OPEN

The Relationship Between Acuity of Organ Failure and Predictive Validity of Sepsis-3 Criteria

Shrirang M. Gadrey, MBBS, MPH[1,2]; Russ Clay, PhD[1]; Alex N. Zimmet, MD[1,2]; Alexander S. Lawson, MD[1]; Samuel F. Oliver, MD[1]; Emily D. Richardson, MD[1]; Vernon J. Forrester, MD[1]; Robert T. Andris, BS[1,2]; Garret T. Rhodes, MS[1]; John D. Voss, MD[1]; Christopher C. Moore, MD[1,2]; J. Randall Moorman, MD[1,2]

Objectives: The Sepsis-3 taskforce defined sepsis as suspicion of infection and an acute rise in the Sequential Organ Failure Assessment score by 2 points over the preinfection baseline. Sepsis-3 studies, though, have not distinguished between acute and chronic organ failure, and may not accurately reflect the epidemiology, natural history, or impact of sepsis. Our objective was to determine the extent to which the predictive validity of Sepsis-3 is attributable to chronic rather than acute organ failure.

Design: Retrospective cohort study.

Setting: General medicine inpatient service at a tertiary teaching hospital.

Patients: A total of 3,755 adult medical acute-care encounters (1,864 confirmed acute infections) over 1 year.

Interventions: None.

Measurements and Main Results: We measured the total Sequential Organ Failure Assessment score at the onset of infection and separated its components (baseline and acute rise) using case-by-case chart reviews. We compared the predictive validities of acuity-focused (acute rise in Sequential Organ Failure Assessment $\geq$ 2) and conventional (total Sequential Organ Failure Assessment $\geq$ 2) implementations of Sepsis-3 criteria. Measures of predictive validity were change in the rate of outcomes and change in the area under receiver operating characteristic curves after adding sepsis criteria to multivariate logistic regression models of baseline risk (age, sex, race, and Charlson comorbidity index). Outcomes were inhospital mortality (primary) and ICU transfer or inhospital mortality (secondary). Acuity-focused implementations of Sepsis-3 were associated with neither a change in mortality (2.2% vs 1.2%; $p = 0.18$) nor a rise in area under receiver operating characteristic curves compared with baseline models (0.67 vs 0.66; $p = 0.75$). In contrast, conventional implementations were associated with a six-fold change in mortality (2.4% vs 0.4%; $p = 0.01$) and a rise in area under receiver operating characteristic curves compared with baseline models (0.70 vs 0.66; $p = 0.04$). Results were similar for the secondary outcome.

Conclusions: The evaluation of the validity of organ dysfunction-based clinical sepsis criteria is prone to bias, because acute organ dysfunction consequent to infection is difficult to separate from preexisting organ failure in large retrospective cohorts.

Key Words: cohort studies; hospital mortality; inpatients; organ dysfunction scores; prognosis; sepsis

The Sepsis-3 taskforce advanced the ideas that: 1) the difference between sepsis and uncomplicated infections is the presence of a life-threatening dysregulated host response and 2) the best marker of sepsis is acute organ dysfunction, even when present only to a modest degree (1). This was a bold departure from older conceptualizations that viewed sepsis as a state of excessive inflammation operationalized with the systemic inflammatory response syndrome criteria (2, 3). The taskforce operationalized the life-threatening component of their definition by resting the validity of the clinical criteria, in large part, on their discrimination for mortality (4). Furthermore, they operationalized the organ dysfunction component by recommending the Sequential Organ Failure Assessment (SOFA) score as the clinical criterion.

The SOFA score rates six categories of organ failure (cardiovascular, respiratory, neurologic, renal, hepatic, and coagulation), each on a scale of 0–4, and takes on values from 0–24. The total SOFA score in any acutely infected patient can be separated into

[1]Department of Medicine, University of Virginia School of Medicine, Charlottesville, VA.

[2]Center for Advanced Medical Analytics, University of Virginia, Charlottesville, VA.

Crit Care Expl 2020; 2:e0199

DOI: 10.1097/CCE.0000000000000199

two components. The first of these is a preinfection baseline SOFA component that captures chronic or preexisting organ failure. The second is an acute SOFA component that captures an acute rise in SOFA associated with the onset of sepsis. Sepsis is, by definition, an acute syndrome. Therefore, diagnosing sepsis using the preinfection baseline SOFA score is clinically inconsistent. Therefore, the taskforce sensibly defined sepsis as an acute rise in the SOFA score of 2 or more over the preinfection baseline.

Automated queries of large datasets can determine the total SOFA at the time of infection. However, for accurate operationalization of the new definition, it is also necessary to determine the preinfection baseline SOFA and subtract it from the total SOFA. The task force suggested using a baseline SOFA score of 0 unless the patient is known to have preexisting organ dysfunction (1). They did not, however, specify a method for retrospective assignment of baseline SOFA scores to patients that are known to have preexisting organ dysfunction. This has led to an inconsistency in research methods and to uncertainty about whether studies report on the outcomes of infected patients with preexisting organ failure mixed with patients with sepsis (**Fig. 1**).

Importantly, all large assessments of the predictive validity of Sepsis-3, including the original taskforce-commissioned work (4), did not specifically assess this distinction. They have conventionally implemented Sepsis-3 criteria by assigning a baseline SOFA of 0 to all patients, thereby assuming that all SOFA points are acutely acquired. This conventional implementation reports the predictive validity of total SOFA at the time of infection rather than the taskforce-defined acute rise in SOFA due to a life-threatening dysregulated host response (4–6).

The Sepsis-3 taskforce did conduct a post hoc analysis where a change in SOFA was calculated of 2 points or more from up to 48 hours before to up to 24 hours after the onset of infection—a metric often referred to as the "delta SOFA" (4, 7). However, this post hoc implementation is not consistent with the taskforce recommendation of a change in SOFA of 2 points or more from the preinfection baseline SOFA. For example, if a patient with preinfection SOFA of 0 presents with acute encephalopathy, elevated bilirubin, and hypotension from cholangitis, their acute rise in SOFA relative to preinfection baseline is 3 (one each for neurologic, hepatic, and cardiovascular components). They clearly meet the Sepsis-3 operational criteria for sepsis. However, if the mentation, liver tests, and blood pressure rapidly improve after fluid resuscitation, antibiotics, and an endoscopic retrograde cholangio-pancreatography, then their subsequent SOFA score is lower than the presenting SOFA score. In such a case, the post hoc analysis would assign a delta SOFA less than 2, thereby labeling this patient as nonseptic despite the presence of infection-associated acute organ dysfunction.

Other attempts have involved indirect adjustments for preinfection baseline SOFA scores by imputing preinfection baseline SOFA with multiple imputation algorithms (4), assigning arbitrary baseline SOFA scores to billing codes (5) and considering the lowest SOFA during the encounter as the baseline (8). None have been validated as accurate measures of the preinfection baseline SOFA required to implement the taskforce definition of an acute rise by 2 points or more. The most rigorous practice we encountered was the calculation of preinfection baseline SOFA on a case-by-case basis using physician chart reviews. An instance of its use is a recent study, examining the preventability of 568 sepsis-associated deaths (9). We did not encounter any studies that examined the predictive validity of Sepsis-3 criteria using this method.

Thus, the impact of overlooking the distinction between acute and chronic organ failures on estimated predictive validity of the Sepsis-3 criteria remains unknown. To test this impact, we studied a cohort where the baseline SOFA had been ascertained on a case-by-case basis by physician chart reviewers. We used this dataset to compare the predictive validity of an acuity-focused implementation of Sepsis-3 criteria (acute SOFA ≥ 2) with that of a conventional implementation (any SOFA ≥ 2).

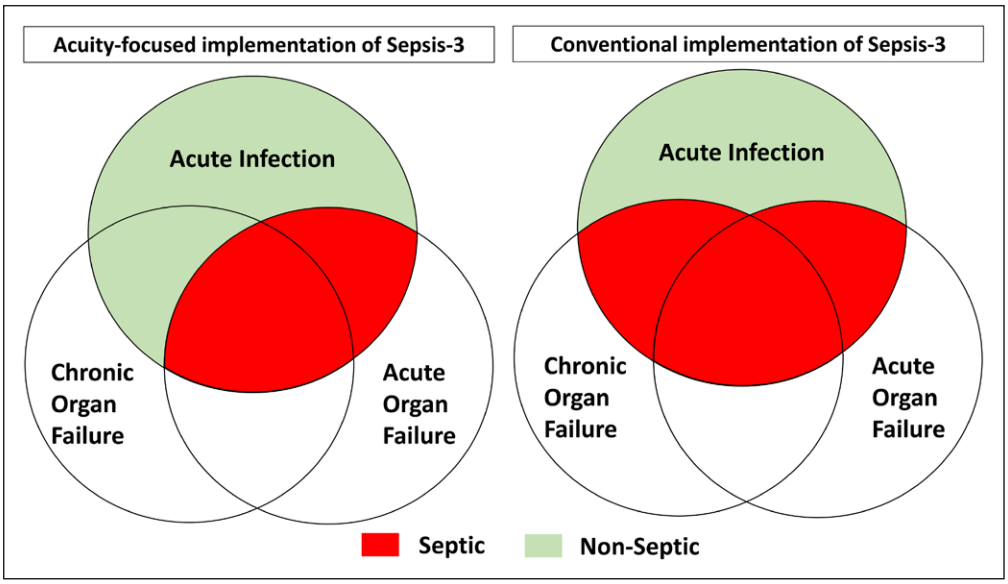

Figure 1. This Venn diagram depicts the difference between the acuity-focused and conventional Sepsis-3 implementations that we compared in this study. The parts shaded *red* are the cases that are labeled as septic and the parts shaded *green* are labeled as nonseptic or uncomplicated infections by the implementation. The acuity-focused implementation used in this study relied on a case-by-case adjudication of baseline Sequential Organ Failure Assessment (SOFA) to ensure that only acutely acquired SOFA points were used to define sepsis. The conventional implementation involved assuming a baseline SOFA of 0 for all patients, thereby assuming that all SOFA points are acute. The difference lies in the classification of acutely infected patients with preexisting chronic organ failure but no acute organ failure (20% in our cohort) (Table 1). Such patients are at risk of adverse outcomes for many reasons unrelated to sepsis. Our study evaluated the hypothesis that this misclassification is an important source of bias in contemporary sepsis research.

MATERIALS AND METHODS

Study Design and Patient Population

We performed a retrospective cohort study among adult (age ≥ 18) medical acute-care encounters. The final cohort (**Fig. 2** and **Table 1**) included all such encounters that occurred

between July 2016 and June 2017 on the General Medicine service at the University of Virginia medical center, an academic tertiary-care center.

Interrater Reliability

We relied on manual chart review to determine presence of infection and to separate baseline SOFA points from acute SOFA points. Six of the authors (S.M.G., A.N.Z, A.S.L., S.F.O., E.D.R., and V.J.F.) were the reviewers. To ensure optimal reliability of manually abstracted data, we adhered to established best practices (10). We established clear operational definitions of the variables to be manually abstracted. We created standardized abstraction forms using REDCap electronic data capture tools hosted at the University of Virginia (11, 12). Before beginning the chart review process, we conducted three rounds of pilot reviews, where all six reviewers would rate the same encounters. Reviewers discussed the disagreements that emerged in these pilot rounds and ways to avoid these modes of disagreement. A brief procedural manual was used with rules that are outlined in **Appendix Table 1** (http://links.lww.com/CCX/A338).

We randomly sampled 10% of each reviewer's chart for blinded second rater reviews. We performed this sampling three times (at 6-wk intervals) until the chart reviews were complete. We calculated percent agreement and Krippendorff alpha (13) for each of the three samples and cumulatively. We selected Krippendorff alpha, because it is suited for more than two reviewers. Interrater reliability (IRR) is high when alpha is greater than or equal to 0.8, moderate when $0.8 > \text{alpha} \geq 0.67$, and poor when alpha less than 0.67. With this process, we ensured high reliability of our data (**Table 2**).

Determining Presence of Suspected Acute Infection

We started with a cohort of 3,755 encounters and selected 1,864 encounters where the presence of suspected acute infection had been ascertained with high confidence (Fig. 2). We defined suspicion of infection as the administration of a new therapeutic antimicrobial regimen for at least 3 consecutive days. We used a minimum of 3 days, because the shortest courses used in clinical practice are three days (uncomplicated cystitis or acute bronchitis in chronic obstructive pulmonary disease [COPD]). We accepted all systemic routes of antimicrobial administration. We did not link inclusion to culture orders. We only studied the first infection per encounter.

Implementing Sepsis-3 Criteria

For the conventional implementation, we used automated queries to determine the total (maximally deranged) values of SOFA at the onset of infection—a time window that we defined to start 2 days before and end 1 day after the initial antimicrobial orders. We assumed a baseline SOFA of 0 for all cases and defined as septic any patient whose total SOFA was 2 or more. We emphasize that this implementation was conventional in its assumption that baseline SOFA was zero in all encounters. However, some aspects of the implementation were unique. For example, because our suspicion of infection was physician-adjudicated, we did not limit inclusion based on relative timing of cultures and antibiotics. As a result, onset of infection was a window defined around the time of antimicrobials rather than earlier of antibiotics or culture orders.

For the acuity-focused implementation of Sepsis-3 criteria, the physician chart reviewers first determined the baseline SOFA before the onset of infection on a case-by-case basis. We then subtracted baseline SOFA from total SOFA to determine the acute SOFA score and defined patients as septic if the acute rise in SOFA was 2 or more.

When a patient's SOFA component was 0 around the onset of infection, we considered the baseline to be 0. For patients with deranged SOFA components, we used the latest known estimates of baseline SOFA for that component. We did not restrict older records, because: 1) any-time cutoff would necessarily be arbitrary and 2) older documentation of chronic organ failure often remains pertinent, for example, previously noted dementia. For infections whose onset was greater than 2 days after hospitalization, SOFA values from the same encounter were used to estimate baseline SOFA so long as they occurred more than 2 days prior to antimicrobial dosing. In the end, therefore, baseline SOFA estimates varied in their proximity to onset of infection, just as they often do in clinical practice.

When data needed to adjudicate the baseline SOFA were not available within our records, we also accessed external records through our point-to-point health information exchange system (Epic System's Care Everywhere) (14). This search included a review of external notes

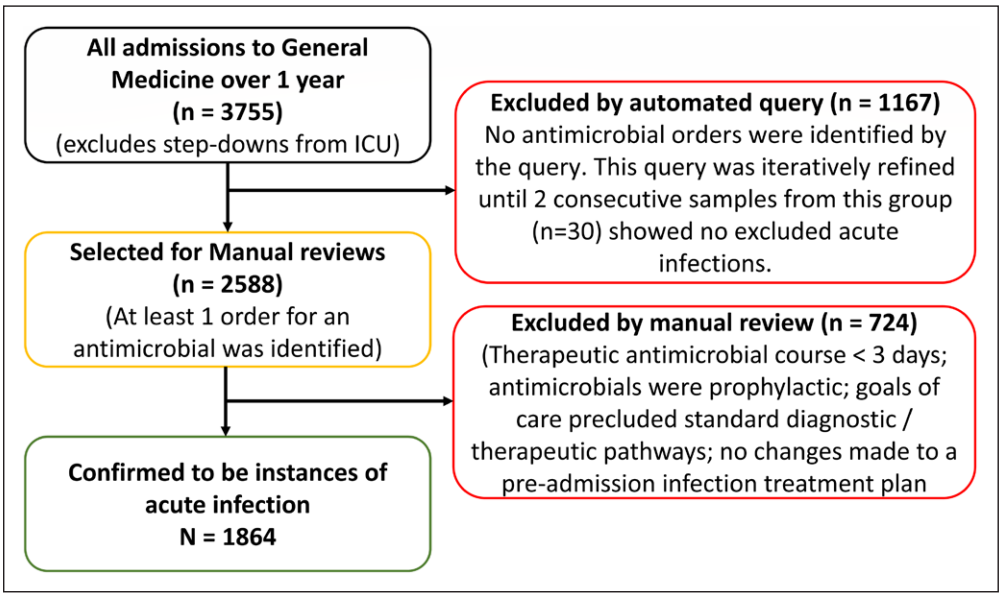

Figure 2. The process we used to select encounters for acute infection. It ensured that the presence of an acute infection and its time of onset had been ascertained with high confidence. Only the first infection per encounter was studied.

TABLE 1. Distribution of Pertinent Demographic and Clinical Variables in Our Cohort

Clinical Values	Confirmed Acute Infections (n = 1,864)
Age, yr, median (IQR)	62 (49–75)
Male, *n* (%)	901 (48.3)
Race, *n* (%)	
White	1,428 (76.6)
Black	406 (21.8)
Other	30 (1.6)
Charlson comorbidity index, median (IQR)	6 (4–9)
Type of Infection, *n* (%)	
Respiratory	510 (27)
Urinary tract	467 (25)
Skin, soft tissue, or musculoskeletal	454 (24)
Intra-abdominal	187 (10)
Blood stream infection	97 (5)
CNS	25 (1.3)
Other	186 (10%)
Onset of Infection, *n* (%)	
Present at admission	1,741 (93.4)
Onset > 2 d after admission	123 (6.6)
Type of organ failure at onset of infection, *n* (%)	
No organ failure	439 (23.6)
Chronic organ failure only	370 (19.8)
Acute and chronic organ failure	485 (26)
Acute organ failure only	570 (30.6)
Inhospital mortality, *n* (%)	33 (1.8)
Inhospital mortality or ICU transfer, *n* (%)	154 (8.3)

IQR = interquartile range.

(for descriptions of home oxygen levels or baseline neurologic examinations) and laboratory reports (baseline creatinine, bilirubin, and platelet levels).

We were able to adjudicate baseline values for SOFA components in greater than 95% of the instances. Only in the remaining cases where no records of a baseline SOFA were available did we assume a baseline SOFA component of 0 (2.4% for renal, 3.9% for hepatic, 3.1% for respiratory, 1% for neurologic, and 2.7% for coagulation components). We subjected the inability to determine a baseline SOFA component leading to an assumed score of 0 to the same IRR checks.

We note that the physician chart reviewer did not exercise any subjective judgment regarding presence of sepsis, because there is no gold standard to allow for diagnostic certainty and physician error is not uncommon. The chart reviewer only focused on ascertaining: 1) presence of infection and 2) acuity of organ failure.

Appendix Table 1 and **Appendix Figure 1** (http://links.lww.com/CCX/A338) contain a detailed account of this process for each component of the SOFA score (15). We note that there are no validated methods to assign baseline SOFA scores. In this context, we concluded that the least unreliable method of determining baseline SOFA would be the one that closely resembles a clinician's approach at the bedside. We designed our method of separating baseline from acute SOFA scores consistent with this guiding principle.

Diagnostic Accuracy of Conventional Implementation of Sepsis-3 Criteria

We conducted three tests for diagnostic accuracy of conventional implementation Sepsis-3 criteria. First, we determined the accuracy of the conventional criteria for suspicion of infection (administration of an antibiotic either 72 hr after or 24 hr before a body fluid culture) for detecting the true infections determined by chart reviewing physicians. Second, among the 1,864 encounters for infection (physician confirmed), we determined the accuracy of the conventional implementation of Sepsis-3 criteria (any SOFA ≥ 2) by comparing it with the acuity-focused implementation (acute rise in SOFA ≥ 2), which is consistent with the taskforce recommendations. Third, we estimated the overall accuracy of sepsis event identification by serially applying the infection and sepsis criteria.

Impact of Implementation on Estimated Predictive Validity for Adverse Outcomes

We used two measures of predictive validity to understand whether the predictive validity of conventional implementation of Sepsis-3 criteria differed from that of the acuity-focused one.

First, we determined the change in rate of adverse outcomes associated with criteria positivity. We compared the rate of outcome among criteria positive patients with that among criteria negative ones using two population tests of proportion. We considered p values from two-sided tests to be significant at $p < 0.05$.

Second, we computed the areas under the receiver operating characteristic (AUROC) curves to test whether adding Sepsis-3 criteria significantly improved the performance of a baseline risk model. The rise in AUROC was considered significant for p values of less than 0.05 using Delong test for comparing two or more correlated AUROCs (16). We first conducted this test using the conventional implementation (any SOFA ≥ 2). We repeated the test using the acuity-focused implementation (acute rise in SOFA ≥ 2).

We used R Version 3.5.1 to perform all analyses (17). The University of Virginia's Institutional Review Board for Health Sciences Research approved the study with a waiver of informed consent (Protocol 20249).

RESULTS

Cohort Characteristics

The demographic and clinical characteristics of our cohort are outlined in Table 1.

Interrater Reliability

Table 2 shows the results of blinded second reviews of each of the three random samples of charts as well as the cumulative results.

TABLE 2. Interrater Reliability of Manually Abstracted Variables[a]

Manually Abstracted Variable	Sample 1 ($n = 76$)		Sample 2 ($n = 82$)		Sample 3 ($n = 101$)		Cumulative ($n = 259$)	
	Agreement (%)	Krippendorf alpha	Agreement (%)	Krippendorf alpha	Agreement (%)	Krippendorf alpha	Agreement (%)	Krippendorf alpha
Presence of infection	92	0.91	98	0.90	93	0.86	94	0.88
Baseline SOFA: PF ratio	100	1	100	1	96	0.65	99	0.90
Baseline SOFA: creatinine	94	0.90	92	0.91	95	0.94	94	0.93
Baseline SOFA: total bilirubin	96	0.54	100	1	100	1	99	0.88
Baseline SOFA: platelets	100	1	98	0.81	92	0.83	97	0.81
Baseline SOFA: neurologic	88	0.72	90	0.73	97	0.92	92	0.77
Peak SOFA: neurologic	88	0.72	93	0.91	96	0.84	92	0.81

SOFA = Sequential Organ Failure Assessment.

[a]We randomly sampled 10% of each reviewer's charts for blinded second rater reviews. We performed this sampling three times (at 6-wk intervals) until the chart reviews were complete. Reported in this table are results for each of the three samples and the overall results.

For all variables, the final raw agreement rate was over 90%. The IRR was high (Krippendorff alpha > 0.8) for all variables except baseline neurologic SOFA. The reliability for baseline neurologic SOFA was moderate. Yet, the final raw agreement rate was over 90%, making it acceptable for inclusion in analysis.

Diagnostic Accuracy of Conventional Implementation of Sepsis-3 Criteria

Our major findings were as follows. First, the conventional Sepsis-3 infection criteria (combination of antibiotics and body fluid cultures) had a positive predictive value of 78% for physician confirmed acute infection. In the remaining 22% instances, concurrence of antibiotics and cultures did not represent an acute infection. For example, a patient who is on prolonged azithromycin therapy for severe COPD or lifelong suppressive antibiotics for a past orthopedic hardware associated infection may get blood cultures ordered for altered mentation before its etiology is determined to be an acute stroke. Second, the positive predictive value of conventionally implemented Sepsis-3 criteria for presence of sepsis (physician confirmed infection + acute rise in SOFA ≥ 2) was 59% (**Fig. 3**).

Impact of Implementation on Estimated Predictive Validity for Adverse Outcomes

The conventional implementations of Sepsis-3 criteria showed better predictive validity for our primary and secondary outcomes than the acuity-focused implementations (**Table 3**).

When using conventionally implemented criteria, the mortality rate was six-fold higher among criteria positive patients than among criteria negative patients (2.4% vs 0.4%; $p = 0.01$). However, with acuity-focused implementations, there was no mortality rate difference between the criteria positive and negative patients (2.2% vs 1.2%; $p = 0.18$). Similarly, when using conventional implementations, the rate of secondary outcome (ICU transfer or mortality) was 3.5-fold higher among criteria positive patients than among criteria negative patients (10.1% vs 2.9%; $p < 0.01$). However, with acuity-focused implementations, the difference was narrower (10.3% vs 5.6%; $p < 0.01$).

In multivariate regression analysis, we assessed whether the AUROC of the predictive models significantly increased after sepsis criteria were added to baseline predictors (age, sex, race, and Charlson comorbidity index). The Charlson comorbidity index,

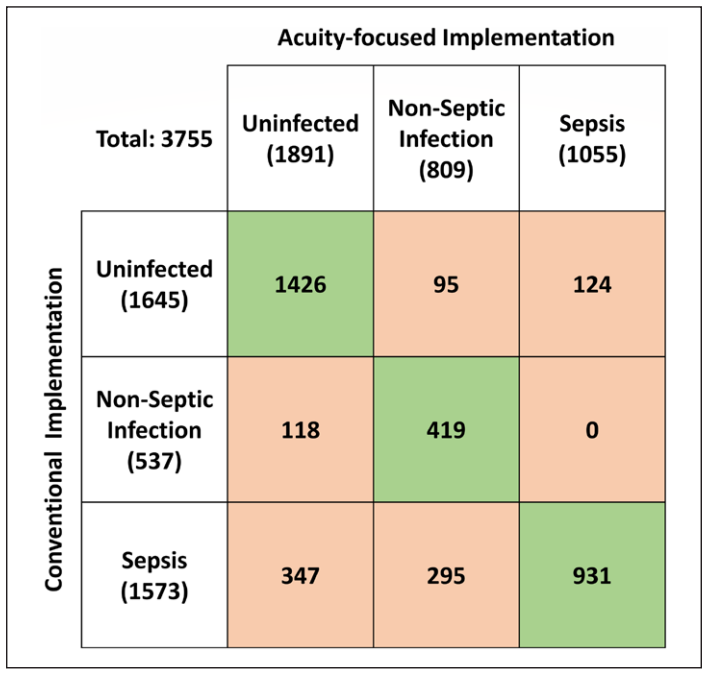

Figure 3. The accuracy of conventional Sepsis-3 implementations for infection and sepsis measured against the acuity-focused implementation rooted in physician chart review. Of the 3,755 encounters, the conventional implementation correctly classified 2,776 encounters (74%) along the *green diagonal*. For suspicion of infection, the positive predictive value was 78% (1,645/2,110) and negative predictive value was 87% (1,426/1,645). For sepsis, the positive predictive value was 59% (931/1,573) and negative predictive value was 94% (2,058/2,182). We note that in the 1,891 patients that were deemed uninfected by physicians (column 1), the charts were not reviewed further to separate acute and chronic organ failure. Therefore, in the 465 encounters where automated implementations falsely detected infections (column 1, rows 2 and 3), the distinction between nonseptic (118) and septic (347) was made based on the assumption the frequency of total Sequential Organ Failure Assessment rising over two is constant in infected and uninfected patients.

TABLE 3. Impact of Sepsis-3 Implementation on Estimated Predictive Validity

Implementation	Outcome	Outcome Rate (Criteria Positive Patients) (%)	Outcome Rate (Criteria Negative Patients) (%)	*p*
Conventional (any SOFA ≥ 2)	Mortality	2.4	0.4	0.01[a]
Acuity-focused (acute SOFA ≥ 2)	Mortality	2.2	1.2	0.17
Conventional (any SOFA ≥ 2)	ICU transfer or mortality	10.1	2.9	< 0.01[a]
Acuity-focused (acute SOFA ≥ 2)	ICU transfer or mortality	10.3	5.6	< 0.01[a]
Implementation	**Outcome**	**AUROC (Baseline Model)**	**AUROC (Baseline Model + Sepsis Criteria)**	***p***
Conventional (any SOFA ≥ 2)	Mortality	0.66	0.70	0.04[a]
Acuity-focused (acute SOFA ≥ 2)	Mortality	0.66	0.67	0.75
Conventional (any SOFA ≥ 2)	ICU transfer or mortality	0.61	0.65	0.01[a]
Acuity-focused (acute SOFA ≥ 2)	ICU transfer or mortality	0.61	0.63	0.18

AUROC = area under the receiver operating characteristic curve, SOFA = Sequential Organ Failure Assessment.
[a]Statistically significant difference.

outlined in **Appendix Table 2** (http://links.lww.com/CCX/A338), is a validated estimator of mortality risk attributable to chronic conditions (18, 19). The AUROC of baseline risk models predicting mortality was 0.66. This rose significantly with the addition of conventionally implemented criteria to the model (0.70 vs 0.66; $p = 0.04$) but not with acuity-focused implementations (0.67 vs 0.66; $p = 0.75$). Similarly, the AUROC of baseline risk models predicting ICU transfer or mortality was 0.61. This rose significantly with the addition of conventionally implemented criteria (0.65 vs 0.61; $p < 0.01$) but not with acuity-focused implementations (0.63 vs 0.61; $p = 0.17$).

DISCUSSION

The Sepsis-3 task force presented a compelling case that it is a dysregulated host response rather than excessive inflammation that makes sepsis life-threatening. They recommended that organ failure-based clinical sepsis criteria are best suited to operationalize this conceptual framework (1). They supported this position with the impressive finding that even a modest degree of organ dysfunction at the onset of infection was associated with a mortality rate comparable with medical emergencies like ST-elevation myocardial infarction (4). Specifically, they reported that an SOFA score of 2 or higher identified a risk of mortality that is increased manifold compared with an SOFA score less than 2 (4).

Unless the distinction between the acute and chronic organ failures is carefully dissected in the infected patient, it is impossible to determine which one of them confers the increased risk of mortality. There is a difference between SOFA points due to chronic derangements in organ failure and those brought about by a life-threatening dysregulation of the host response to infection.

We found that chronic organ failure, quantified by the preinfection baseline SOFA score, earmarked the infected patients likely to die more so than acute organ failure did. In our cohort, the predictive validity of the SOFA score was driven by the preinfection baseline SOFA score. This finding, if widely replicated, would not in any way invalidate the Sepsis-3 taskforce conceptualization that a dysregulated host response makes sepsis lethal. It would, though, lead to deeper scrutiny of the notion that acute organ dysfunction defined as a rise in the SOFA score is the best way to detect the lethal dysregulated host response.

Our study particularly raises important concerns about the conventional practice of assuming a baseline SOFA of 0 in all patients as a pragmatic compromise to enable enormous sample sizes. It demonstrates that this assumption likely introduces significant bias in the estimation of predictive validity of acute organ dysfunction for sepsis-related adverse outcomes. The use of alternate markers of baseline risk (age, gender, race, and Charlson comorbidity index) did not alleviate this bias. Studies that differ in their assumptions regarding acuity of organ failure may not be comparable. Future efforts to develop and validate organ failure-based clinical sepsis criteria must better address this problem.

An additional contribution of our study is to measure the diagnostic accuracy of conventional implementations of Sepsis-3 criteria. In our series, cultures and antibiotics were often ordered concurrently for reasons other than clinical suspicion of an acute infection (22% of instances). Combining this with the inaccuracies introduced by assuming that the preinfection baseline SOFA score was 0 led to an overall positive predictive value of 59% for the conventionally implemented Sepsis-3 criteria. The result of these inaccuracies is to dilute datasets of "infected" and "septic" patients with large numbers of uninfected, nonseptic patients, adding noise to statistical estimators and uncertainty to conclusions about epidemiology, natural history, and impact of sepsis.

A limitation of this study was the variable proximity of baseline SOFA estimates to onset of infection, being missing in some encounters. This problem is largely unavoidable and the one that is encountered in the everyday practice of sepsis care. Another limitation is that it was conducted at a single center and with a small sample size relative to the large-scale Sepsis-3 validation studies. This was necessitated by the labor-intensive nature of chart reviews. As such, our study must be viewed as an initial exploration into the role of preinfection organ failure as a source of bias in

contemporary sepsis research. A deeper understanding into this problem and validated solutions will only emerge from prospective multicenter studies.

The study of sepsis, a common and dire disease of increasing impact on the population and the healthcare system, continues to face a core problem of identifying cases with certainty. Clinicians know how hard it can be to decide if an individual patient is septic, let alone how hard it is to face millions of encounters with infection and to decide who was really septic. Without question, the large-scale studies reveal much about the disease, but smaller studies such as ours with more finely characterized patients must inform the community of the degree of the uncertainty present in larger studies and keep us vigorous in the pursuit of better ways to study and to understand sepsis.

CONCLUSIONS

The evaluation of the validity of organ dysfunction-based clinical sepsis criteria is prone to bias, because acute organ dysfunction consequent to infection is difficult to separate from preexisting organ failure in large retrospective cohorts. Studies that do not separate the acute and preexisting components of clinical sepsis criteria may not accurately report the epidemiology, natural history, or impact of sepsis.

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).

Center for Advanced Medical Analytics, University of Virginia, is supported by the Frederick Thomas Advanced Medical Analytics Fund.

Dr. Moorman has ownership interests in AMP3D (Charlottesville, VA), a company that provides advanced predictive analytic solutions. The remaining authors have disclosed that they do not have any potential conflicts of interest.

For information regarding this article, E-mail: smg7t@hscmail.mcc.virginia.edu

---

REFERENCES

1. 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
2. Bone RC, Balk RA, Cerra FB, et al: Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine. *Chest* 1992; 101:1644–1655
3. Levy MM, Fink MP, Marshall JC, et al; SCCM/ESICM/ACCP/ATS/SIS: 2001 SCCM/ESICM/ACCP/ATS/SIS international sepsis definitions conference. *Crit Care Med* 2003; 31:1250–1256
4. 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
5. Raith EP, Udy AA, Bailey M, et al; Australian and New Zealand Intensive Care Society (ANZICS) Centre for Outcomes and Resource Evaluation (CORE): Prognostic accuracy of the SOFA score, SIRS criteria, and qSOFA score for in-hospital mortality among adults with suspected infection admitted to the intensive care unit. *JAMA* 2017; 317: 290–300
6. Kovach CP, Fletcher GS, Rudd KE, et al: Comparative prognostic accuracy of sepsis scores for hospital mortality in adults with suspected infection in non-ICU and ICU at an academic public hospital. *PLoS One* 2019; 14:e0222563
7. García-Gigorro R, Sáez-de la Fuente I, Marín Mateos H, et al: Utility of SOFA and Δ-SOFA scores for predicting outcome in critically ill patients from the emergency department. *Eur J Emerg Med* 2018; 25:387–393
8. Rhee C, Zhang Z, Kadri SS, et al; CDC Prevention Epicenters Program: Sepsis surveillance using adult sepsis events simplified eSOFA criteria versus sepsis-3 Sequential Organ Failure Assessment criteria. *Crit Care Med* 2019; 47:307–314
9. Rhee C, Jones TM, Hamad Y, et al; Centers for Disease Control and Prevention (CDC) Prevention Epicenters Program: Prevalence, underlying causes, and preventability of sepsis-associated mortality in US acute care hospitals. *JAMA Netw Open* 2019; 2:e187571
10. Matt V, Matthew H: The retrospective chart review: Important methodological considerations. *J Educ Eval Health Prof* 2013; 10:12
11. 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
12. 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
13. Krippendorff K: Estimating the reliability, systematic error and random error of interval data. *Educ Psychol Meas* 1970; 30:61–70
14. Winden TJ, Boland LL, Frey NG, et al: Care everywhere, a point-to-point HIE tool: Utilization and impact on patient care in the ED. *Appl Clin Inform* 2014; 5:388–401
15. 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
16. 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
17. R Core Team: R: A Language and Environment for Statistical Computing. Vienna, Austria, R Foundation for Statistical Computing, 2018. Available at: https://www.R-project.org/. Accessed July 2, 2018
18. Charlson ME, Pompei P, Ales KL, et al: A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *J Chronic Dis* 1987; 40:373–383
19. Quan H, Sundararajan V, Halfon P, et al: Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care* 2005; 43:1130–1139