline SOFA; alpha was 0.84 for CCI and 0.90 for baseline SOFA. This met our prespecified IRR threshold.

Risk of Clinical Deterioration Associated With Overt Hypoxemia

Overt hypoxemia, operationalized using ePFR, independently predicted clinical deterioration within 24 hours (AOR: 0.990 [UVA; 95% CI, 0.984–0.996], 0.995 [Emory; 95% CI, 0.993–0.997]; $p < 0.01$ at both sites). Adding ePFR to the baseline risk model resulted in model discrimination (AUROC: 0.76 [UVA]; 0.71 [Emory]) that was better than Spo$_2$ (AUROC: 0.65 [UVA]; 0.66 [Emory]), oxygen flow rate (AUROC: 0.73 [UVA]; 0.69 [Emory]) and comparable to S/F ratio (AUROC: 0.76 [UVA]; 0.70 [Emory]). At both sites, ePFR outperformed NEWS (AUROC: 0.70 [UVA]; 0.70 [Emory]) and SOFA (AUROC: 0.68 [UVA]; 0.65 [Emory]) (**Fig. 2**).

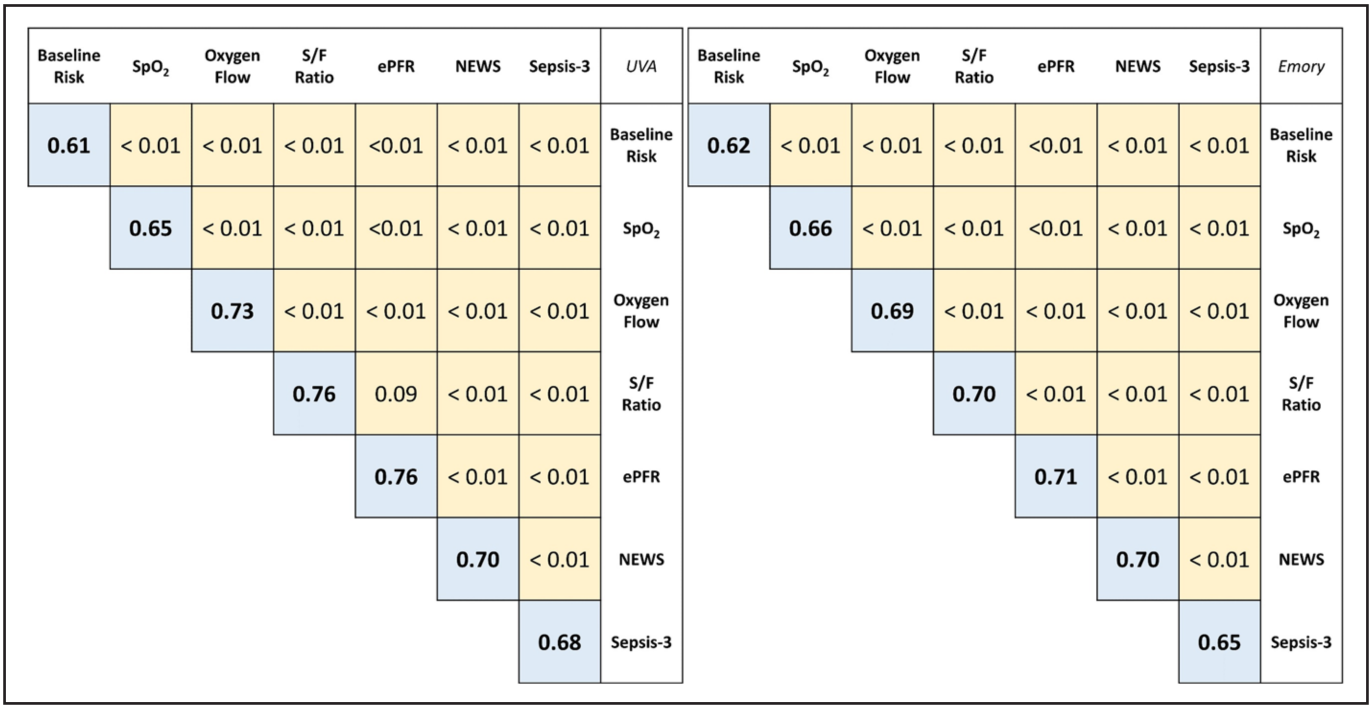
Racial Disparities Attributable to Occult Hypoxemia From Pulse Oximetry

For all hypoxemia measures (Spo$_2$, S/F ratio, and ePFR) and for both sites (UVA and Emory), we observed that the ECDF were "right-shifted" in Black patients relative to non-Black patients; that is, Black patients

TABLE 1.
Cohort characteristics

	University of Virginia Cohort		Emory Cohort	
Clinical Variable	All Patients (1,100)	Outcome Positive (177)	All Patients (4,219)	Outcome Positive (791)
Age, yr, median (interquartile range)	55 (38–68)	67 (57–77)	55 (39–68)	64 (52–75)
Male, *n* (%)	545 (50)	101 (57)	2,016 (48)	453 (57)
Race/ethnicity, *n* (%)				
White, non-Hispanic	446 (40)	89 (50)	987 (23)	215 (27)
Black	320 (29)	54 (31)	2,515 (60)	422 (54)
Hispanic	285 (26)	31 (17)	327 (8)	64 (8)
Other	49 (5)	3 (2)	390 (9)	90 (11)
Charlson Comorbidity Index, *n* (%)				
0	526 (48)	49 (28)	1,713 (41)	116 (15)
1–2	299 (27)	55 (31)	1,701 (40)	320 (40)
≥ 3	275 (25)	73 (41)	805 (19)	355 (45)
Baseline Sequential Organ Failure Assessment, *n* (%)				
0	722 (66)	76 (43)	NA	
1–2	271 (24)	63 (36)	NA	
≥ 3	107 (10)	38 (21)	NA	
In-hospital mortality, *n* (%)	49 (5)	49 (28)	240 (6)	240 (30)
ICU transfer, *n* (%)	161 (15)	161 (91)	694 (16)	694 (88)
Composite outcome, *n* (%)	177 (17)	177 (100)	791 (19)	791 (100)

NA = not available.



Baseline Risk	SpO_2	Oxygen Flow	S/F Ratio	ePFR	NEWS	Sepsis-3	*UVA*
0.61	< 0.01	< 0.01	< 0.01	<0.01	< 0.01	< 0.01	Baseline Risk
	0.65	< 0.01	< 0.01	<0.01	< 0.01	< 0.01	SpO_2
		0.73	< 0.01	< 0.01	< 0.01	< 0.01	Oxygen Flow
			0.76	0.09	< 0.01	< 0.01	S/F Ratio
				0.76	< 0.01	< 0.01	ePFR
					0.70	< 0.01	NEWS
						0.68	Sepsis-3

Baseline Risk	SpO_2	Oxygen Flow	S/F Ratio	ePFR	NEWS	Sepsis-3	*Emory*
0.62	< 0.01	< 0.01	< 0.01	<0.01	< 0.01	< 0.01	Baseline Risk
	0.66	< 0.01	< 0.01	<0.01	< 0.01	< 0.01	SpO_2
		0.69	< 0.01	< 0.01	< 0.01	< 0.01	Oxygen Flow
			0.70	< 0.01	< 0.01	< 0.01	S/F Ratio
				0.71	< 0.01	< 0.01	ePFR
					0.70	< 0.01	NEWS
						0.65	Sepsis-3

Figure 2. Discrimination of estimated ratio of Pao_2/Fio_2 (P/F ratio) for clinical deterioration in patients with COVID-19. This figure compares the area under the receiver operating characteristic curve (AUROC) of multivariable logistic regression models for clinical deterioration (transfer to ICU or mortality within 24 hr) from COVID-19. The *blue boxes* show the AUROC for a model and the *yellow boxes* show *p* values from pairwise comparison (DeLong test). Results from University of Virginia (UVA) are on the *left* and those from Emory are on the *right*. The baseline risk model used age, sex, race, Charlson Comorbidity Index, and preinfection baseline Sequential Organ Failure Assessment (SOFA) score as predictors (baseline SOFA was only available at UVA). The model for each criterion was created by adding that criterion to the baseline risk predictors. The estimated P/F ratio (ePFR) had optimal model discrimination, and it outperformed National Early Warning Score (NEWS) and SOFA (acute rise in SOFA score at UVA and total SOFA in Emory) models. S/F ratio = ratio of oxygen saturation from pulse oximetry/Fio_2, Spo_2 = oxygen saturation from pulse oximetry.

appeared to have better oxygenation (higher Spo_2, S/F ratios, and ePFR) than non-Black patients. Yet, Black patients had worse outcomes for comparable degrees of apparent oxygenation (**Fig. 3**; and **eFig. 2**, http://links.lww.com/CCX/B110).

In two important ways, this racial disparity was better revealed by ePFR and S/F ratio than by Spo_2. First, the Spo_2 distribution showed a narrower right shift (K-S distance: 0.09 [UVA], 0.15 [Emory]; $p < 0.01$) than was revealed by the S/F ratio and ePFR distributions (K-S distance: 0.17 [UVA and Emory]; $p < 0.01$). Second, a racial influence on relationship between overt hypoxemia and outcomes (i.e., evidence of occult hypoxemia) was revealed much better by ePFR and S/F ratio than by Spo_2. At UVA, when we modeled clinical deterioration using race, Spo_2, and other baseline predictors (age, sex, and comorbidity), race was not found to be a significant predictor ($p = 0.14$). In contrast, when Spo_2 was replaced by ePFR or S/F ratio in the model, race was a strong predictor (AOR, 2.2–2.3; $p < 0.01$). Similarly, in the Emory data, race was a stronger predictor when clinical deterioration was modeled with ePFR or S/F ratio (AOR, 1.20; $p < 0.01$) than with Spo_2 (AOR, 1.04; $p < 0.01$).

Sensitivity Analyses at UVA

Repeating the analysis with restricted cubic splines for temperature, respiratory rate, heart rate, and mean arterial pressure did not significantly affect the predictive validity of the ePFR. When we extended the prediction horizon, the ePFR continued to outperform NEWS and SOFA (**eFig. 3**, http://links.lww.com/CCX/B110). The results were not meaningfully impacted by: 1) excluding patients who died without transfer to ICU; 2) excluding patients who were discharged without outcome in less than 24 hours (most likely to be ED visits and brief observation stays); and 3) varying of censoring time. When we compared Black patients with White patients (instead of non-Black patients in primary analysis), we observed a similar disparity as was observed in the primary analysis (Fig. 3).

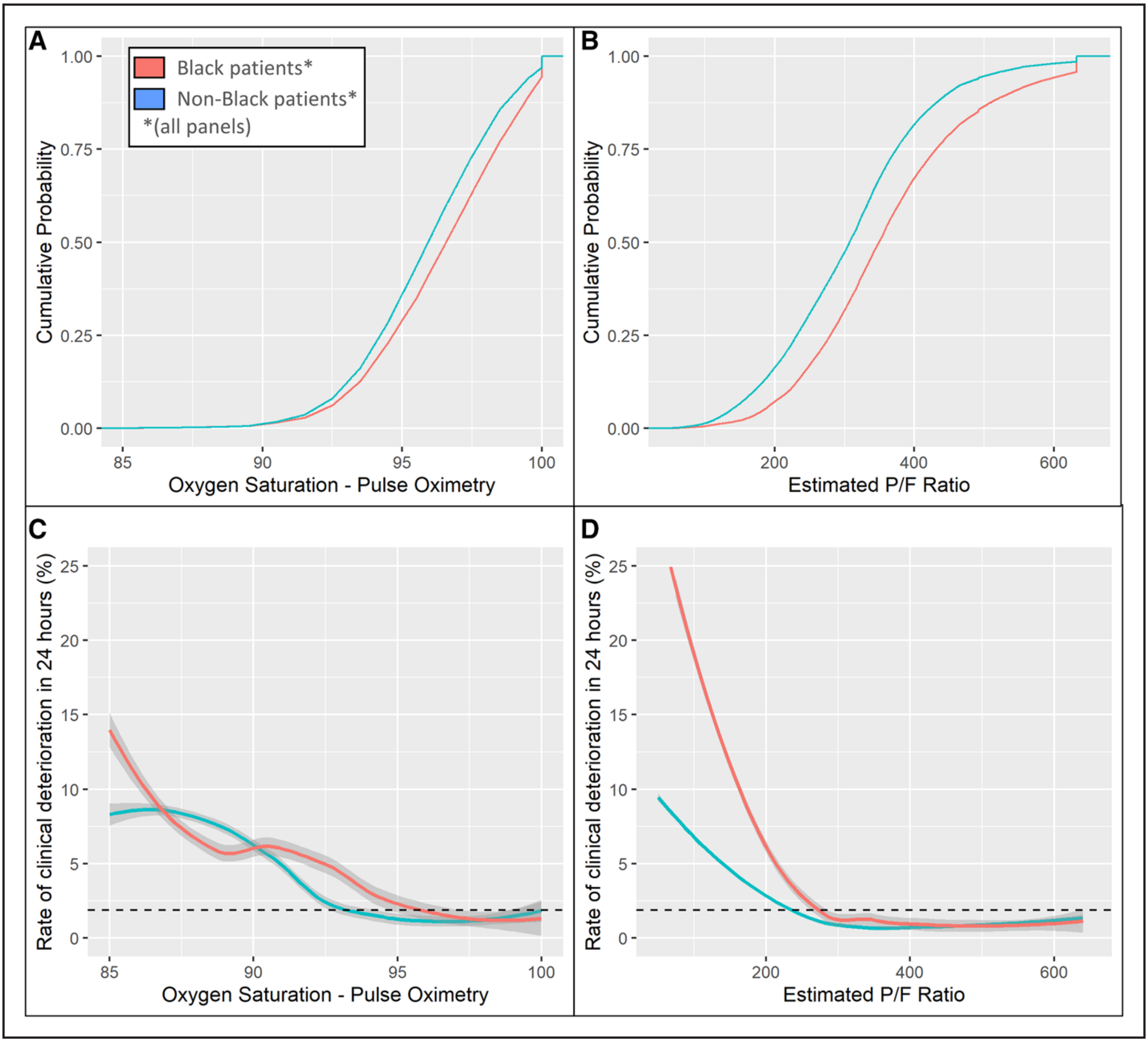


Figure 3. Characterizing the impact of racially biased pulse oximetry measurements. **A** and **B**, Empirical cumulative distribution functions (ECDFs) for oxygen saturation from pulse oximetry (Spo_2) and estimated Pao_2/Fio_2 ratio (ePFR), respectively. This figure depicts the results from University of Virginia. Corresponding results from Emory are shown in eFigure 2 (http://links.lww.com/CCX/B110). Race is encoded by color (*red*—Black patients, *blue*—others). The separation in Spo_2 distributions was narrow (being minimal at $Spo_2 < 92\%$), suggesting an equitable clinician effort to prevent oxygen desaturation. Yet, the separation in the ePFR distribution was wide at all values. This suggests that, on average, clinicians were achieving their Spo_2 targets with lower Fio_2 settings in Black patients (**eFig. 4**, http://links.lww.com/CCX/B110). For comparable ePFR values, outcomes were worse for Black patients than others (**D**). Together, these findings reveal that clinicians were likely undertreating hypoxemia due to an overestimation of Spo_2. Significantly, this disparity remained undetected when the Spo_2 was studied (**C**) instead of ePFR (**D**). To make the plots directly comparable despite the varying scales of the hypoxemia measures, we used Spo_2 values ranging from 85% to 100% and the corresponding range from a minimum ePFR of 50 (representing a Spo_2 of 85% on 100% Fio_2) to a maximum ePFR of 633 (representing a Spo_2 of 100% on room air). To smoothen the ECDFs, we converted Spo_2 from integer to continuous by adding uniformly distributed noise (± 0.5% with a maximum Spo_2 of 100%). To calculate the rate of clinical deterioration at a particular level, we used a window centered at that level with width equal to one SD (2.5 for Spo_2 and 120 for ePFR). The *dashed horizontal lines* (**C** and **D**) mark the rate of clinical deterioration in the entire dataset (1.85%). P/F ratio = ratio of Pao_2/Fio_2.

DISCUSSION

We studied how noninvasive measures of oxygenation inform on the clinical course of hospital patients with COVID-19. Our major findings are that a P/F ratio estimated by applying a model of the oxygen dissociation curve to pulse oximetry data (ePFR) had strong predictive validity for COVID-19 outcomes and that pathologic hypoxemia can be hidden in Black patients.

The AOR of 0.990–0.995 for a 1-point rise in ePFR reflects a strong relationship with clinical deterioration, considering the degree of variability that is typically observed in the ePFR (SD around 120). It is equivalent to an odds ratio for deterioration of 1.7–2.7 for a 100-point decrease in the ePFR. On its own, ePFR outperformed complex multi-system dysfunction models like NEWS and SOFA in predicting deterioration. This likely reflects the uniqueness of COVID-19 as a syndrome in which acute deterioration occurs predominantly from impaired oxygenation. In syndromes like sepsis, which consist of a multiple system organ dysfunction, the incorporation of ePFR into clinical criteria may enhance their performance.

As demonstrated in Figure 1, conventional markers of hypoxemia can be shown to have poor construct validity in common clinical scenarios. In some scenarios, these measures detect changes in hypoxemia when none exist. This may lead to false alarms and alarm fatigue. Even more concerning are the scenarios where these markers fail to sound early alarms about worsening hypoxemia. Such errors may lead to missed opportunities for early intervention and adverse patient outcomes. The ePFR is less prone to these problems. Interestingly, we found that the improvements in construct validity were not necessarily associated with improvements in predictive validity. For example, the predictive validity of the ePFR was similar to that of the S/F ratio. This finding is likely attributable to our retrospective study design and our choice of AUROC as the measure of predictive validity. In prospective studies that measure the promptness of hypoxemia alerts, the benefits of improved construct validity may be more prominently noted. The ePFR, therefore, outperformed other operational markers of hypoxemia in at least one domain of performance (availability, construct validity, predictive validity, and ability to characterize racial disparities) and was noninferior in all other domains. It may, therefore, be the preferred alternative when measured P/F ratios are missing.

Importantly, this study validates the ePFR as a tool to demonstrate the real-world effects of racially biased pulse oximetry readings. We found no disparities in the probability of significant oxygen desaturation (such as $\text{Spo}_2 < 90$), which suggests that clinicians were equitable in their efforts to prevent desaturation by adjusting supplemental oxygen. Yet, the separation in ePFR distributions was wide even at low values. This suggests that, on average, clinicians were achieving their Spo_2 targets with lower supplemental oxygen settings in Black patients (eFig. 4, http://links.lww.com/CCX/B110). By itself, this finding could suggest that Black patients were hospitalized with less severe respiratory failure than others. But that conclusion is inconsistent with the finding that for comparable levels of oxygenation, Black patients were at higher risk of adverse outcomes than others (AOR, 1.2–2.2). Together, these findings point to a phenomenon like occult hypoxemia, which leads clinicians to use lower Fio_2 settings because of a falsely reassuring Spo_2 reading, leading to worse outcomes. We do acknowledge, however, that other explanations for this finding are plausible, such as disparities in the overall quality of care received by Black patients.

Our approach of comparing empirical cumulative distributions of ePFR is not limited by the need for arterial blood sampling. It will enable research into occult hypoxemia on a larger scale than has been possible to date. This new study design can equip consumers, advocates, politicians, and regulators with evidence of racial disparities attributable to pulse oximetry to create the market forces and/or regulatory climate needed to bring an end to this important, longstanding source of structural inequity in healthcare. Until the time that pulse oximeter performance becomes racially equitable, the ePFR can be used to account for the influence of skin color on hypoxemia severity estimation. However, such adjustments will need to be approached with caution, especially within the context of multiple organ dysfunction models like SOFA. For example, studies have shown that the overall SOFA score may overestimate mortality risk in Black patients (40–42) and that this effect is primarily driven by the renal component (41). If a bias toward underestimation of respiratory risk is corrected without correcting the bias toward overestimation of the renal risk, the overall racial bias of the SOFA model may be aggravated.

The strength of our method for computing ePFRs is that it is grounded in the well-established physiology of the oxygen-hemoglobin dissociation curve. Unlike statistical imputation strategies (like multiple imputation), its reliability is not related to frequency of missing data. Additionally, our method lends itself to convenient implementation in large datasets including electronic medical records. Finally, the reproducibility of findings in diverse clinical settings is a major strength of this work.

A limitation of this work is the use of a care-delivery outcome. The reproducibility of results at diverse sites does improve confidence in findings. Still, several clinical practices may differ between sites and with times, affecting generalizability. Exclusion of patients who were critically ill on arrival (e.g., direct admits to ICU from outside hospitals, transfers to ICU within 4 hr of presentation) may be a source of bias, but this is likely to be small. Another limitation is our broad categorization of patients as Black or non-Black. Skin color is not binary; skin color and racial identity are incongruous, and the race as recorded in the medical record is frequently misaligned with the patient's racial identity (43). Additionally, we use aforementioned broad racial categories as predictors of risk in order to highlight disparities. But there is increasing consensus that stratified studies may be preferable over the use of race as a biological risk predictor (44). As such, the inability of this retrospective study to perform race (or skin color) stratified analyses is a limitation.

CONCLUSIONS

P/F ratios estimated using the oxygen dissociation curve were simple to implement and accurately measured the severity of overt hypoxemic respiratory failure. In patients with COVID-19, they outperformed complex multiple system organ dysfunction models. Estimated P/F ratios may allow real-world modeling of racial disparities in outcomes attributable to occult hypoxemia from pulse oximetry.

1 University of Virginia School of Medicine, Charlottesville, VA.

2 University of Virginia Center for Advanced Medical Analytics.

3 Emory University, Atlanta, GA.

4 Nihon Kohden Digital Health Solutions, Inc, Irvine, CA.

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

Drs. Gadrey (University of Virginia) and Kamaleswaran (Emory) had full access to all of the data from their site used in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. All authors contributed substantially to the study design, data analysis and interpretation, and the writing of the article.

This work was supported by the Manning fund for COVID-19 research.

Dr. Moore is supported by the National Institutes of Health (U01AI150508). Dr. Holder is supported by the National Institutes of Health (K23GM37182), and he has received speaker and consulting fees from Baxter International and Philips, respectively. Dr. Kamaleswaran was supported by the National Institutes of Health (R01GM139967). Dr. Clark is an employee of Nihon Kohden Digital Health Solutions (Irvine, CA). Dr. Moorman has equity in Medical Predictive Science Corporation, Charlottesville, VA, and consults for Nihon Kohden Digital Health Solutions, Irvine, CA, with proceeds donated to the University of Virginia Medical Foundation. The remaining authors have disclosed that they do not have any potential conflicts of interest.

For information regarding this article, E-mail: SMG7T@virginia.edu

REFERENCES

1. Gupta RK, Marks M, Samuels THA, et al; UCLH COVID-19 Reporting Group: Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19: An observational cohort study. *Eur Respir J* 2020; 56:2003498
2. Wynants L, Calster BV, Collins GS, et al: Prediction models for diagnosis and prognosis of Covid-19: Systematic review and critical appraisal. *BMJ* 2020; 369:m1328
3. Knight SR, Ho A, Pius R, et al; ISARIC4C investigators: Risk stratification of patients admitted to hospital with Covid-19 using the ISARIC WHO Clinical Characterisation Protocol: Development and validation of the 4C Mortality Score. *BMJ* 2020; 370:m3339
4. Liang W, Liang H, Ou L, et al: Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. *JAMA Intern Med* 2020; 180:1081–1089
5. Gerotziafas GT, Sergentanis TN, Voiriot G, et al: Derivation and validation of a predictive score for disease worsening in patients with COVID-19. *Thromb Haemost* 2020; 120:1680–1690
6. Saria S, Schulam P, Yeh BJ, et al: Development and validation of ARC, a model for anticipating acute respiratory failure in coronavirus disease 2019 patients. *Crit Care Explor* 2021; 3:e0441
7. Fukuda Y, Tanaka A, Homma T, et al: Utility of SpO_2/FiO_2 ratio for acute hypoxemic respiratory failure with bilateral opacities in the ICU. *PLoS One* 2021; 16:e0245927
8. Singhal L, Garg Y, Yang P, et al: eARDS: A multi-center validation of an interpretable machine learning algorithm of early onset acute respiratory distress syndrome (ARDS) among critically ill adults with COVID-19. *PLoS One* 2021; 16:e0257056
9. Singh V, Kamaleswaran R, Chalfin D, et al: A deep learning approach for predicting severity of COVID-19 patients using a parsimonious set of laboratory markers. *iScience* 2021; 24:103523

10. Galloway JB, Norton S, Barker RD, et al: A clinical risk score to identify patients with COVID-19 at high risk of critical care admission or death: An observational cohort study. *J Infect* 2020; 81:282–288
11. Xie J, Hungerford D, Chen H, et al: Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19. *medRxiv* Preprint posted online April 7, 2020. doi: 10.1101/2020.03.28.20045997
12. Vaid S, Kalantar R, Bhandari M: Deep learning COVID-19 detection bias: Accuracy through artificial intelligence. *Int Orthop* 2020; 44:1539–1542
13. Guillamet MCV, Guillamet RV, Kramer AA, et al. Toward a Covid-19 score-risk assessments and registry. *medRxiv* Preprint posted online April 20, 2020. doi: 10.1101/2020.04.15.20066860
14. Olsson T, Terent A, Lind L: Rapid Emergency Medicine score: A new prognostic tool for in-hospital mortality in nonsurgical emergency department patients. *J Intern Med* 2004; 255:579–587
15. Prytherch DR, Smith GB, Schmidt PE, et al: ViEWS--Towards a national early warning score for detecting adult inpatient deterioration. *Resuscitation* 2010; 81:932–937
16. Smith GB, Prytherch DR, Meredith P, et al: The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation* 2013; 84:465–470
17. Singer M, Deutschman CS, Seymour CW, et al: The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA* 2016; 315:801–810
18. Jakobsen JC, Gluud C, Wetterslev J, et al: When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Med Res Methodol* 2017; 17:162
19. Seymour CW, Liu VX, Iwashyna TJ, et al: Assessment of clinical criteria for sepsis: For the third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA* 2016; 315:762–774
20. Gadrey SM, Clay R, Zimmet AN, et al: The relationship between acuity of organ failure and predictive validity of Sepsis-3 criteria. *Crit Care Explor* 2020; 2:e0199
21. Gadrey SM, Lau CE, Clay R, et al: Imputation of partial pressures of arterial oxygen using oximetry and its impact on sepsis diagnosis. *Physiol Meas* 2019; 40:115008
22. Hill AV: Proceedings of the physiological society: January 22, 1910. *J Physiol* 1910; 40(Suppl): iv–vii
23. Goutelle S, Maurin M, Rougier F, et al: The Hill equation: A review of its capabilities in pharmacological modelling. *Fundam Clin Pharmacol* 2008; 22:633–648
24. Severinghaus JW: Simple, accurate equations for human blood O_2 dissociation computations. *J Appl Physiol* 1979; 46:599–602
25. Jubran A, Tobin MJ: Reliability of pulse oximetry in titrating supplemental oxygen therapy in ventilator-dependent patients. *Chest* 1990; 97:1420–1425
26. Ebmeier SJ, Barker M, Bacon M, et al: A two centre observational study of simultaneous pulse oximetry and arterial oxygen saturation recordings in intensive care unit patients. *Anaesth Intensive Care* 2018; 46:297–303
27. Sjoding MW, Dickson RP, Iwashyna TJ, et al: Racial bias in pulse oximetry measurement. *N Engl J Med* 2020; 383:2477–2478
28. Wong A-KI, Charpignon M, Kim H, et al: Analysis of discrepancies between pulse oximetry and arterial oxygen saturation measurements by race and ethnicity and association with organ dysfunction and mortality. *JAMA Netw Open* 2021; 4:e2131674
29. Vesoulis Z, Tims A, Lodhi H, et al: Racial discrepancy in pulse oximeter accuracy in preterm infants. *J Perinatol* 2021; 1:7
30. Henry NR, Hanson AC, Schulte PJ, et al: Disparities in hypoxemia detection by pulse oximetry across self-identified racial groups and associations with clinical outcomes*. *Crit Care Med* 2022; 50:204–211
31. Valbuena VSM, Barbaro RP, Claar D, et al: Racial bias in pulse oximetry measurement among patients about to undergo extracorporeal membrane oxygenation in 2019-2020: A retrospective cohort study. *Chest* 2022; 161:971–978
32. Burnett GW, Stannard B, Wax DB, et al: Self-reported race/ethnicity and intraoperative occult hypoxemia: A retrospective cohort study. *Anesthesiology* 2022; 136:688–696
33. Fawzy A, Wu TD, Wang K, et al: Racial and ethnic discrepancy in pulse oximetry and delayed identification of treatment eligibility among patients with COVID-19. *JAMA Intern Med* 2022; 182:730–738
34. Matt V, Matthew H: The retrospective chart review: Important methodological considerations. *J Educ Eval Health Prof* 2013; 10:12
35. Harris PA, Taylor R, Thielke R, et al: Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009; 42:377–381
36. Harris PA, Taylor R, Minor BL, et al; REDCap Consortium: The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform* 2019; 95:103208
37. Krippendorff K: Estimating the reliability, systematic error and random error of interval data. *Educ Psychol Meas* 1970; 30:61–70
38. DeLong ER, DeLong DM, Clarke-Pearson DL: Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 1988; 44:837–845
39. R Core Team: R: A Language and Environment for Statistical Computing. 2018. Available at: https://www.R-project.org/. Accessed July 2, 2018
40. Ashana DC, Anesi GL, Liu VX, et al: Equitably allocating resources during crises: Racial differences in mortality prediction models. *Am J Respir Crit Care Med* 2021; 204:178–186
41. Miller WD, Han X, Peek ME, et al: Accuracy of the Sequential Organ Failure Assessment score for in-hospital mortality by race and relevance to crisis standards of care. *JAMA Netw Open* 2021; 4:e2113891
42. Sarkar R, Martin C, Mattie H, et al: Performance of intensive care unit severity scoring systems across different ethnicities. *medRxiv* Preprint posted online January 20, 2021. doi: 10.1101/2021.01.19.21249222
43. Klinger EV, Carlini SV, Gonzalez I, et al: Accuracy of race, ethnicity, and language preference in an electronic health record. *J Gen Intern Med* 2015; 30:719–723
44. Vyas DA, Eisenstein LG, Jones DS: Hidden in plain sight - reconsidering the use of race correction in clinical algorithms. *N Engl J Med* 2020; 383:874–882